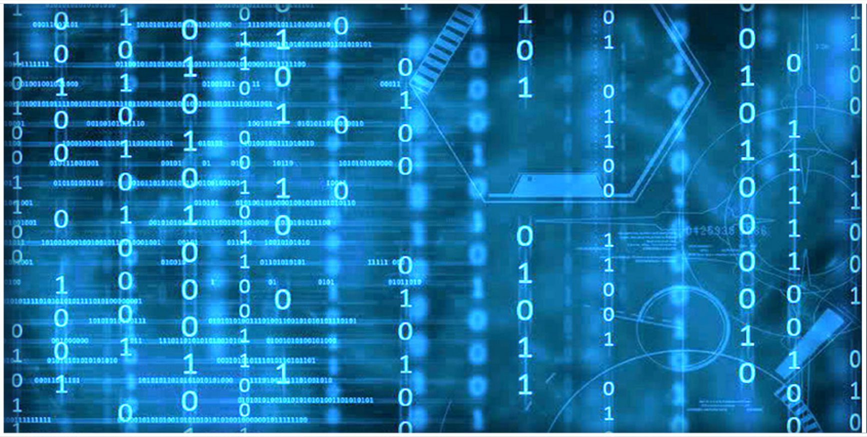
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 15 Issue 9

September 2024



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA

Volume 15 Issue 9 September 2024
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Evolving Software Architectures from Monolithic Systems to Resilient Microservices: Best Practices, Challenges and Future Trends

Martin Kaloudis
Provadis School of International Management and Technology, Frankfurt, Germany

Abstract—Microservice architecture has emerged as a widely adopted methodology in software development, addressing the inherent limitations of traditional monolithic and Service-Oriented Architectures (SOA). This paper examines the evolution of microservices, emphasising their advantages in enhancing flexibility, scalability, and fault tolerance compared to legacy models. Through detailed case studies, it explores how leading companies, such as Netflix and Amazon, have leveraged microservices to optimise resource utilisation and operational adaptability. The study also addresses significant implementation challenges, including ensuring data consistency and managing APIs. Best practices, such as Domain-Driven Design (DDD) and the Saga Pattern, are evaluated with examples from Uber's cross-functional teams and Airbnb's transaction management. This research synthesises these findings into actionable guidelines for organisations transitioning from monolithic architectures, proposing a phased migration approach to mitigate risks and improve operational agility. Furthermore, the paper explores future trends, such as Kubernetes and AIOps, offering insights into the evolving microservices landscape and their potential to improve system scalability and resilience. The scientific contribution of this article lies in the development of practical best practices, providing a structured strategy for organisations seeking to modernise their IT infrastructure.

Keywords—Service-Orientated Architecture; SOA; microservices; monolithic architecture; migration

I. Introduction

Microservice architectures [1] have gained significant traction in recent years, primarily due to their ability to address the scalability and flexibility limitations of traditional monolithic systems. This architectural paradigm shift is driven by the growing need to manage complex, distributed applications efficiently. By decomposing applications into independently deployable services, microservices offer enhanced modularity, fault tolerance, and adaptability, positioning themselves as a superior alternative to monolithic architectures in large-scale, dynamic environments.

While the benefits of microservices, including independent scalability, enhanced fault isolation, and faster deployment cycles, are well-documented, their adoption is not without challenges. Ensuring consistency in data across distributed services remains a critical issue, particularly in environments where services manage their own databases. Furthermore, the operational overhead of managing an increasing number of APIs can result in significant complexity, particularly as systems grow in scale. These issues underscore the need for robust strategies to mitigate the operational challenges inherent in microservice architectures.

This study contributes to the ongoing discourse by proposing a structured approach to the transition from monolithic to microservice architectures, focusing on best practices derived from industry case studies. While existing literature extensively covers the theoretical benefits of microservices, there is a notable gap in actionable, empirically validated strategies for managing the complexities associated with their implementation. By analyzing case studies from industry leaders such as Netflix and Amazon, this research offers a phased migration strategy that minimizes risks and operational disruption. The novelty of this study lies in its practical framework for managing the inherent challenges of microservices, particularly in the context of large-scale enterprise systems.

II. Theoretical Basics

A. Definition and Characteristics of Microservice Architecture

Microservice architecture is a software development approach in which an application is developed as a collection of small, independent services. Each service fulfils a specific business requirement and communicates with other services via precisely defined APIs [2]. Microservices are small, independent services that fulfil specific business requirements. In [3] it is emphasised that this architecture simplifies the development and maintenance of complex systems due to its loose coupling and high cohesion. The author in [4] emphasises that microservices are particularly suitable for systems that place high demands on scalability and flexibility. One of the features of microservice architecture is decentralisation, in which services, functions and data are decentralised, resulting in a loose coupling of components. This promotes the application's reliability and fault tolerance. Another feature is independent development and deployment, which means that each service can be developed, tested and deployed independently. Errors in one service do not affect the entire application, which increases fault tolerance. Services can be reused in different applications, which increases efficiency and development speed [5].

B. Monolithic Architecture

Monolithic architecture is a traditional approach to software development in which all components of an application are integrated into a single, cohesive code base. This tight integration means that the application is developed, tested and deployed as an inseparable whole. A key feature of this approach is dependent scaling: if one component of the application experiences a higher load and requires more resources, the entire application must be scaled. This can be inefficient and resource-intensive, as the underutilised parts of the application also have to be scaled.

Another feature of monolithic architecture is that development cycles tend to be longer. As all components are closely interlinked, a change in one part of the application potentially affects many other parts. This requires extensive testing and can delay the release of new features. Any change, no matter how small, often requires a redeployment of the entire application, which is not only time-consuming but can also lead to downtime. This downtime can be particularly critical if the application provides business-critical functionality. Monolithic systems (see Fig. 1) have traditionally been favoured for their consistency and simplicity of implementation and management. Developers only have to deal with one code base and one deployment process. This can speed up initial development and simplify management, especially for smaller applications or teams. The clear structure and centralised management of dependencies and configurations make monolithic architectures attractive for many use cases.

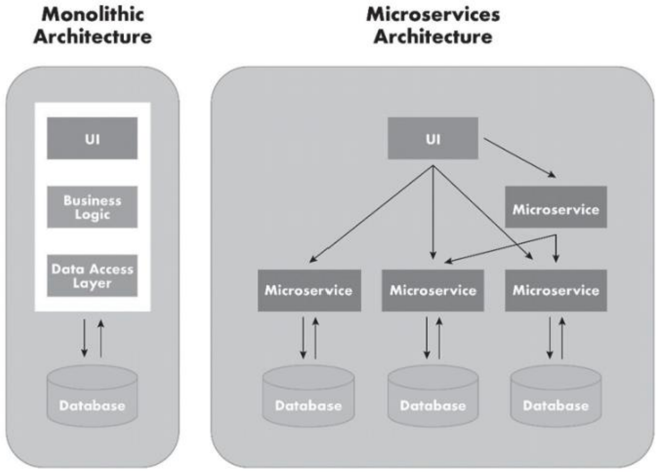


Fig. 1. Monolithic vs. Microservices architecture from [6].

To summarise, although monolithic architecture offers advantages due to its simplicity and consistency, it has significant disadvantages in an increasingly dynamic and scaled IT landscape. The lack of flexibility and scalability as well as the potential risks due to the tight integration of components make it unsuitable for many modern use cases. These disadvantages have led to the development and spread of more flexible and scalable architectures such as microservice architecture, which eliminate the specific weaknesses of the monolithic approach.

C. Microservices Architecture

In contrast to monolithic architecture, microservices divide an application into a collection of loosely coupled, independently deployable services, each of which fulfils specific business requirements. This architecture offers better scalability, flexibility and fault tolerance, but requires advanced knowledge of distributed systems development and DevOps practices. Each microservice has its own database and can be developed, tested and deployed independently, reducing infrastructure complexity and enabling more efficient resource utilisation [2].

1) Decentralisation and loose coupling: The microservice architecture is characterised by a fundamental decentralisation of services, functions and data. Instead of developing a monolithic application that combines all functions and processes in a single, closely linked structure, the microservice architecture breaks down the application into a large number of smaller, independent services. Each of these services, also known as a microservice, is designed to fulfil a specific business requirement or functionality. These services are not only functionally independent, but also often operate in isolated runtime environments and have their own databases. This means that each microservice manages and stores its own data, which ensures better data consistency. This decentralisation leads to a loose coupling of the components. Loose coupling means that the individual services are only minimally dependent on each other. Changes or errors in one service therefore have little to no impact on the other services. This decoupling enhances the overall resilience of the application, defined as its capacity to maintain operational continuity in the presence of faults or failures. Resilience describes the ability of a system to remain functional despite errors or faults. In a monolithic architecture, an error in one component can affect the entire application, whereas in a microservice architecture, an error remains isolated and the other services continue to function normally. This not only reduces fault tolerance, but also increases the overall reliability of the application [2].

One of the critical features of microservice architecture is its ability to scale individual services independently. As each microservice has its own database and is operated independently of the other services, each service can be scaled individually depending on the specific requirements and the load to be managed. This is particularly beneficial in cloud environments where resources can be allocated dynamically. For example, if a particular service has a high volume of traffic, it can be scaled independently of the other services without having to scale the entire application. This leads to more efficient resource utilisation and lower operating costs. Separation into independent services also improves fault isolation. Fault isolation means that problems in one service do not directly affect other services. If a microservice fails or a problem occurs, this error is limited to the affected service and does not affect the entire application. This not only makes troubleshooting easier, but also increases the reliability of the application. Developers can focus on fixing the specific problem without having to worry about changes to one service negatively impacting other parts of the application.

By decentralising services, functions and data, the microservice architecture offers considerable advantages in terms of reliability, fault tolerance, data consistency and scalability. The loose coupling of the services leads to increased robustness of the application, as errors remain isolated and the other services can continue to work undisturbed. Independent scalability enables efficient resource utilisation and reduces operating costs, while improved fault isolation and recovery increases the overall reliability of the application [6].

2) Scalability and fault tolerance: The ability to scale independently is one of the most outstanding features of microservice architecture and brings significant benefits in terms of resource management and increased efficiency. In a conventional monolithic architecture, all components of an application must be scaled together, even if only a small part of the application actually experiences an increased load. This leads to inefficient resource utilisation and increased costs, as not all parts of the application require the same scalability. In contrast, microservices enable targeted and demand-orientated scaling of individual services. Each microservice can be scaled independently of the other services, based on the specific requirements and the current load [7]. Splitting the application into independent services also has a positive effect on the fault tolerance of the entire system architecture. In a monolithic system, an error in one component can affect the entire system and lead to a total failure. This is because the components are closely interconnected and there is a dependency that disrupts the entire operating process. Microservices, on the other hand, isolate these errors to the affected service. If a microservice fails or an error occurs, this has no impact on the other services. The application remains functional and the affected microservice can be analysed and repaired in isolation.

Another aspect that increases fault tolerance is the ability to recognise and rectify errors automatically. Modern microservice architectures often utilise monitoring and management tools that continuously monitor the status of the services and react automatically in the event of anomalies or errors. This can be done by restarting the faulty service, switching to redundant services or dynamically reallocating resources. These automated processes reduce downtimes and improve the overall availability of the application [8]. The resilience, i.e. the ability of a system to recover from disruptions, is significantly improved by the microservice architecture. The loose coupling of the services means that they can work largely independently of each other. This independence allows the system to react flexibly to changes or failures without affecting the entire application. If a service is overloaded by a sudden increase in requests, it can be scaled in isolation to cope with the increased load. Should a service nevertheless fail, alternative services or failover mechanisms can be activated to ensure the continuity of business processes.

3) Independent development and provision: Microservice architecture facilitates the independent development and deployment of software components, yielding considerable improvements in both the efficiency and agility of the development process. In traditional monolithic architectures, all parts of an application must be developed, tested and deployed as a single unit. This means that even small changes to a component require extensive testing and full deployment of the entire application. This dependency leads to longer development cycles, an increased risk of errors and downtime as well as limited flexibility when implementing new functions [5]. A key advantage of this independent development and deployment is improved fault isolation. In a monolithic architecture, an error in one component can affect the entire application, which can lead to extensive downtime and difficult troubleshooting. In a microservice architecture, an error in one service remains limited to that specific service and does not affect the other parts of the application. This independence of services also encourages parallel development by different teams. In monolithic systems, development teams must coordinate their work closely to avoid conflicts, which can slow down development processes. In a microservice architecture, different teams can work on different services at the same time without their work interfering with each other.

The independent development and provision of microservices also supports better scalability of development resources. In monolithic systems, the scaling of development teams is often limited, as all teams have to work on the same code base and coordinate changes. In a microservice architecture, development teams can be scaled flexibly as they work independently on different services. This allows organisations to use their development resources more efficiently and respond more quickly to business requirements, resulting in faster implementation and improving the flexibility and agility of development processes. Improved fault isolation, parallel development by different teams and support for CI/CD practices lead to faster and more reliable releases, higher productivity of development teams and better scalability of development resources [5].

4) Reusability and flexibility in technology selection: Flexibility in technology selection allows teams to develop customised solutions that are optimised for their specific business needs. For example, a team working on a data-intensive analytics service might choose a programming language such as Python, which is known for its powerful data science libraries and frameworks. Another team developing a high-performance, critical real-time service might choose a language like Go or Rust, which are known for their efficiency and low latency. This freedom in technology choice leads to a better customisation of solutions to the specific needs of each service and therefore to business requirements [5]. Another advantage of this flexibility is the ability to introduce and use specialised technologies that are particularly suitable for specific tasks. Teams can select technologies that best fit the requirements and challenges of their specific microservices without having to consider the rest of the application. This can lead to a significant improvement in performance and efficiency. For example, a team working on a machine learning model could use specific frameworks and hardware acceleration to optimise training times and model accuracy [9].

The reusability and technological independence of microservices also help to reduce technical debt. Technical debt arises when short-term solutions are chosen that lead to higher maintenance costs in the long term. By using proven and reusable services, development teams can create consistent and maintainable code bases that reduce long-term maintenance efforts. In addition, flexibility in technology selection allows teams to continuously use the best tools and practices to minimise technical debt [4].

These advantages contribute to the microservice architecture being a favoured choice for modern, scalable and flexible software development projects [7].

III. Why and How Split Monoliths?

Industries such as retail, travel and transport and automotive have increasingly begun to break up their monolithic applications into microservices in order to become more flexible and scalable. This change is being driven by the need to respond more quickly to market changes and reduce total cost of ownership (TCO). However, the transition from monolithic to microservice architectures is a complex process that requires careful planning and a systematic approach. Fig. 2 shows cost-based determination of granularity services.

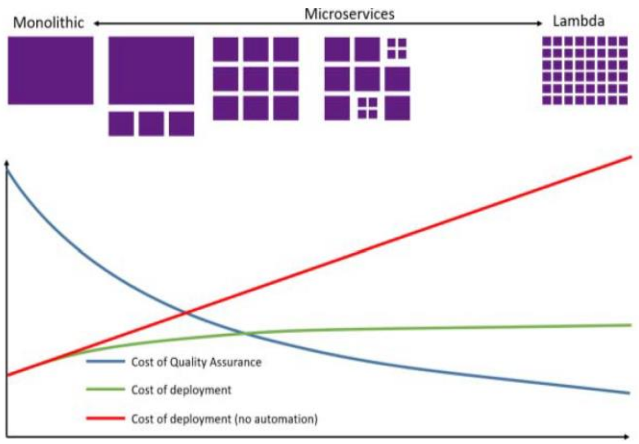


Fig. 2. Cost-based determination of granularity services [10].

A. *Reasons for the Switch to Microservices*

An important reason for this migration is the greater flexibility that microservices offer. In retail, for example, companies can develop and introduce new functionalities faster by splitting their applications into smaller, independent services. This is particularly important in a market that is constantly evolving and where competition is fierce. Retailers need to be able to respond quickly to new trends and customer demands, be it by introducing new payment methods, optimising supply chains or personalising the shopping experience.

B. *Systematic Analysis and Step-by-Step Migration*

The process of migrating from monolithic to microservice architectures often begins with a comprehensive and systematic analysis of the existing architecture. The aim of this analysis is to identify the current dependencies, bottlenecks and weak points. Based on these findings, companies can develop a clear migration strategy that is implemented step by step. A complete switch to microservices in a single step is usually too risky and too complex. Therefore, many companies prefer a step-by-step migration in which they gradually break down the application into microservices.

C. *Identification of Business Areas*

An important aspect of migration is the definition of business units. Business units are functional areas within an organisation that have clearly defined responsibilities. For example, a retailer might define business domains such as inventory management, order fulfilment, customer service and payment processing. Each of these domains can then be implemented as an independent microservice. This clear demarcation makes it possible to reduce the complexity of the overall application and clearly define responsibilities.

D. *Formation of Cross-Functional Teams*

Another important step in the migration process is the formation of cross-functional teams. Traditionally, development and operations teams are separate in many organisations, which can lead to communication problems and delays. However, microservice architecture requires close collaboration between these teams. Cross-functional teams consisting of developers, testers, operations experts and other relevant professionals can make the development and deployment of microservices more efficient. These teams are responsible for the entire lifecycle of a microservice, from development and testing to deployment and maintenance.

E. *Introduction of DevOps Processes*

The introduction of DevOps practices is another key component in the transition to microservices. DevOps stands for the integration of development and operations and aims to improve collaboration between these two areas. DevOps practices include Continuous Integration (CI) and Continuous Delivery (CD), which enable faster and more reliable delivery of software. By using automation tools and processes, companies can increase their efficiency, reduce the error rate and shorten the time to market for new functions.

IV. Challenges of Migration

Overall, the transition from monolithic applications to microservices offers significant benefits for many companies in the retail, travel and transport and automotive industries. By conducting systematic analyses, identifying business units, forming cross-functional teams and implementing DevOps practices, companies can make their IT infrastructure more flexible and scalable. A step-by-step migration minimises risks and enables continuous adaptation to changing market requirements. Research and practical reports prove the positive effects of this transformation on the efficiency and competitiveness of companies [10].

Procedure: The migration from a monolithic to a microservice architecture is a complex process that requires careful planning. It usually starts with the identification and extraction of business domains as independent microservices. Business domains, i.e. specific areas within an organisation, form the basis for the new architecture [11].

Analyse the existing architecture: The first step is to analyse the monolithic system to understand the dependencies between the components. Tools can automatically create dependency diagrams that help to visualise the interactions and control the migration process.

Development of a migration plan: Based on this analysis, a detailed migration plan should be created outlining the steps to minimise risk and ensure continuity. The plan should prioritise the domains to be migrated according to their business value and technical complexity.

Identifying and extracting business areas: Business areas, such as the product catalogue or payment processing in an e-commerce platform, must be clearly defined before they can be implemented as independent microservices.

Support for DDD: DDD helps to manage complexity by dividing systems into manageable units. The concept of "bounded context" ensures that each microservice has clear boundaries, which simplifies development and scaling.

Minimising risk and ensuring continuity: During the migration, mechanisms such as APIs or messaging systems help to ensure communication between microservices and the monolith and maintain continuity during the process.

Iterative development and continuous improvement: Migration should be viewed as iterative. Each migrated domain provides insights for optimising the process for future domains.

Static and dynamic analysis: Static analysis checks the source code to determine dependencies, while dynamic analysis monitors runtime behaviour and helps to prioritise services based on usage patterns.

By combining these approaches, companies can reduce system complexity and build scalable, maintainable architectures.

Development of a comprehensive migration plan: Based on the findings from the static and dynamic analysis, a comprehensive migration plan can be developed. This plan should take into account the identified services and interfaces as well as the prioritised usage patterns. It contains detailed steps for carrying out the migration, including the order of migration of the individual services, the necessary changes to the infrastructure and the implementation of transition mechanisms to ensure business continuity.

Static and dynamic analyses are crucial methods for preparing a successful migration from monolithic to microservice architectures. While static analysis reveals the structure and dependencies of the existing system, dynamic analysis provides valuable insights into the actual usage and performance of the application. By combining both approaches, organisations can develop a solid and low-risk migration strategy that takes into account both technical and operational aspects.

V. Advantages of the Integration of Microservices

Microservices offer numerous advantages that improve software development, reduce operating costs and simplify maintenance. Key benefits include independent development and deployment, efficient resource utilisation, technological flexibility, fault isolation and reusability of services.

One of the most important advantages of microservices is the ability to develop and deploy services independently of each other. In contrast to monolithic architectures, where every change requires extensive testing and a complete deployment of the entire system, microservices allow individual services to be updated independently of each other. This speeds up development, reduces errors and facilitates continuous deployment [11].

Efficient use of resources: Microservices enable independent scaling of services and thus optimise resource allocation. For example, if a service experiences increased demand, it can be scaled independently without affecting the rest of the system. This improves performance and reduces costs, especially compared to monolithic systems where the entire application has to be scaled [12].

Technological flexibility: Each microservice can be developed with the technology best suited to its needs. This allows development teams to innovate and customise solutions more effectively. Teams working on performance-critical components may opt for faster, more efficient languages, while others may favour simple development [13].

Isolation of errors: Errors in a microservice do not affect the entire application, which increases reliability. Isolated errors make it easier to diagnose and rectify problems without causing system-wide downtime [14].

Reusability of services: Microservices are independent units that can be reused in different applications. For example, an authentication service developed for one application can be reused in other projects, which saves development time and ensures consistency and security.

These advantages make microservices a favoured choice for modern, scalable and flexible software systems.

VI. Performance of Microservice Architectures: Case Studies

A. Case Study I: Comparison of Performance

A performance comparison between monolithic and microservice architectures shows clear differences in efficiency and scalability under different load conditions. In this case study, extensive tests were carried out to analyse the performance of the two architectures (see Fig. 3).

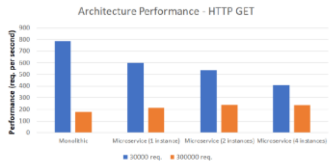


Fig. 8. Architecture performance (HTTP GET) - results

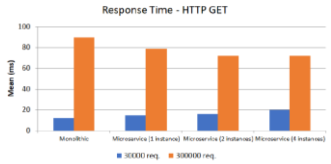
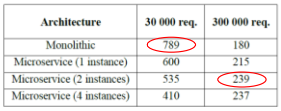


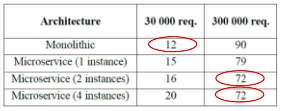
Fig. 9. Response time (HTTP GET) - results

TABLE I. ARCHTECTURE PERFORMANCE – HTTP GET



Architecture	30 000 req.	300 000 req.
Monolithic	789	180
Microservice (1 instance)	600	215
Microservice (2 instances)	535	239
Microservice (4 instances)	410	237

TABLE II. RESPONSE TIME – HTTP GET



Architecture	30 000 req.	300 000 req.
Monolithic	12	90
Microservice (1 instance)	15	79
Microservice (2 instances)	16	72
Microservice (4 instances)	20	72

Fig. 3. Performance test from [15].

Performance under lower load: The tests showed that monolithic architectures work more efficiently under lower load. This is because monolithic architectures combine all components and functions into a single, integrated application. This tight integration enables optimised use of resources and minimal communication latency between components. With low user numbers and few requests, the monolithic architecture is therefore able to deliver stable and fast response times. The reduced complexity and the lack of communication effort between distributed services contribute to greater efficiency under low load.

Performance at higher loads: As the load increases, however, the results change in favour of the microservice architecture. The tests showed that microservices scale better at higher loads and therefore achieve better performance results. In scenarios with increasing numbers of users and requests, the microservice architecture was able to handle the load more efficiently thanks to horizontal scaling. This means that additional instances of the individual microservices were provided to meet the increased demand. Of particular note is the use of replication, where multiple copies of a microservice are operated simultaneously to evenly distribute the load and increase availability. This ability to scale flexibly and on demand leads to improved performance under high load compared to monolithic systems. Fig. 4 shows the performance tests.

Monolithic architectures can therefore be more efficient at low loads, while microservice architectures show their strengths at high loads and scalability. The choice of architecture should therefore be based on the specific load requirements and the expected usage patterns.

B. *Case Study II: Scalability and Reliability*

The scalability and reliability of an application are critical factors for its performance and usability. This case study highlights the differences between horizontal and vertical scaling and shows how the choice of scaling strategy influences the scalability and reliability of the application.

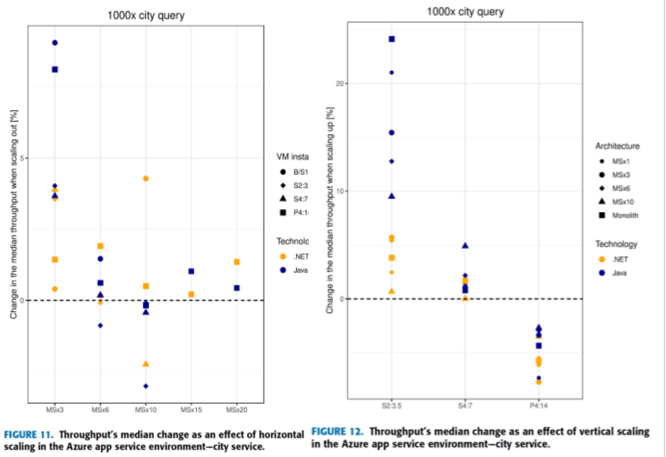


Fig. 4. Performance test from [16].

Horizontal scaling: With horizontal scaling, also known as "scaling out", additional instances of a service are added to cope with the load. This method is particularly suitable for applications with a high load and the need for flexible scalability. Horizontal scaling involves running multiple copies of the service in parallel, which distributes the load evenly. This not only improves performance, but also increases fault tolerance, as the failure of one instance can be compensated for by the other instances. An example of this could be a web server that is supported by additional server instances when data traffic increases in order to distribute requests efficiently and minimise response times.

Vertical scaling: Vertical scaling, also known as "upscaling", involves adding additional resources to a single instance of a service, e.g. more CPU, RAM or storage space. This method is better suited to low to medium load applications where requirements can be met by upgrading existing hardware. Vertical scaling can be easier to implement as no changes to the software architecture are required. However, it comes up against physical and economic limits, as the performance of a single instance cannot be increased indefinitely. A typical example would be a database that is scaled by adding more memory and more powerful processors to enable the processing of larger amounts of data.

C. *Decision Criteria for the Scaling Strategy*

The decision between horizontal and vertical scaling depends on various factors, including the specific requirements of the application, the expected load patterns and the existing infrastructure. Horizontal scaling offers more flexibility and higher fault tolerance, but is more complex to implement and manage. Vertical scaling is easier to implement, but has limited scalability and can reach its limits with very high loads.

Horizontal scaling may therefore be ideal for applications with high loads and flexible scaling requirements, while vertical scaling may be suitable for applications with moderate loads and specific hardware requirements. The decision in favour of one or the other strategy should be carefully weighed up based on the individual requirements and objectives of the application [15].

VII. Results and Best Practices

The scientific literature and documented case studies, for example on the performance of monolithic and microservices architectures, which this article provides an overview of, are diverse. What has been missing so far is a best-practice approach that generically describes how to proceed with a monolithic application landscape in order to achieve a decentralised and resilient microservices architecture - in other words, a kind of "recipe". This is proposed in this article and inductively derived from the existing articles cited above, consolidated and outlined below.

Migrating from monolithic systems to a microservice architecture is a strategically challenging task that requires not only technical expertise, but also careful planning and implementation. There is no "best-of-breed" approach because, as described above, the procedural and technological complexity of application architectures in companies is individual.

However, reference can be made to examples from which a generic migration path can be derived. Based on the case studies analysed above and the scientific work, the author recommends the following best practice derived from case studies and thus a strategy for the analysis and implementation of microservice architectures.

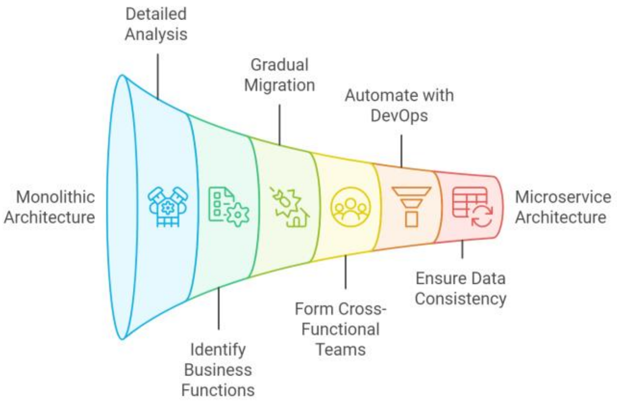


Fig. 5. Migration from monolithic to microservices.

1) Detailed analysis of the existing architecture: A thorough analysis of the existing monolithic architecture is an important first step on the way to a successful migration to a microservice architecture (see Fig. 5). Without a deep understanding of the current components, their dependencies and interactions, splitting them into microservices can lead to unforeseen complications. The migration process should therefore begin with a comprehensive analysis of both the system architecture and the underlying code. Two primary approaches play a central role in this analysis: static and dynamic analysis.

Static analysis: In static analysis, the source code is analysed to determine the dependencies between individual modules and their relationships. This method helps to map the existing structure of the monolith and visualise the connections between the components. The tools used for static code analysis can create dependency diagrams that provide a clear overview of how the system works as a whole. These insights are crucial for identifying potential microservices and ensuring that the modularisation of the system is effective and sustainable [10].

Dynamic analysis: While static analysis focuses on the structure, dynamic analysis captures the behaviour of the application during runtime. Monitoring the real-time behaviour of the system allows engineers to understand usage patterns and critical business processes supported by specific modules. This method provides insight into performance bottlenecks and areas of the system in need of optimisation, which can inform which components should be prioritised for migration into standalone microservices.

Example: Amazon carried out a comprehensive analysis of its monolithic architecture before switching to a microservice architecture. Critical areas such as the product catalogue and the payment system were identified and spun off as independent microservices, which significantly improved the scalability of the system [13].

2) Identification and delimitation of business functions: Decomposing a monolithic system into well-defined business functions is essential for creating clear service boundaries when migrating to microservices. The domain-driven design approach ensures that each microservice corresponds to a logical business domain, resulting in loosely coupled services that can operate independently.

DDD: DDD is a methodological approach that focuses on mastering complexity by modelling business domains. A key concept in DDD is the "bounded context", which delineates a specific area within a business domain where a consistent model is applied. Defining these bounded contexts ensures that microservices have clear responsibilities, reducing interdependencies and simplifying development, maintenance and scaling.

Example: Netflix used DDD to separate user management from its video streaming service. This logical separation enabled independent development and provision of functions and minimised the risk of system-wide failures [13].

3) Gradual migration and minimisation of risks: A phased or step-by-step migration strategy is critical to minimising the risks associated with the transition from a monolithic architecture to microservices. A "big bang" migration - where the entire system is migrated at once - can lead to serious system failures and business disruption. Instead, migrating smaller, less critical components initially allows for a smoother and safer transition.

Step-by-step migration strategy: The aim of a step-by-step migration is to divide the migration process into smaller steps so that individual components of the monolith can be migrated gradually. This approach allows each microservice to be thoroughly tested to ensure that it works independently and integrates smoothly with the remaining monolith. By prioritising low-risk areas, companies can significantly reduce the likelihood of critical failures during migration [12].

Example: Spotify opted for a step-by-step migration by first converting its playlist management system to a microservice. This approach enabled the company to tackle problems in isolation and minimise the risk of widespread outages [5].

4) Formation of cross-functional teams: In addition to the technical changes, the migration to microservices also requires organisational restructuring. The formation of cross-functional teams is essential for the efficient development and maintenance of microservices. These teams consist of employees from different disciplines - development, operations and quality assurance - who work together to provide services more effectively.

Autonomous teams: Each cross-functional team has full responsibility for one or more microservices and works independently within the scope of the services assigned to it. This autonomy enables the teams to develop, deliver and optimise their services without being dependent on other teams, which increases productivity and flexibility within the company.

Example: Uber formed cross-functional teams that were responsible for individual microservices, such as route calculation. These teams worked independently of each other, which enabled faster adaptation to changes in the business model or technology, which significantly improved productivity [12].

5) Automate with DevOps practices: Automation plays a central role in the successful provision and management of microservices. Continuous Integration (CI) and Continuous Delivery (CD) are two key DevOps practices that ensure reliable and rapid delivery of microservices. CI/CD pipelines automate the processes of testing, building and deploying software, enabling faster releases and greater system stability.

DevOps practice: DevOps emphasises collaboration between development and operations teams to ensure continuous improvement of both development and deployment processes. By automating testing and deployment, human error is minimised, allowing teams to release frequent, smaller updates that improve the quality and reliability of the overall system.

Example: Facebook implemented CI/CD pipelines to provide several microservices every day. This enabled faster releases and greater agility in the development process [13].

6) Consideration of data consistency and error tolerance: Ensuring data consistency in distributed systems is one of the biggest challenges in microservice architectures (see Fig. 6). In a microservice system, each service often manages its own database, which can make consistency across the entire system difficult. Two important techniques for overcoming this challenge are event sourcing and the saga pattern.

Event sourcing: With event sourcing, the state of the application is saved as a sequence of events that change this state instead of saving the current state directly in the database. These events are stored in an event log that can be replayed as required to reconstruct the current state. This ensures that all changes are traceable and recoverable and that data consistency is maintained across distributed systems.

Saga pattern: The Saga pattern enables the decomposition of long transactions into smaller, atomic transactions that can be managed by individual microservices. Each transaction is designed to either complete or roll back in the event of an error, ensuring that all services involved either reach a consistent state or return to their previous state, thereby avoiding inconsistencies.

Example: Airbnb uses the Saga pattern to coordinate transactions such as bookings across multiple microservices. This pattern ensures that the system remains consistent even if errors occur in individual services [3].

7) Summary of the 6 phases: Successful migration from a monolithic to a microservice architecture requires a strategic, well-planned approach that takes both technical and organisational aspects into account.

Firstly, a thorough analysis of the existing system is crucial. Using static and dynamic analysis techniques ensures a comprehensive understanding of component dependencies and enables informed decisions on how to split the monolith into manageable microservices. Once the architecture is understood, it is important to identify and delineate the business functions using approaches such as DDD. This ensures that the microservices are aligned with the logical business domains, resulting in clear boundaries, reduced dependencies and more effective scaling. To minimise risks, a step-by-step migration strategy is recommended. Gradually migrating smaller, low-risk components allows for testing and optimisation, reducing the likelihood of critical errors and ensuring a smooth transition. On the organisational side, the formation of cross-functional teams is essential. These teams, made up of experts from development, operations and quality assurance, should be able to manage individual microservices independently to increase both productivity and flexibility. Automating processes through DevOps practices, particularly CI and CD, ensures that microservices are deployed efficiently and reliably. Automation minimises human error and speeds up the development cycle, allowing for frequent, smaller updates. Finally, ensuring data consistency and fault tolerance in distributed systems is a major challenge that can be addressed with techniques such as event sourcing and the Saga pattern. These methods help maintain consistent states and handle failures gracefully to ensure system reliability and robustness.

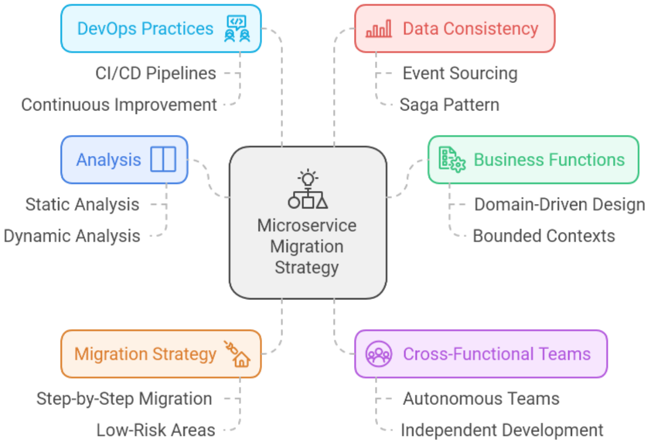


Fig. 6. Challenges and best practices in migrating from a monolithic to a microservice architecture.

If companies follow these best practices an included recommendations, they can minimise the risks of switching to microservices and at the same time benefit from the advantages of a scalable, flexible and resilient system architecture.

VIII. Discussion

A. Data Consistency and Management of Distributed Data

Microservice architecture's distributed nature enables greater flexibility in data management, allowing each service to optimize its data storage based on specific requirements. However, maintaining data consistency across distributed services presents significant challenges. Techniques like event sourcing and the Saga pattern are often employed to ensure synchronized, up-to-date transactions across services. While effective, these methods increase the complexity of system architecture, adding to operational overhead and necessitating advanced technical knowledge [3]. This complexity may not always justify the benefits for smaller organisations or systems with lower scalability needs, where a monolithic approach could be more practical [6].

B. Trade-offs Between Microservices and Traditional Architectures

Microservices offer clear advantages in scalability and fault tolerance, but there are notable trade-offs. For example, smaller organisations may struggle with the overhead of managing numerous independent services, APIs, and maintaining data consistency. Monolithic architectures, by contrast, can offer faster development cycles and simpler management, especially in smaller applications that don't require significant scalability [6]. As discussed by [12], the benefits of microservices tend to emerge in large-scale environments where scalability is critical. Thus, the decision to adopt microservices should be weighed against the organisation's size, technical capabilities, and future growth plans [7].

C. API Management and Overheads

While microservices communicate through APIs, promoting modular design and interoperability, the management of multiple APIs can become a burden as systems grow. This issue, often termed API proliferation, can increase operational complexity and management costs [5]. Solutions like API gateways and service meshes help centralize API management and streamline communication, providing advanced features such as load balancing and security. However, these tools also introduce new layers of infrastructure, which may pose challenges for smaller organisations without the technical resources to maintain them [10].

D. Testing Strategies for Microservices

One advantage of microservices is their ability to support isolated unit testing, reducing the risk of system-wide failures [18] [20]. Automated testing tools can efficiently validate microservices before deployment. However, ensuring that multiple microservices function as a cohesive system requires extensive integration testing, which can slow deployment cycles and increase operational complexity. As emphasized, integration testing in microservices introduces an additional layer of complexity not present in monolithic systems [17].

E. Increased Focus on Emerging Trends

Technologies like Kubernetes, AIOps, and serverless computing are shaping the future of microservices by offering advanced automation and orchestration capabilities. Kubernetes, for instance, simplifies the management of microservices by providing container orchestration tools for scaling and fault tolerance [18]. However, Kubernetes' complexity often requires specialised expertise, making it more suitable for larger enterprises. AIOps—which integrates machine learning to predict system failures and optimize performance—offers significant potential for improving microservice reliability, but also introduces additional complexity [19]. As pointed out, the success of these technologies depends on how well organisations can manage this complexity [8].

F. Linking Case Studies More Critically

The adoption of microservices by companies like Netflix, Amazon, and Uber has been widely documented, but the unique contexts that facilitated their success must be critically evaluated [6]. For example, Netflix's need for high reliability in streaming services and Amazon's demand for rapid scalability due to their e-commerce platform both necessitated the use of microservices [13]. However, smaller companies, or those in industries with lower scalability needs, may not experience the same benefits. Case studies of large companies should therefore be viewed with caution when attempting to generalise these strategies to smaller firms [5].

G. Potential Gaps in Existing Research

Although microservices have been widely adopted across various sectors, there remain gaps in research concerning their implementation in highly regulated industries like fintech and healthcare, where data security and regulatory compliance are crucial [11]. These industries face challenges in adopting decentralised architectures due to the need for strict data governance. More research is needed to explore how microservice best practices can be adapted for these sectors. Moreover, there is limited empirical data evaluating the long-term performance of microservices in such contexts [16].

H. More Quantitative Evaluation

Empirical studies have shown that monolithic systems tend to perform more efficiently under lower loads, while microservice architectures excel at higher loads, thanks to their ability to scale horizontally. Quantitative data on resource utilization, fault tolerance, and operational costs would provide a stronger foundation for decision-making when comparing these architectures [15]. For instance, valuable insights into the performance benefits of microservices under varying conditions is provided [16].

I. Critical Look at Migration Strategies

A phased migration approach, where organisations gradually transition from monolithic to microservice architectures, is often recommended to mitigate risks [7]. This approach allows for continuous testing and ensures that each microservice functions independently before migrating the entire system. However, this can also extend the migration [21] process and lead to technical debt, as both systems must be maintained during the transition. In some cases, a "big bang" migration, where the entire system is migrated at once, might be more efficient, especially for smaller systems [5]. Organisations must carefully assess their specific needs, technical capacity, and risk tolerance before deciding on the best migration strategy [10].

IX. Conclusion

To summarise, the evolution from monolithic to service-oriented and finally to microservice architectures represents a significant advance in the development and maintenance of modern software applications. Microservice architectures address many of the challenges associated with traditional monolithic systems and service-oriented architectures and offer significant improvements in flexibility, scalability and fault tolerance. By splitting complex applications into independent, loosely coupled services, microservices enable organisations to better respond to changing business requirements and optimise resource utilisation. As highlighted in the best practices and recommendations, the key benefits of microservices include their modular structure, which enables independent development, deployment and scalability. This modular approach allows organisations to scale services based on specific requirements without impacting the overall system.

However, data consistency and Application Programming Interface (API) management remain a major challenge and require sophisticated strategies such as event sourcing and the Saga pattern to maintain synchronisation across distributed systems.

The recommended step-by-step migration strategy helps to minimise the risks associated with the transition from monolithic systems to microservices. This step-by-step approach, together with cross-functional teams, ensures a smoother migration process and promotes collaboration between development, operations and quality assurance. In addition, the adoption of DevOps practices such as CI and CD increases the efficiency and reliability of microservice delivery, even though this requires a high level of technical maturity. While the benefits of microservices are obvious, the increased complexity and operational overhead created by their distributed nature, as well as the need for advanced skills in DevOps and container orchestration, present challenges that must be carefully managed. However, the integration of new technologies such as AIOps, Kubernetes and serverless computing increases the potential of microservice architectures and positions them as the dominant model for scalable, flexible and resilient software systems in the future.

Companies that strategically apply these best practices and recommendations will be better equipped to overcome the challenges of microservice architectures while maximising the benefits of scalability, flexibility and fault tolerance in their IT infrastructures.

References

[1] Valdivia, J.A., Lora-González, J., Limón, X., Cortes-Verdin, K., & Ocharán-Hernández, J.O. (2020): "Patterns related to microservice architecture: a multivocal literature review". Programme. Comput. Software, 46, 594-608.

[2] Newman, S. (2021). "Building Microservices". O'Reilly Media, Inc.

[3] Abdelfattah, A.S. & Cerny, T. (2023): Roadmap to reasoning in microservice systems: a rapid review. Appl. Sci, 13, 1838.

[4] Cadavid, H., Andrikopoulos, V., & Avgeriou, P. (2020): "The architecture of systems of systems: A tertiary study". Inf. Software Technol. 118, 106202.

[5] Lewis, J., & Fowler, M. (2014). "Microservices: A definition of this new architectural term".

[6] Vecherskaya, S.E. (2023). "Tasks and evoluti on of microservice architecture", pp. 37-43, Complex systems: models, analysis, management, Bulletin of the Russian New University.

[7] Baškarada, S., Nguyen, V., & Koronios, A. (2020). „Architecture of Microservices: Practical Opportunities and Challenges".

[8] Van Eyk, E., Iosup, A., Seif, S., & Thömmes, M. (2017). „The spec cloud group's research vision on faas and serverless architectures". In WOSC 2017 - Proceedings of the 2nd International Workshop on Serverless Computing, Part of Middleware 2017 (pp. 1-4). Association for Computing Machinery, Inc. https://doi.org/10.1145/3154847.3154848.

[9] Zimmermann, O. "Microservices tenets". Comput Sci Res Dev 32, 301-310 (2017).

[10] Gouigoux, P., & Tamzalit, D. (2017). "From Monolith to Microservices: Lessons Learned on an Industrial Migration to a Web Oriented Architecture".

[11] Evans, E. (2004). Domain-Driven Design: "Tackling Complexity in the Heart of Software". Addison-Wesley Professional.

[12] Baškarada, S., Nguyen, V., & Koronios, A. (2020). „Architecture of Microservices: Practical Opportunities and Challenges".

[13] Newman, S. (2015). "Building Microservices". O'Reilly Media.

[14] Dragoni, N., Lanese, I., Larsen, S.T., Mazzara, M., Mustafin, R., Safina, L. (2018). Microservices: "How To Make Your Application Scale". In: Petrenko, A., Voronkov, A. (eds) Perspectives of System Informatics. PSI 2017. Lecture Notes in Computer Science(), vol 10742. Springer, Cham. https://doi.org/10.1007/978-3-319-74313-4_8.

[15] Gos, K., & Zabierowski, W. (2020). "The comparison of microservice and monolithic architecture".

[16] Blinowski, G., Ojdowska, A. and Przybyłek, A., 2022, "Monolithic vs. Microservice Architecture: A Performance and Scalability Evaluation", in IEEE Access, vol. 10, pp. 20357-20374, 2022, doi: 10.1109/ACCESS.2022.3152803.

[17] Tran, H.K.V., Unterkalmsteiner, M., Börstler, J., & bin Ali, N. (2021). „Evaluating the quality of test artefacts - a tertiary study". Inf. Software Technol.

[18] Arvanitou, E.M., Ampatzoglou, A., Bibi, S., Chatzigeorgiou, A., & Deligiannis, I. (2022): "Application and exploration of DevOps: A tertiary study". IEEE Access, 10, 61585-61600.

[19] Liu, X., Li, S., Zhang, H., Zhong, C., Wang, Y., Waseem, M., & Babar, M.A. (2022): "Microservice architecture research: A tertiary study". SSRN Electron. J.

[20] Alaasam, A.B., Radchenko, G., Tchernykh, A., & González Compeán, J.L. (2020): "Analytical study on containerisation of stateful stream processing as a microservice to support digital twins in fog computing. Programme". Comput. Software, 46, 511-525.

[21] Fritzsch, J., Bogner, J., Wagner, S., & Zimmermann, A. (2019). „Microservices Migration in the Industry: Intentions, Strategies, and Challenges".

The Effects of IDS/IPS Placement on Big Data Systems in Geo-Distributed Wide Area Networks

Michael Hart[1], Eric Richardson[2], Rushit Dave[3]
College of Science, Engineering, & Technology, Minnesota State University, Mankato, United States[1, 3]
College of Health and Human Services, University of North Carolina Wilmington, United States[2]

***Abstract*—Geographically-distributed wide-area networks (WANs) offer expansive distributed and parallel computing capabilities. This includes the ability to advance Wide-Area Big Data (WABD). As data streaming traverses foreign networks, intrusion detection systems (IDSs) and intrusion prevention systems (IDSs) play an important role in securing information. The authors anticipate that securing WAN network topology with IDSs/IPSs can significantly impact wide-area data streaming performance. In this paper, the researchers develop and implement a geographically distributed big data streaming application using the Python programming language to benchmark IDS/IPS placement in hub-and-spoke, custom-mesh, and full-mesh network topologies. The results of the experiments illustrate that custom-mesh WANs allow IDS/IPS placements that maximize data stream packet transfers while reducing overall WAN latency. Hub-and-spoke network topology produces the lowest combined WAN latency over competing network designs but at the cost of single points of failure within the network. IDS/IPS placement in full-mesh designs is less efficient than custom-mesh yet offers the greatest opportunity for highly available data streams. Testing is limited by specific big data systems, WAN topologies, and IDS/IPS technology.**

***Keywords*—Information security; network topology; wide-area big data; wide-area networks; wide-area streaming**

I. INTRODUCTION

Increasingly, organizations must collect large amounts of data that is located in physically distanced data centers. Geographically-distributed big data server clusters provide massive scale data analytic capabilities across wide-area networks (WANs). Several big data frameworks in use at the time of this writing such as Apache Spark are deployed within single data centers [1]. However, big data clusters that run in local area networks (LANs) do not necessarily have the same challenges as WANs. For instance, LANs have certain advantages like bandwidth, shorter distance routing, and highly available communication at cheaper costs. LANs also have limitations spanning from local resources to global connectivity [2].

WANs enlarge the capabilities of LANs, offering expansive resources and connectivity for geo-distributed data streaming. For instance, WANalytics research is investigating how to optimize distributed structured query language (SQL) queries across WANs [3]. Subsequently, unsupervised machine learning provides several possibilities to enhance geo-distributed data streaming. For example, a sliding version of the hidden Markov model (SlidHMM) improves bottleneck detection in WAN data analytics [4]. Despite the latter progress, a survey on geo-distributed frameworks found that research is lacking in several areas. This includes decentralized architecture, data streaming, multi-clusters, information security, and privacy [1]. The objective of this work is to investigate the role of information security in geo-distributed big data analytic framework literature and provide subsequent steps toward securing this infrastructure in future research.

Organization of the paper is as follows. The authors perform a review of literature on the influence of information security on geographically-distributed big data systems in Section II. A methodology develops from the review that identifies procedures to test the performance of secured WAN topologies in Section III. The results of the testing and a discussion are given in Section IV and Section V respectively. Finally, Section VI concludes the study.

II. RELATED WORK

To better understand big data frameworks and their geographically-distributed contributions, Bergui [1] performed a survey of existing literature. A theme in progression centers around optimizing big data systems for the ever-increasing changes in network topology. Tuning these systems for WANs is complex, yet not always clear in existing literature. For example, in [1], bandwidth-aware systems do not always use resource managers like yet another resource negotiator (YARN) with specific WAN tuning capabilities. The researchers also emphasize further work is necessary to study information security and system architectures in geo-distributed big data systems. Trust models become more complex when distributing data between different governments. Researchers encourage designing authentication strategies and decentralized architecture frameworks capable of supporting more complex geo-distributed clusters [1].

To better understand research that helps optimize big data systems, the writers review data querying, transfer, placement, and their environments, which includes network topology.

A. Querying Data

Research on optimal geo-distributed computing architectures is ongoing. In study [3], the researchers introduce the term WANalytics, which they contrast with Wide-Area Big Data (WABD). WABD typically copies data from multiple data centers to a single data center where data analytics transpire. WANalytics is designed to support massive scale geo-distributed analytics across multiple data centers. Its goal focuses on reducing expensive WAN bandwidth while maintaining compatibility with data sovereignty restrictions [3].

Initial experimentations demonstrate that WANalytics can reduce data transfer costs by as much as 360 times compared to centralized data center methods. This occurs by allowing users to test SQL queries between data centers in Europe, North America, and South-East Asia. While WANalytics shows tremendous progress toward optimizing geo-distributed computing architectures, information security appears to be distant in this literature [3].

Demand for wide-area data analytics enforces the need to advance the capabilities of geo-distributed big data systems. For instance, Wang and Li [4] propose the Lube system framework to monitor, detect, and resolve bottlenecks in in geo-distributed data analytic queries. Benchmarks show optimizing scheduling policies across distributed data centers can lower query response times up to 33 percent when compared to other big data systems like Apache Spark. Similar to Lube [4], Turbo [5] has the ability to improve geo-distributed data analytics queries at runtime. Using machine learning, Turbo optimizes data analytic query execution plans across multiple physically distanced data centers. In a geo-distributed Google Cloud environment that spanned eight regions, Turbo lowered query completion times by 41% [5].

In study [6], the authors focus on common executions in wide-area network streaming analytics queries. Examples of common execution elements include shared data processing and input data. While improvements are achievable using common query executions in streaming analytics, researchers emphasize that without WAN awareness, weaker performance can exist in geo-distributed data center communications. WAN-aware multi-query optimizations that leverage common executions can reduce WAN bandwidth as much as 33% in contrast to systems that fail to use shared execution components. Therefore, multi-query efficiency may have some dependency on WAN-awareness [6]. Despite the advancements of wide-area data analytics in geo-distributed analytics, many questions remain. Researchers in [1] note that further work is beneficial to address variations in the structure of data, determine the optimal features to reduce query completion times, and construct a larger range of performance metrics to measure bottlenecks. Another complementary vein of research focuses on bulk data transfer.

B. Bulk Data Transfer

Transferring bulk data within inter-datacenter networks requires efficient strategies to reduce associated costs. Multimedia big data such as video streams and gaming content, compete for leftover bandwidth in backbone transport networks that connect geographically-distributed data centers. However, the exponential increase of data transfer these services need can degrade backbone networks [7]. Though certain algorithms can efficiently manage guaranteed traffic and reassignment [8], it is well understood that middleware and control plane protocols are amongst several layers of the architecture that require greater attention in research [9]. As one example of progress toward the latter goal, software defined networking (SDN) helps dissociate the control plane from data paths. This leads to more dynamic adjustment of data routing as network environmental attributes change [10].

Particularly when sending bulk data transfers between geo-distributed data centers, researchers in study [10] highlight three primary services. This includes 1) task admission control, 2) data routing, and 3) store-and-forward. Task admission control rejects or accepts network transfer requests based upon whether they can be completed by a specified deadline. Data routing must choose the best path data should take to reach its destination, which can include rerouting through intermediate data centers. The concept of store-and-forward decides whether it is more efficient to store data temporarily within intermediate data centers and forward it at a more optimal time than the immediate time of execution. If so, decisions must be made to determine where the data is temporarily stored until it reaches its destination [10].

C. Data Placement

Subsequent focus on efficiently distributing data between data centers are algorithms that calculate cloud service provider (CSP) costs [11]. Certain data sent between CSPs can tolerate delays, which can be transferred using store-and-forward intermedia storage nodes with off-peak internet service provider (ISP) bandwidth that is already financed [12]. Multi-rate bandwidth on-demand (BoD) brokers employ scheduling algorithms to optimize the use of this residual bandwidth. As an example, the BoD broker in study [6] uses standby wavelengths within the wavelength division multiplexing (DWDM) layer to decrease peak network bandwidth. Adjustments are possible based on delay-intolerant and delay-tolerant transfer requests. Compared to relational algorithms like First-Come-First-Served (FCFS), more precise use of time slots in all wavelengths is optimal when peak bandwidth results in delayed or blocked requests [7].

When inter-datacenter networks are congested, certain storage decisions can help reduce additional network load. This includes the use of intermedia storage (IS) and edge storage (ES). ES allows certain types of jobs like bulk data transfers to leverage storage at the edge of network domains and forward it during periods of off-peak CSP bandwidth. In study [13], as network load increases there is a linear decrease in the success of bulk transfers. Bulk data transfers are optimal when the allowed wait time is twice the aggregated network load. In summary, the authors found that ES and IS perform similar when peak bandwidth times are small. Medium or less network load results in little difference between ES and IS. However, in this research IS performed significantly better than ES in times of high network load [13].

Research on bulk data transfer across low latency or congested links is helping advance several needs including scheduling optimization [9], bandwidth costs [11], and delay tolerance [12]. In reviewing related literature, information security is not a central component of big data transfers between geo-distributed cloud data centers [7, 10], inter-datacenter bulk transfers [11, 13], or research networks [8]. Additionally, while certain testing considers differences in specific network topology [7, 10] others do not [8, 12]. Therefore, opportunities may exist to study the influence of information security and network topology on bulk data transfers in low latency network environments. To explore this further, the authors turn to the role of network topology and geo-distributed big data systems.

D. Network Topology

Network topology influences several dimensions of geo-distributed big data systems, including the elasticity of nodes in a cluster [12]. A challenge of big data streaming is resource provisioning across shared cloud infrastructure. Particularly when the cloud tenant does not own infrastructure, it can be challenging to decipher the cause of poor performance on collective physical hardware that runs virtual machines (VMs). In study [14], the authors highlight the need for the dynamic rescheduling of big data streaming tasks using multitenant-aware resource provisioning that is independent of the VM hypervisor. Software defined networking (SDN) plays an impactful role in this provisioning by supporting load balancing between cloud-based VM clusters. In contrast to other network topologies, SDN can define its topology in real time. This in turn allows for additional cloud node elasticity [14]. In study [12], researchers focus on optimizing bulk data transfer in a geo-distributed data center system using SDN architecture. SDN elasticity promotes dynamic routing decisions using bulk data transfers in pieces in contrast to handling transfers as endless flows [12].

Like [13], researchers in study [15] highlight a need to optimize big data streaming strategies between geo-distributed data centers. The authors note that traditional methods for distributed data streaming such as task assignment are insufficient when high throughput data exists along with low latency WAN links [15]. Researchers also emphasize the need to perform data mining on data sent between WANs from streaming applications that perform user-clicks, social networks, and Internet of Things (IoT) hardware [16]. A proposed advancement is an SDN-based resource provisioning framework capable of monitoring WANs, identifying an optimal selection of big data worker nodes, and more efficiently assigning tasks to the chosen nodes. In initial tests, SDN resource provisioning results in minimal processing time that is 1.64 times faster on the tested environment, which included Apache Flink, Apache Spark, and Apache Storm [15].

One of the challenges of geo-distributed and wide-area network data analytics streaming is identifying performance problems when infrastructure is not under the control of the customer. Multitenant-aware resource provisioning using SDN network topology is a proposed solution when cloud computing hardware is shared amongst multiple customers [14]. Monitoring and increasing performance of multitenant streaming analytics also requires more advanced worker node and IoT placement strategies in low latency network topology. Streaming platforms like S4, Apache Storm, and Apache Spark were not initially designed for low latency analytics shared between users and applications in distributed IoT systems. However, improvements are being made in the streaming platforms. For instance, Apache Spark supports structured streaming via PySpark, a Python API. Spark streaming has the capability to stream data in micro-batches [1]. In study [16], the GeeLytics platform is introduced as an alternative streaming platform to address low latency networks. This includes more dynamic mechanisms to balance real-time streaming in the cloud and network edges. The proposed design is expected to reduce edge-to-cloud bandwidth use for IoT data analytics. It is also engineered to increase customer insight into multi-tenancy system efficiency [16].

Proposed in study [17], a worker node placement framework focuses on wide-area streaming analytics. It builds upon the Simple Additive Weighting (SAW) method. In this model, a central global manager determines how tasks are assigned across multiple edge data centers using a proposed SAW-based Node Ranking (SNR) algorithm. Task slots are determined based upon the amount of input data and processing power of each slot. Additionally, task slots communicate over the WANs links. This allows the global manager to maintain the status of key link metrics including cost, delay, and bandwidth as well as identify network topology changes. Researchers tested the SNR algorithm on Apache Flink, Apache Spark, and Apache Storm using small, medium, and large graphs to simulate different network sizes. Each big data system shows performance improvements compared to other worker node placement strategies [17].

WAN traffic costs are central to several recent advancements in geo-distributed streaming analytics research. Costs are influenced by network design. For instance, the hub-and-spoke design includes several network edges that interconnect via WANs to a central data warehouse. Popular streaming analytics service providers use this model at the time of this writing [18]. Important to this network model is determining the optimal amount of computation that should exist at the center of the topology or the edge. Based on the hub-and-spoke network topology, researchers have identified staleness or the delay in retrieving data results and WAN traffic as pivotal metrics. Experiments using common analytics from large CDNs highlight the need to minimize both latter metrics [18].

AggNet is a subsequent advancement in research focused on reducing WAN traffic costs. Developed on the Apache Flink framework, AggNet [19] reduces WAN bandwidth by aggregating a percentage of real-time data analytics closer to the location of end users. Aggregation from AggNet implementation has shown 47% to 83% decreases in traffic costs when compared to traditional costs from relevant industry organizations that included Akamai and Twitter [19].

Although the hub-and-spoke network model is a cornerstone in recent geo-distributed streaming analytics work [18-19], researchers understand current network topology must change to meet the future needs of big data analytics. In study [20], researchers argue that high communication cost, data sovereignty, and data privacy challenge the feasibility of central data center designs. Proposing the concept of geo-distributed machine learning (Geo-DML), parameter server (PS) placement remains a challenge for distributing raw machine learning data between WANs. A proposed solution is using approximation algorithms capable of selecting the optimal data center for training using network cost. Results of using this strategy reduce communication cost up to 21.78% over other Internet network topology. However, the potential effect of IDS/IPS hardware is unknown [20].

E. Summary

Several advancements are occurring that improve geo-distributed big data systems. AggNet helps reduce WAN traffic by placing data closer to end users [19]. In study [18], researchers develop a hybrid online algorithm to determine optimal computation at the network edges versus the center in a hub-and-spoke WAN model. In small to large network topology, the SNR algorithm shows capability to optimize tasks across geo-distributed data centers using the simple-additive weighting method [17]. Subsequently, an approximation algorithm finds the best data center as the parameter server for machine learning training on two network topologies, which included a Google private WAN and a United States Internet with nine interconnected data centers [20]. Like research on bulk data transfer between geo-distributed data centers, little emphasis exists on information security in these papers [17-20]. Additionally, network topologies are limited to only a few different types of WANs [20] as well as traditional hub-and-spoke designs [18]. SDN-based networks also show promise in helping optimize resource provisioning but may need additional consideration as they gain more traction in geo-distributed WAN analytics [14].

The research that follows presents an elementary investigation into whether IDS/IPS placement impact the performance of big data systems operating between low-latency network topologies.

III. Methodology

The research design follows the information systems research framework outlined in study [21]. Three pillars of the framework include the environment, information systems research, and the coinciding knowledge base. Within the environment stage of the latter research methodology, this paper focuses on building modern IT infrastructure to support massive-scale data analytics. Subsequently, the research stage focuses on WAN simulations to evaluate supporting network topology for capable big data systems. The researchers add to the existing knowledge base by reporting on the effects of IDS/IPS placement on real-time data streaming systems in network topologies able to migrate into modern SDN-enabled WANs.

Following the design science methodology, business needs are the driver for building new information system artefacts [21]. Wide-area data analytics is gaining traction due to the increased need for businesses to analyze real-time data streams in multiple physical locations [6]. Notably, big data systems in geo-distributed data centers provide immense opportunity to support streaming massive amounts of data on low-latency WAN connections. Provisioning resources across modern SDN WAN architectures, provides big data systems like Apache Spark with more expansive horizontal scalable than centralized data centers [15]. To support the growing business need for geo-distributed streaming, the researchers design and implement current WAN topologies capable of efficiently and securely connecting physically distanced big data systems.

Investigators design and implement three well recognized Software-defined-wide area network (SD-WAN) arbitrary topologies outlined by study [22], including hub-and-spoke, full-mesh, and custom-mesh. Each are implemented across ISP leased lines. The applied network topology uses the specifications engineered by Cisco Systems in their Cisco Extended Enterprise SD-WAN Design Guide. These are located in Fig. 7 Hub-and-Spoke Topology with Cisco IR1101 and Fig. 8 Mesh Topology with Cisco IR1101 and SD-WAN in study [23].

A. Experimental Network

The experimental Cisco Systems network resides in an enterprise-class data center. Within the research network, the authors design and implement wide area network (WAN) data centers in four major United States cities. The central data center is located in New York, New York. From the New York data center, WAN links connect to data centers via routers in the cities of Orlando, Florida, Los Angeles, California, and Seattle, Washington. Router placement and configuration for each WAN parallel Cisco IR1101 and SD-WAN in [23]. WAN network latency between the latter data centers equals averages, at the time of this writing, in milliseconds (MS) published by AT&T in study [24]. The full-mesh network topology, which includes network latency for all WAN links, is outlined in Fig. 1.

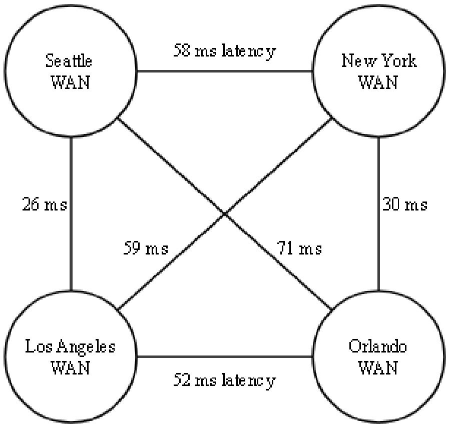


Fig. 1. WAN network latency.

B. Big Data System Architecture

The experimental environment includes four big data server clusters in each data center. Clusters are connected by the WANs and secured by intrusion detection systems (IDSs) and intrusion prevention systems (IPSs). All server and router hardware are the same make and model. Hardware has precisely the same specifications including physical CPUs, memory, and solid state disks. Each data center houses a Dell PowerEdge server running an updated Microsoft Hyper-V Server 2019. Virtual machines hosted in Hyper-V consist of Intel Xeon processors with five physical CPU cores and 24 gigabytes of memory.

Fig. 2 shows the big data system architecture for cluster one (C1) connected to the New York WAN. Each of the four system clusters parallel this architecture. The clusters consist of six big data system VMs running the Ubuntu 22.04 Long Term Support (LTS) server operating system. Two VMs are dedicated Apache Hadoop name nodes. The primary and secondary name nodes connect to four data nodes with a replication factor of three. Data nodes are configured as both Apache Hadoop and Apache Spark worker nodes. Name nodes connect to the WAN through a router and an IDS/IPS. The WAN routers at each site also have one external facing Dell PowerEdge server with 5 physical CPU cores and 24 gigabytes of memory. The latter WAN-connected Ubuntu 22.04 LTS servers measure and collect performance data between the geo-distributed data centers. The edge servers are also the source of all external data streams sent to the big data clusters. Table I shows the corresponding software and versions of the big data systems.

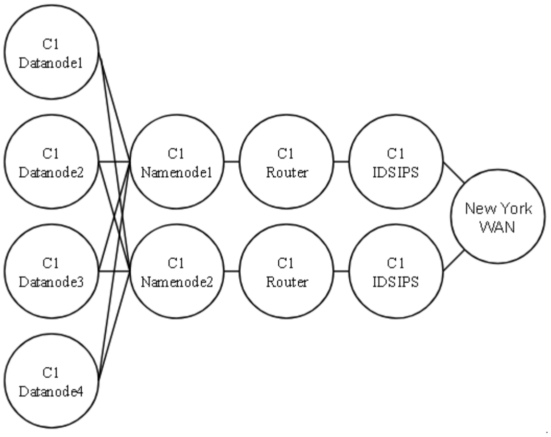


Fig. 2. Cluster architecture.

Experiments use Suricata for the intrusion detection system (IDS) and intrusion prevention system (IPS). Suricata is well supported by the open-source community as a modern world-class IDS/IPS [25]. It allows researchers to customize packet bundling techniques to analyze stream data sets efficiently and effectively [26]. Suricata is compiled with the emerging threats open ruleset [25]. Specific Suricata rules allow the unique public IP addresses of the streaming clients to connect to a primary and secondary name node in each data center cluster. Streams and associated IPS rules use the customized TCP port range of 9990 – 9999 on each big data cluster. With the exception of the experimental data streams and SSH for system administrator IP addresses, no other external traffic is allowed into the data center networks by the IDSs/IPSs.

TABLE I. EXPERIMENTAL SOFTWARE VERSIONS

Software	Version
Hadoop	3.3.6
Iptables	1.8.7
Nmon	16
OpenJDK	8u412
Pdsh	2.31-3
Pyspark	3.5.1
Python	3.10.12
Spark	3.5.1
Suricata	6.0.4
Tcpdump	4.99.1
Ubuntu	22.04.4

C. Streaming Architecture

Within the big data system architecture, the primary and secondary name nodes are configured as Apache Spark streaming servers. Fig. 3 outlines the big data streaming architecture. From an Ubuntu server on each WAN, 1 GB streams are sent to the primary and secondary name nodes. To process the data streams the authors developed a big data streaming application using Apache PySpark. The application facilitates the unstructured data streams to Apache Spark on the primary and secondary name nodes. It uses the Spark context object and PySpark streaming class instudy [27] to develop the streaming functions. Each application instance processes word counts on the data streams. Word counts are aggregated using key value pairs using Spark in-memory computation and subsequently written across the Hadoop Distributed File System (HDFS) for long-term data analytics. HDFS block sizes are configured for 128 MBs.

D. Benchmarking Technologies

Simulation is one of several methods in the design science research framework [21] that helps assess and refine novel artifacts. Central to this work is determining how modern IDS/IPS placements impact the performance of geo-distributed big data system clusters. The researchers use raw network performance statistics between connecting WANs to evaluate real-time data streams. In study [28] researchers evaluate raw network performance using httping and iperf3 on anonymous circuit-based communications. The networking utilities were able to effectively measure the average latency and throughput between hubs in a metropolitan area. Iperf3 is also used in WAN environments to test network capacity. Researchers investigated the transfer of science big data across WANs in study [29] using NVMe over Fabrics (NVMe-oF). NVMe-oF is able to provide enhanced non-volatile memory functionality for storage networking fabrics. Methods in the study successfully use iperf3 to test for bottlenecks in the networks [29].

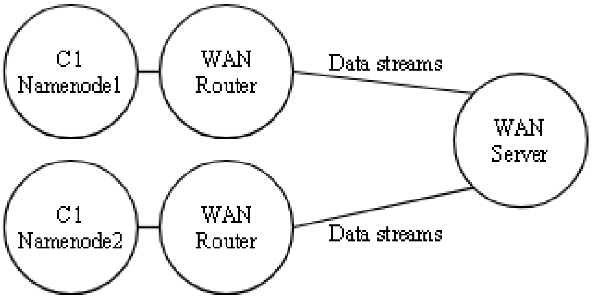


Fig. 3. Streaming architecture.

Like [29] iperf3 measures latency between the geo-distributed data centers in this study. In Fig. 2, iperf3 resides on the name node servers, IDS/IPS servers, and the WAN servers. Network latency is measured between the edge of each WAN and the name node clusters. Similar to study [26], the authors combine TCP packets into streams to analyze the network data. Libpcap, tcpdump [30], and Nigel's performance Monitor for Linux (nmon) [31] facilitate the raw network packet captures. Nmon uses the “-s” option to collect network packets every second throughout the duration of the Apache Spark streaming tests.

In addition to using network bitrate to test data streaming performance, it also determines the optimal location of the IDSs/IPSs in this study. To establish the optimal IDS/IPS placement in each network topology, researchers iteratively run the experiments with each recommendation in the Cisco Extended Enterprise SD-WAN Design Guide [23]. The authors base the final IDS/IPS location selection on the best raw network bitrate for each topology in the testing that follows.

The proposed research methodology uses a design science approach to investigate the impact of IDS/IPS placement on geo-distributed big data systems. It outlines the system architectures and benchmarking processes in the coinciding experiments. Next, the authors implement the proposed tests and report the results of the evaluations.

IV. Results

Hub-and-spoke in Fig. 4 is the first experimental network topology (T1) that tests the IDS/IPS performance of geo-distributed big data systems. WAN connections source from a central data center in New York, NY to the remote cities of Seattle, Los Angeles, and Orlando. The authors automated the tests using the Python programming language and Bourne-Again SHell (bash) scripting. This includes a start and stop script.

A. Experimental Environment

A start script prepares a consistent experimental environment for each iteration of the performance testing. A stop script resets the environment to the original state, ensuring each test begins with the same configuration. The start script begins by starting each Suricata IDS/IPS service and checking the compiled security rules. After the IDS/IPS is functioning properly, the script starts Apache Hadoop and Spark. At this stage, a health check ensures HDFS is operating correctly across the clusters. If the distributed file system is unhealthy, it exits after logging error codes. If HDFS is healthy, TCP ports 9990-9999 open for Apache Spark streaming.

Each name node on four geo-distributed big data clusters runs a parallel Python application that facilitates the system and network performance benchmarking. The Python application invokes the PySpark streaming application, establishing 1 GB data streams to the primary and secondary names nodes. Throughout the experiments, a health check monitors the Apache Hadoop and Spark logs. If the Python application fails in-memory processing or HDFS writes at any time during the real-time stream, the application exits after logging error codes. The start script sleeps for 30 seconds following invocation of the Python application to ensure streaming is functional.

Following successful execution of streaming services, a series of bash shell commands collect and aggregate raw network performance statistics using libpcap, tcpdump, nmon, and iperf3. Data aggregation is per cluster. For example, data combines from the two name nodes and two IDSs/IPSs for each site into a single file. Measurement and results are from transmission control protocol (TCP) network traffic. Tcpdump and nmon results are collected from real-time TCP traffic. Intervals for each tool are set to write performance data every second. Nmon executes with the default settings with the exception of the “-s” syntax for seconds. Iperf3 uses the IP address of each server, the connecting port, and the interval in seconds, and the bidirectional traffic syntax.

Tests invoke in parallel across each cluster using the start script. To ensure saturation, the authors ran the tests ten times for twelve minutes each. Each test produces 720 unique rows of data, of which the middle 600 rows are selected for analysis to avoid potential anomalies at the beginning or ending of the testing. Data analysis begins and ends on the same timestamp for each cluster.

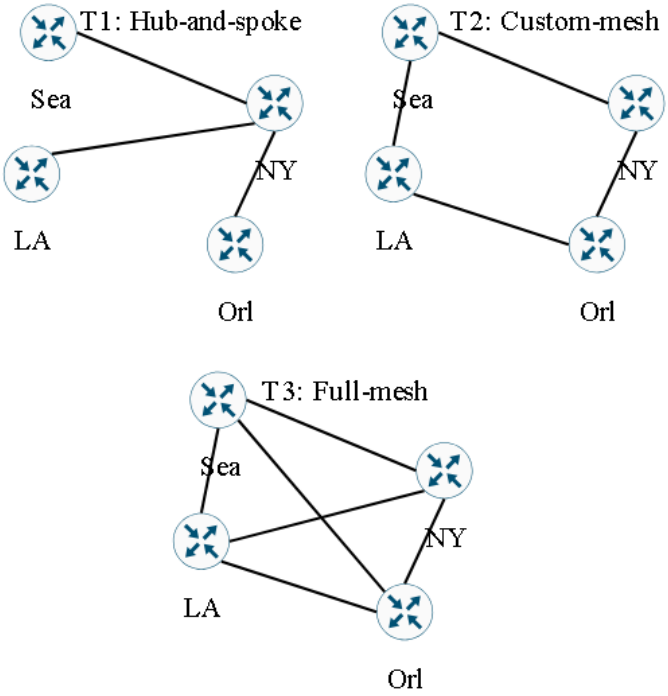


Fig. 4. WAN network topology (T).

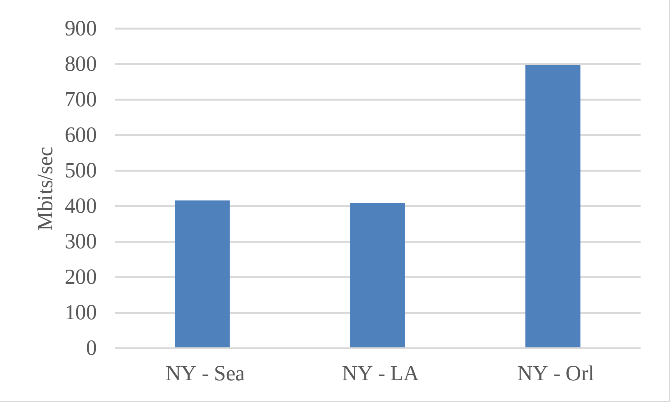


Fig. 5. Hub-and-spoke topology (T1) bitrate.

B. Hub-and-Spoke Topology

Topology 1 (T1) represents the hub-and-spoke WAN experiments. The New York data center connects to Orlando, Seattle, and Los Angeles. WAN latency is 30 milliseconds to Orlando, 58 milliseconds to Seattle, and 59 milliseconds to Los Angeles. Consistent with the cluster architecture in Fig. 2, Spark streams run from the WAN VM through dual Suricata IDSs/IPSs before reaching the primary and secondary Apache Hadoop name nodes. Data streams over three WANs are sent to the primary and secondary name nodes of each big data cluster. The name nodes load balance 128 MB HDFS block writes with a replication factor of three across the data nodes.

Fig. 5 outlines the network bitrate from the WANs to the name nodes measured in megabits per second (mbits/sec). From the New York data center, the rates are 416.496 mbits/sec to Seattle, 409.346 mbits/sec to Los Angeles, and 796.833 mbits/sec to Orlando. The mean bitrate for the hub-and-spoke topology is 540.892 mbits/sec.

C. Custom-Mesh Topology

Topology 2 (T2) represents the custom-mesh WAN experiments. In the custom-mesh network topology, the IDSs/IPSs protect the big data systems at the edge of each LAN. Dual routes exist through each IDS/IPS to the primary and secondary Hadoop name nodes. WANs have redundant paths to each LAN, allowing data streams alternative routes in case of a network failure. Testing establishes a total of eight data streams to Apache Spark. For example, in Fig. 4, New York has a data stream from Seattle and Orlando.

In the custom-mesh topology, the New York data center connects to Orlando and Seattle. WAN latency is 30 milliseconds to Orlando and 58 milliseconds to Seattle. Data streams from New York to Los Angeles route through either Seattle or Orlando. The Los Angeles data center connects to Seattle and Orlando. WAN latency from Los Angeles is 26 milliseconds to Seattle and 52 milliseconds to Orlando.

Fig. 6 outlines the network bitrate from the WANs to the name nodes measured in megabits per second (mbits/sec). From the New York data center, the rates are 795.578 mbits/sec to Orlando and 415.931 mbits/sec to Seattle. Rates from Los Angeles to Seattle are 915.41 mbits/sec and Los Angeles to Orlando 464.22 mbits/sec. The mean bitrate for the custom-mesh topology is 647.729 mbits/sec, which is 106.837 mbits/sec greater than the hub-and-spoke network topology.

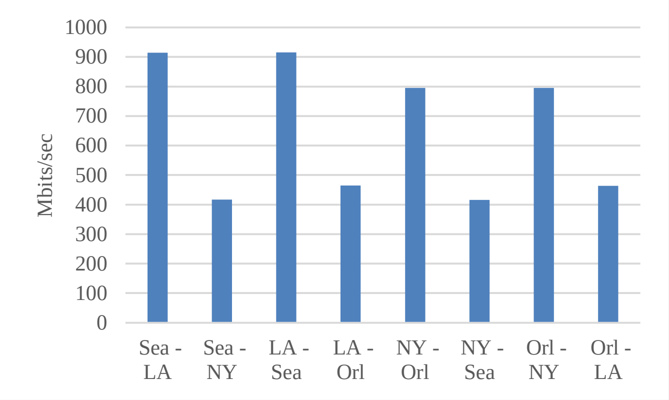


Fig. 6. Custom-mesh topology (T2) bitrate.

D. Full-Mesh Topology

Topology 3 (T3) is a full-mesh WAN design. As highlighted in Fig. 4, data centers have WAN paths to each city, providing the most redundancy of the designs. Twelve data streams are sent to the primary and secondary name nodes of each big data cluster through dual IDSs/IPSs. This is shown in the cluster architecture in Fig. 2.

Within the full-mesh topology, New York has WAN connections to data centers in Orlando, Seattle, and Los Angeles. In sequence, WAN latency from the New York data center to Orlando is 30 milliseconds, to Seattle 58 milliseconds, and to Los Angeles 59 milliseconds. Likewise, the Los Angeles data center connects to Seattle, Orlando, and New York. Los Angeles WAN latency to Seattle is 26 milliseconds and 52 milliseconds to Orlando. Orlando to Seattle WAN latency is the largest at 71 milliseconds.

Fig. 7 shows the network bitrate from the WANs to the name nodes measured in megabits per second. New York data center bitrates are 761.068 mbits/sec to Orlando, 414.98 mbits/sec to Seattle, and 409.33 mbits/sec to Los Angeles. Los Angeles data center bitrates are 462.771 mbits/sec to Orlando and 870.065 mbits/sec to Seattle. Seattle bitrates are 882.696 mbits/sec to Los Angeles, 414.995 mbits/sec to New York, and 341.19 mbits/sec to Orlando. The mean bitrate for the full-mesh topology is 544.637 mbits/sec. The mean rate is 3.745 mbits/sec more than the hub-and-spoke topology and 103.092 mbits/sec less than the custom-mesh topology.

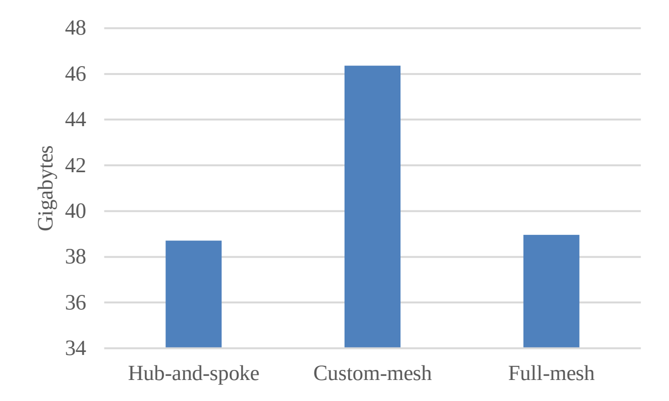


Fig. 7. Mean WAN data stream transfers in gigabytes.

E. Streaming Data Transfers

Fig. 8 highlights the mean data transfer rates of the Apache Spark streams through the WAN links. Fig. 9 illustrates the total data transfer rates of the Apache Spark streams through the WANs.

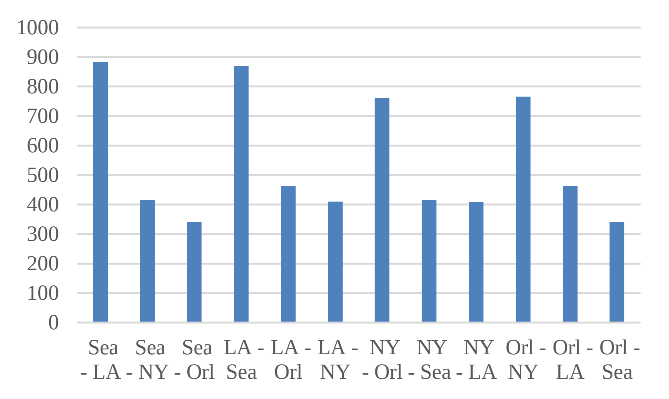


Fig. 8. Full-mesh topology (T3) bitrate in mbits/sec.

Gigabytes were converted from megabytes for the total data stream transfers. Total gigabytes transferred across the WAN network links for the hub-and-spoke network topology is 116.116. Mean gigabytes transferred between the data center sites is 38.705. Custom-mesh produces a mean of 46.364

gigabytes and a total of 370.915 gigabytes. Full-mesh network topology delivers a mean of 38.956 gigabytes and a total data transfer of 467.474 gigabytes.

Full-mesh has a mean data transfer rate slightly greater than hub-and-spoke. On the contrary, mean custom-mesh data transfer produces 7.659 gigabytes more than the hub-and-spoke network topology and 7.408 gigabytes more than the full-mesh topology.

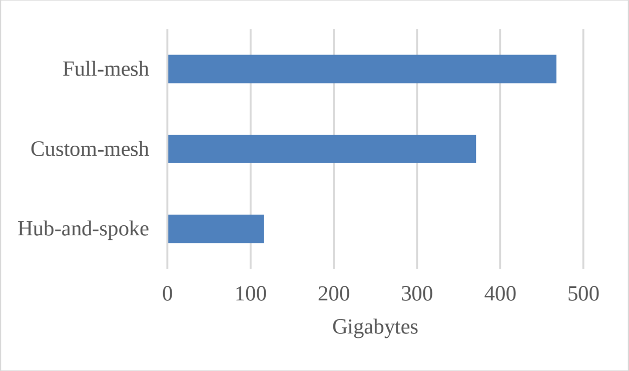


Fig. 9. Sum WAN data stream transfer in gigabtyes.

F. WAN Performance

Table II highlights the total WAN latency of each network topology along with the total amount of data transfer from the data streams. Table II also notes the number of internet service provider (ISP) leased lines used for each WAN topology in the experiments. Hub-and-spoke network topology results in an average of 38.705 gigabytes of data transfer per ISP leased line. Custom-mesh has an average data transfer of 92.728 gigabytes per leased line while full-mesh has an average data transfer of 77.912 gigabytes per leased line.

TABLE II. WAN LATENCY VERSUS DATA TRANSFER

Topology	ISP Leased Lines	Total WAN Latency	Total Data Transfer
Hub-and-spoke	3	147 ms	116.1166 Gbs
Custom-mesh	4	166 ms	370.9158 Gbs
Full-mesh	6	296 ms	467.4745 Gbs

G. Summary

To measure whether IDS/IPS placement impacts geo-distributed big data systems, the researchers study WAN connections between the remote cities of Los Angeles, Orlando, New York, and Seattle. Data centers in each city host big data clusters running Apache Hadoop and Spark. Data streams are sent through the IDSs/IPSs from the WANs to each of the four big data clusters. The researchers develop a novel Python application that uses PySpark streaming classes to facilitate real-time geo-distributed massive data streaming. Performance measures use raw network traffic data to demonstrate the results of three prominent network designs; hub-and-spoke, custom-mesh, and full-mesh. Results illustrate the ability to load balance data streams through IDS/IPS locations with the lowest WAN latency in custom-mesh topology while continuing to offer alternative network paths to geo-distributed data centers. Next, the authors discuss these results.

V. Discussion

Live data streams across four unique geo-distributed data centers show variability in real-life scenarios. Though researchers were able to optimize bandwidth through three different WAN topologies, there are clear performance differences that decision makers should consider when architecting secure clusters for WABD.

A. Geo-Distributed IDS/IPS Placement Performance

Researchers were able to achieve the fastest data streams across geo-distributed data centers using a custom-mesh network design. IDS/IPS placement in the custom-mesh network topology achieves a mean of 106.836 mbits/sec more network bitrate than the optimized hub-and-spoke topology. Similarly, on average the custom-mesh topology is 103.091 mbits/sec faster than the full-mesh design.

In this study, IDS/IPS placement within the full-mesh network design results in slightly faster mean bandwidth available for WABD data streams than the hub-and-spoke network topology. Full-mesh benefits from a mean of 3.745 additional mbits/sec across the WAN architecture. While full-mesh network topology has additional benefits over both hub-and-spoke and custom-mesh such as more fault tolerance, this comes at the cost of expensive WAN bandwidth [22].

In the experiments, hub-and-spoke has three ISP leased lines. Custom-mesh has four leased lines while full-mesh has six leased lines. When reviewing Table II, custom-mesh is able to transfer 54.023 more gigabytes of streaming data through the IDSs/IPSs per leased line than the hub-and-spoke network topology. This comes at a cost of only one additional ISP leased line in these experiments. However, it also adds an extra path of redundancy between each site, eliminating potential single points of failure in the hub-and-spoke network topology.

Custom-mesh also transfers 14.916 gigabytes more data per leased line than the full-mesh topology. Despite this result, full-mesh benefits from an additional redundant path to subsequent data centers. Similar to custom-mesh, full-mesh provides more bandwidth than the hub-and-spoke topology. In comparison, full-mesh produces 39.206 gigabytes more data per leased line than hub-and-spoke. While data centers in the full-mesh design could experience several network failures before losing complete connectivity to another site, it also comes at the cost of three additional ISP leased lines over the custom-mesh topology.

B. Limitations

This paper does not address pricing, which limits the analysis of geo-distributed IDS/IPS placements specific to big data streaming. Although the results of this study give some indication of potential efficiency of various IDS/IPS locations for geo-distributed big data systems, it is financially inconclusive as many variables determine the costs of implementing and maintaining each of the network designs in real-life environments. For instance, in study [15], custom topology resulted in considerable pricing differences for data transfer alone, ranging from $0.02 to $0.25 per GB of data transfer.

Research efforts are advancing big data worker node placement using several available data points. For example, in study [17] the simple-additive weighting method strategically places data streaming tasks using data transmission cost, latency, and bandwidth. However, algorithms lean upon available network data without considering human factors. Future research is important to consider more closely defined pricing models for IDS/IPS placement specific to geo-distributed WAN data streaming.

This paper is also limited to initial benchmarking of three traditional WAN topologies that use manual IDS/IPS placement methods. To advance this research, existing algorithms could consider IDS/IPS latency within avant-garde WAN topologies. For example, the approximate parameter server placement (APSP) algorithm proposed by study [20] could be tested in IDS/IPS environments to identify if the randomized rounding method is still applicable. Similarly, future research could test IDS/IPS locations using WAN topology-aware frameworks introduced in study [15] and study [17].

Finally, IDS/IPS benchmarking is limited to a Python streaming application engineered for Apache Spark. Similar to study [17], researchers may consider other big data streaming systems like Apache Flink and Apache Storm along with varied streaming applications developed in Scala and/or Java.

VI. Conclusion

This paper develops a PySpark streaming application in Python capable of benchmarking geo-distributed data centers secured by IDSs/IPSs. The application sends data streams across the WAN topologies of hub-and-spoke, custom-mesh, and full mesh. In each topology, the researchers optimize IDS/IPS placement using industry best practices and experimentation. The proposed placements show several tradeoffs. Hub-and-spoke has the least aggregate WAN latency and the fewest number of ISP leased lines but at the cost of single points of failure within the WAN topology. Custom-mesh network topology benefits from the fastest raw network performance. It also has dual paths to geo-distributed data centers at a cost of only one additional ISP leased line. Full-mesh offers the most fault tolerance and raw data streaming bandwidth. However, it requires a minimum of two additional ISP leased lines over custom-mesh. In summary, IDS/IPS placement in custom-mesh network topology allows engineers to customize the amount of high availability across WANs while reducing associated costs of leased lines. Advancing this work could include evolving network topology for WANalytics, automating IDS/IPS placement, testing alternative big data streaming systems, and incorporating financial costs into IDS/IPS placement determination. Subsequently, researchers may consider testing existing or new worker node placement algorithms in WABD IDS/IPS environments.

References

[1] M. Bergui, S. Najah, and N. S. Nikolov, “A survey on bandwidth-aware geo-distributed frameworks for big-data analytics,” *Journal of Big Data*, vol. 8, no. 40, pp. 1-26, Feb. 2021, doi: 10.1186/s40537-021-00427-9.

[2] “Cluster Mode Overview,” The Apache Software Foundation, June, 2023. [Online]. Available: https://spark.apache.org/docs/latest/cluster-overview.html.

[3] A. Vulimiri, C. Curino, P. Godfrey, T. Jungblut, K. Karanasos, J. Padhye, and G. Varghese, “WANalytics: Geo-distributed analytics for a data intensive world,” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, in SIGMOD ’15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1087–1092. doi: 10.1145/2723372.2735365.

[4] H. Wang and B. Li, “Mitigating bottlenecks in wide area data analytics via machine learning,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 155–166, 2020, doi: 10.1109/TNSE.2018.2816951.

[5] H. Wang, D. Niu, and B. Li, “Turbo: Dynamic and decentralized global analytics via machine learning,” in *Proceedings of the ACM Symposium on Cloud Computing*, in SoCC ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 14–25. doi: 10.1145/3267809.3267812.

[6] A. Jonathan, A. Chandra, and J. Weissman, “Multi-query optimization in wide-area streaming analytics,” in *Proceedings of the ACM Symposium on Cloud Computing*, New York, NY, USA, 2018, pp. 412–425. doi: 10.1145/3267809.3267842.

[7] A. Yassine, A. A. N. Shirehjini, and S. Shirmohammadi, “Bandwidth on-demand for multimedia big data transfer across geo-distributed cloud data centers,” *IEEE Transactions on Cloud Computing*, vol. 8, no. 4, pp. 1189–1198, Dec. 2020, doi: 10.1109/TCC.2016.2617369.

[8] K. Rajah, S. Ranka, and Y. Xia, “Advance reservations and scheduling for bulk transfers in research networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 11, pp. 1682–1697, Nov. 2009, doi: 10.1109/TPDS.2008.250.

[9] K. Rajah, S. Ranka, and Y. Xia, “Scheduling bulk file transfers with start and end times,” *Computer Networks*, vol. 52, no. 5, pp. 1105–1122, Apr. 2008, doi: 10.1016/j.comnet.2007.12.005.

[10] Y. Wu, Z. Zhang, C. Wu, C. Guo, Z. Li, and F. C. M. Lau, “Orchestrating bulk data transfers across geo-distributed datacenters,” *IEEE Transactions on Cloud Computing*, vol. 5, no. 1, pp. 112–125, Mar. 2017, doi: 10.1109/TCC.2015.2389842.

[11] T. Nandagopal and K. P. N. Puttaswamy, “Lowering inter-datacenter bandwidth costs via bulk data scheduling,” in *2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, May 2012, pp. 244–251. doi: 10.1109/CCGrid.2012.70.

[12] N. Laoutaris, G. Smaragdakis, R. Stanojevic, P. Rodriguez, and R. Sundaram, “Delay-tolerant bulk data transfers on the internet,” *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1852–1865, Dec. 2013, doi: 10.1109/TNET.2012.2237555.

[13] S. Yue, X. Lin, W. Sun, and W. Hu, “Modeling sparse store-and-forward bulk data transfers in inter-datacenter networks with multiple congested links,” *IEEE Transactions on Cloud Computing*, vol. 11, no. 3, pp. 1–14, 2022, doi: 10.1109/TCC.2022.3225977.

[14] C. Vicentini, A. Santin, E. Kugler Viegas, and V. Abreu, “SDN-based and multitenant-aware resource provisioning mechanism for cloud-based big data streaming,” *Journal of Network and Computer Applications*, vol. 126, Nov. 2018, doi: 10.1016/j.jnca.2018.11.005.

[15] H. Mostafaei and S. Afridi, “SDN-enabled resource provisioning framework for geo-distributed streaming analytics,” *ACM Trans. Internet Technol.*, vol. 23, no. 1, Feb. 2023, doi: 10.1145/3571158.

[16] B. Cheng, A. Papageorgiou, F. Cirillo, and E. Kovacs, “GeeLytics: Geo-distributed edge analytics for large scale IoT systems based on dynamic topology,” in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 2015, pp. 565–570. doi: 10.1109/WF-IoT.2015.7389116.

[17] H. Mostafaei, S. Afridi, and J. Abawajy, “Network-aware worker placement for wide-area streaming analytics,” *Future Generation Computer Systems*, vol. 136, pp. 270–281, Nov. 2022, doi: 10.1016/j.future.2022.06.009.

[18] B. Heintz, A. Chandra, and R. K. Sitaraman, “Optimizing Timeliness and Cost in Geo-Distributed Streaming Analytics,” *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 232–245, 2020, doi: 10.1109/TCC.2017.2750678.

[19] D. Kumar, S. Ahmad, A. Chandra, and R. K. Sitaraman, “AggNet: Cost-Aware Aggregation Networks for Geo-distributed Streaming Analytics,” in *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, 2021, pp. 297–311. doi: 10.1145/3453142.3491276.

[20] Y. Li, C. Fan, X. Zhang, and Y. Chen, "Placement of parameter server in wide area network topology for geo-distributed machine learning," *Journal of Communications and Networks*, vol. 25, no. 3, pp. 370–380, 2023, doi: 10.23919/JCN.2023.000021.

[21] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[22] S. A. Ibrahim Hussein, F. W. Zaki, and M. M. Ashour, "Performance evaluation of software-defined wide area network based on queueing theory," *IET Networks*, vol. 11, no. 3–4, pp. 128–145, May 2022, doi: 10.1049/ntw2.12039.

[23] "Cisco Extended Enterprise SD-WAN Design Guide," *Cisco Systems, Inc.*, July, 2024. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/solutions/Verticals/EE/DG/ee-WAN-dg.pdf.

[24] "U.S. Network Latency," *AT&T*, June, 2024. [Online]. Available: https://ipnetwork.bgtmo.ip.att.net/pws/network_delay.html

[25] "Suricata user guide," *Open Information Security Foundation*, June, 2024. [Online]. Available: https://docs.suricata.io/en/latest/.

[26] M. Kalinin and V. Krundyshev, "Security intrusion detection using quantum machine learning techniques," *Journal of Computer Virology and Hacking Techniques*, vol. 19, no. 1, pp. 125–136, Mar. 2023, doi: 10.1007/s11416-022-00435-0.

[27] "pyspark.streaming.StreamingContext," *The Apache Software Foundation*, June, 2024. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.streaming.StreamingContext.html.

[28] V. Nunes, J. Brás, A. Carvalho, D. Barradas, K. Gallagher, and N. Santos, "Enhancing the Unlinkability of Circuit-Based Anonymous Communications with k-Funnels," *Proceedings of the ACM on Networking.*, vol. 1, pp. 1-26. Nov. 2023, doi: 10.1145/3629140.

[29] S. Y. Yu *et al.*, "Analysis of NVMe over fabrics with SCinet DTN-as-a-Service," *Cluster Computing*, vol. 25, Aug. 2022, doi: 10.1007/s10586-021-03433-x.

[30] "TCPDUMP & LIBPCAP," *The Tcpdump Group*, June, 2024. [Online]. Available: https://www.tcpdump.org/.

[31] "nmon for Linux," *IBM*, June, 2024. [Online]. Available: http://nmon.sourceforge.net.

STAR, a Universal, Repeatable, Strategic Model of Corporate Innovation for Industry Domination

Ronald Berman, Nicholas Markette, Robert Vera, Tim Gehle
Grand Canyon University, College of Doctoral Studies, 3300 W. Camelback Road, Phoenix, AZ. 85017, United States

***Abstract*—Within an existing organization, internal expertise, staffing, compensation, information systems, and market focus may complicate the introduction of new ideas while culture and aversion to risk may completely derail the organizations' ability to innovate. The STAR model for corporate innovation provides a theoretical model on how to develop and execute innovative practices to overcome these obstacles and achieve significant market penetration and value. The model is a theoretical framework that empowers organizations of all sizes to construct the necessary structures and advocacy needed to create products, services, and internal processes that enable them to dominate the industry in which they participate. The model also provides the mechanism to support the identification, acceptance, and rapid deployment of relevant new technologies that offer an opportunity to create an unfair advantage, something that is very hard to replicate.**

Keywords—Corporate entrepreneurship; innovation model; market dominance; competitive advantage

I. Introduction

Few companies have strategic architecture to promote the creation and launch of bold new innovations that lead to market dominance. Without a strategic approach, they focus on fragmented tactical isolated activities. Processes do not exist that differentiate between minor product enhancements and bold new initiatives. While every organization has a culture, few, however, encourage and support innovation by fostering input from multiple diverse stakeholders. Innovation is not rewarded. Expertise is not acknowledged. Risk-taking is discouraged. Significant value creation is not considered. Rapid development and prototyping are not permitted until lengthy specification documents are prepared. Metrics do not exist to assess innovation initiatives. Thus, companies may theoretically support the notion of innovation, but the absence of an integrated strategic model limits their ability to actually innovate. These issues cited represent a small subset of the many challenges existing organizations face when attempting to strategically organize for innovation.

When considering the millions of organizations who attempt to innovate, but ultimately fail, understanding how to better conceptualize and enable innovation within an existing organization is crucially important. To begin to understand this important topic, the authors thoroughly reviewed the extant literature then created and deployed a management survey to assess corporate innovation and the necessity to create a management innovation model suitable for use by existing organizations. Survey respondents cited the benefits of innovation while also reporting the challenges encountered in their own organizations and the need to have some type of model. Based on the survey results, the authors propose the STAR model for corporate innovation described in the remainder of this paper.

II. Literature Review

The literature is replete with tactical approaches for start-up companies to create successful new products rapidly and accurately. Over the last 15 years, much of this discussion has been based on the seminal work by Eric Ries [1], who after concluding that many new products fail, proposed tactical novel approaches to product development. Instead of completing a lengthy specifications document that sometime took years to write, he proposed creating a minimum viable product based on perceived customer need. Then, through rapid development and continuous improvement, the product is refined and marketed thereby resulting in accelerated market entry. He argued that continuous improvement and focusing on delivering value to customers produces better results than the creation of traditional business plans.

Ries [2] stressed the importance of employing innovation accounting and continued the use of the term pivot (introduced in 2011) as a structured course correction. He explained that a "pivot" is a change in direction informed by feedback and data from the market, customers, or other sources. In most cases, it is not a complete change in direction, instead, it is a natural part of the iterative process. Ries [3] created a tactical step-by-step guide for implementing lean methodologies in product development. He expanded the discussion and use of a minimal viable product by incorporating practical tools and techniques to create and test customer value propositions. He also supplemented his approach with the creation of a leader's guide in which he suggests the use of innovation accounting methods to measure progress, how to manage entrepreneurial employees, and how to sustain innovation.

Reis successfully proposes tactical approaches for start-ups to create new products more rapidly. Adapting this approach to an existing company has limitations because it is proposed without consideration of an organization's infrastructure, personnel, expertise, staffing and its market focus. He addresses value creation, rapid development, hypothesis testing, and metrics. However, an existing organization is different from a new company. Ries does not address an existing organizations' structure, reward system, company culture and of course risk aversion. Most importantly, Reis does not address how to incorporate existing company resources [4], nor does he critically examine how to scale the initiative [5]. While it is known that there are tactical targeted approaches to enable innovation in start-up ventures, it is not known whether there is a need for a similar approach for existing organizations. Specifically, it is not known whether existing organizations need an integrated strategic structure that provides the framework to enable innovation on an ongoing basis. To that end, the authors initiated a survey to assess whether management within existing companies in the United States supports the need for such a model.

III. Research Methodology

In 2022, a 32-question innovation survey was distributed to managers within existing companies in the United States to assess the current nature of how innovation is enabled within their respective organizations. In as much as there is little data relative to innovation execution within existing companies, the sample population, which includes multiple industries and organizations of all sizes, creates a balanced initial management assessment relative to the innovation process inherent to these organizations.

The sample population consists of managers employed in organizations of varying sizes. As such, 32.9% of the managers represent organizations with two to 100 employees, 34.7% of the managers represent organizations with 101 to 1,000 employees, and 32.4% of the managers represent organizations with 1001 or more employees. The sample population also includes managers employed in over 40 industrial sectors. The largest industry representation consists of Retail (9.3%), Healthcare (8.8%), Manufacturing (7.9%), Information Technology (6.5%), Food Beverage (5.6%), Education (4.6%), Construction (5.6%), Computer Software (4.6%), and Banking/Finance (4.6%).

U.S. managers in this study clearly state that innovation is critical and essential to their firm's success. Eighty-five percent of U.S. managers agreed or strongly agreed that "Innovation is critical to your firm's success," and 81.9% of those same managers agreed or strongly agreed that "Innovation is essential to your organization's survival." Moreover, 88% agreed or strongly agreed that "Innovation is good for your employees" and "Innovation is good for your customers," respectively.

However, in contrast to management's positive perception, the survey results indicate an absence of a corporate innovation model and the necessity to have one. U.S. managers specifically noted that their organizations "needed" certain elements of corporate entrepreneurship and innovation. Seventy percent of U.S. managers agreed or strongly agreed that their organization needs an interconnected (anti-siloed) innovation structure (i.e., ecosystem, finance, marketing, operations, sales, prototyping, internal and external networking) to bring new products to market or to deploy solutions to achieve market leadership. The need for an interconnected innovation model is supported by managers representing organizations of all sizes (Table 1) and industries (Table 2). It increases as the number of employees in the organization grows beyond 100.

While there is an expressed management need for a corporate innovation system, there is an acknowledgment of the absence of key factors included in such a system that drives innovation. Specifically, only 57.9% of management agree or strongly agree that the firm's reward structure promotes innovation. Additionally, only 63% agree that the ecosystem, which consists of climate, environment, and work orientation is important to innovation in their company. Less than 66% indicate that it is easy for non-managers to introduce new ideas. Only 61% indicate that conflicting ideas are welcome in the organization and are well received, and less than 70% of the managers are encouraged to network and share new ideas that may lead to market leadership. Moreover, only 56% respond that the organization creates customer buy-in before launching new products or solutions.

TABLE I. Organizations that Strongly Agree for the Need for an Innovation Structure

Number of Employees	N	% Strongly Agree
2 – 10 employees	13	54%
11 – 50 employees	31	68%
51 – 100 employees	27	63%
101 – 500 employees	34	74%
501 – 1000 employees	41	73%
1001 – 5000 employees	27	81%
5000+ employees	43	72%
Total	216	

Note: The table above represents the response to the question, "Organization needs interconnected structure," based on the number of employees in the manager's organization

TABLE II. Industry Sectors that Strongly Agree for the Need of an Interconnected Innovation Structure

Industry	N	%Strongly Agree
Education	10	80%
Food Beverage	12	75%
Healthcare	19	63%
Information / Tech	14	71%
Manufacturing	17	82%
Retail / Wholesale	20	55%
Total	216	

Note: The table above represents the response to the question, "Organization needs interconnected structure," based on the industry sector of the manager's organization

A. Discussion

Three findings emerged which illustrate the adverse relationship in organizational innovation needs vs. operational deployment. Management overwhelmingly indicates that innovation is crucial to their firm's success, is essential to their organization's survival, is good for employees, is good for customers, and is good for stakeholders. The second finding is much less positive than the first finding, as management indicates that their company is not overly successful in deploying new innovations into the market. In the third finding, management supports a solution through an interconnected innovation structure to bring new products or solutions to achieve market leadership. Moreover, as the number of employees increases, the expressed need for an interconnected innovation structure also increases.

The rationale for the inverse relationship in organizational needs vs operational deployment is complex; however, management cites several possible explanations. They suggest that the lack of organizational innovation collaboration, unaligned reward system relative to the innovative proposals, insufficient time to consider new approaches, and lack of encouragement to share, test, and pilot new ideas, adversely affects innovation. As such, this may explain the rationale for the expressed management need for a formal innovation structure.

B. Argument in Favor for a Model for Corporate Innovation

Researchers have created models to increase the likelihood of desired outcomes. Consider that as early as 1985, Peter Drucker [6] noted that entrepreneurship was an intentional and systematic discipline. Peter Senge [7], and J. Richard Hackman [8] reinforced this noting that the organizational intentionality—structures—increase the likelihood that organizations can learn, adapt, and innovate Knezović and Drkić [9] also commented that specific determinants—psychological empowerment, decision-making process, and organizational processes—precede innovative work behaviors. While some organizations appear to have to some extent an internal innovation model and structure such as Amazon [10], there appears to be paucity in the availability of a universally applicable corporate entrepreneurship and innovation models that can be replicated by organizations in multiple industries and sizes.

Based on system thinking and with recognition of previous research and current managerial survey results, the need for a universal model of corporate entrepreneurship and innovation emerged. What follows is an argument suggesting the creation of The STAR Model for Corporate Entrepreneurship™ (STAR™), a replicable corporate innovation model that relies on predefined organizational structures that increase the likelihood of successful corporate entrepreneurship and innovation.

IV. STAR Model

STAR is a universal, repeatable, integrated strategic model for corporate innovation that empowers companies to create products, services and processes that enable them to dominate the industry in which they participate. STAR is a theoretical framework constructed on the foundation of four building blocks. Organizational structures (S) are the principles and practices that influence organizational outcomes. Think (T) is the process that empowers anyone in the organization to envision and propose bold new ideas that can have the potential to deliver market domination. Advocate (A) is a process that solicits support throughout the organization. Run (R) is a replicable process that provides the framework to go to market at the right time with the right resources. The acronym STAR provides a framework for managers, employees, consultants and academic researchers to conceptualize how to best enable development and execution of innovative practices that result in significant market penetration. The broad universal applicability of the STAR Model is that all tactical elements for market dominating innovation reside in one of the components of the model.

Building on the foundational structural and process focused thinking of thought leaders like Senge, Porter, Kotter, and Hackman, STAR increases likelihood of replicable innovation success. STAR incorporates major tenets necessary to enable successful innovation. This model begins by establishing a structure to foster innovation, followed by the creative proposal to create scalable solutions that address big problems that can propel the company beyond market leadership to market dominance. Throughout the process, advocates are sought out to provide guidance, input, and support to influence senior management approval to go to market at the right time with the right resources.

A. Application of STAR

Implementation of the STAR model produces extraordinary results! In 2019 Levi Conlow co-founded Letric e-Bikes with his childhood friend Robby Deziel. Under the auspices of Grand Canyon University’s Canyon Ventures Center for Innovation and Entrepreneurship, and using the tenets inherent to the STAR model. Mr. Conlow guided Letric e-Bikes to become a $250+ million company in just under three years and the largest and fastest-growing e-bike brand in the United States. In 2021 Forbes took notice of the company’s success and recognized Mr. Conlow as one of the nation’s top 30 under 30 innovators [11]. Lectric’s success can be attributed to excellent leadership and early implementation of the four components of STAR: Structures, Think, Advocate and Run.

The following sections provide additional details relative to each tenet in the STAR model. Corporate entrepreneurs working in institutions of all sizes may adopt the model as an architecture to achieve competitive advantages in the marketplace leading to market disruption and or domination.

V. Structures

Structures are principles and practices that influence organizational outcomes. They are a bias for action. Effective corporate entrepreneurs consciously build organizational structures that align reward systems and energize work into a dynamic innovation engine that seeks bold initiatives. As the foundation of the STAR model, Structures defines how work is done, who is responsible, and how information is shared within the organization. It is defined relative to the (a) creation of the organization's culture, (b) definition of the reward system, (c) creation of work structures, and (d) adoption of an information system.

A. Culture

Culture is the fabric that binds together what is acceptable and what is not acceptable. Culture are those things that we always do, never do, and that which we celebrate and correct. According to Schein [12], culture is present in artifacts, espoused values, and deeply held understandings. It shapes individual behavior through shared values, beliefs, and practices. It provides guideposts of what to do when explicit directions are not readily present. Culture is interwoven with the overall management system and is the unwritten rule book that works alongside the formal organizational structure. A culture that enables independent thought becomes the catalyst for creative thinking.

B. Rewards

Rewards, financial and non-monetary, reinforce and promote the creation of innovative thoughts and actions. Proposing new ideas, questioning old ideas, and exploring new technologies are risky and should be positively recognized. Employees who are rewarded for behaviors consistent with the organization's vision produce more predictable outcomes. If an organization wants empowered, creative, and innovative people, they reward it. As noted by Emilia Bratu [13] when writing about Lockheed's Skunk Works, "Reward performance, not status."

Moreover, when members of a team pivot from their original innovation, innovative organizations reward this. Consider the stories behind both the microwave oven and the Post-it note. Both were happy accidents, but Raytheon and 3M corporation, respectively, rewarded people for applying ideas in new ways. When the developer of a defense radar noticed his candy bar melted in his pocket, he developed the microwave oven. Similarly, when the inventor of a not-very-sticky glue used it on small notes to help him keep his place while singing in the church choir, he created the Post-it note. But, in both cases, there was someone within the organization that rewarded them for their creativity and encouraged them to run with their idea.

C. Work Structure

Work structures define core norms of conduct to balance control and chaos. Too much structure stifles creativity, and too little structure may diffuse outcomes. Work structures are aligned with the organization's culture and reward system. Organizations that find the balance between too much and too little structure unlock the potential of their teams [14]. Moreover, the work structures must reinforce the meaningful nature of the work, the value of the employee's personal contribution, and the opportunity and requirement to propose meaningful enhancements and innovations and receive feedback regarding the results of his or her work.

Work structures must be balanced with the end in mind. Before organizing work, leaders should ask, "What is non-negotiable, and where does the flexibility exist?" The organization of the work must allow for the next microwave oven or Post-it note invention while not detracting from the organization's vision.

D. Information Systems

Information Systems provide the organizational repository for innovative ideas, proposals, and projects, both successful and unsuccessful. The information system provides an objective measure of the level of innovation present within an organization. It offers the ultimate visible feedback loop providing insights into new opportunities. The system enables institutional visibility for anyone to view previously submitted ideas or to propose new ideas. Proposed ideas are automatically routed for review to assess whether the idea provides bold new opportunities or enhancements to an existing product. Ideas are presented to multiple departments within the organization for review. The use of information systems answers the question: What is the level of innovation in the company? Does innovation have the potential to deliver market dominance? What projects are in the process? What projects have been successful? What projects were not successful (and why)?

Information systems inform the organization where it is compared to where it wants to be. Tracking the submission of ideas can help organizations to foster a culture of innovation by demonstrating that the organization values and is committed to supporting new ideas and initiatives.

E. Application of Structures

At the very onset, Conlow through his management team at Lectric E-bikes recognized the need to establish structure to foster innovation and to enable it pervasively throughout the organization. He adopted the STAR model as described in the following statement:

We believe that if you have the right structure, you get the right outcomes. To this end, we push innovation down into every area of the company from design to customer service. At Letric innovative ideas are expected and rewarded. This structure empowers the entire company to focus on solving customer problems that are then incorporated into product design and the sales process.

VI. Think

Think represents the iterative process that empowers anyone in the organization to envision and propose bold new ideas that have the potential to deliver market domination. Building on the structures that precede the outcomes, the Think process rests on concepts well established in literature and industry, then integrates them into a clear model. However, in many cases, Think is viewed in isolation. Specifically, the Think process which is described below includes five stages.

A. Think Deeply

Think Deeply is an inquisitive, creative approach to uncovering important new sizable market opportunities. This is accomplished by critically questioning business, consumer, political, capital, and technology assumptions. Although business assumptions were once true, are they still true? Think deeply to challenge convention. Think to examine what is known. Think to discover what is not known. Think deeply to ask, "What if?" Successful organizations encourage thinking deeply about common things at all levels within the organization. Only one percent of the workforce is in top management. If thinking about innovation is exclusively limited to the top management, organizations will miss the ideas of 99% of their most vital resource—people.

Effective thinking encourages the conflict of ideas without allowing it to become a conflict between people. When ideas are allowed to clash, innovation follows. Good leaders protect the disruptors within their organizations because it is vital to challenge conventional thinking. Innovation comes from ideas, and ideas come from thoughts. Successful organizations encourage thought. Thinking organizations want the best ideas from 100% of their people, not only the top one percent.

B. Hypothesize and Prototype

The hypothesis is a proposal to create a new product or process that addresses a significant market opportunity. It is the natural outgrowth of Thinking Deeply. Hypotheses evolve into early prototyping [15]. At this early phase, a prototype developed within your team may be a presentation, a process flow document, a partial simulation including the market assessment (from the previous phase), or all combined. It does not need to be a fully functioning product or process. Prototypes should not be constrained by company directives and guidelines (except for legal) due to their potentially disruptive nature to the organization. Consequently, successful prototyping circumvents traditional barriers placed in organizations, which are adherence to the status quo operations, existing organizational structure, and the risk-averse nature of human beings.

In some cases, prototypes fail. However, creative organizations are willing to “Fail early to succeed sooner” [16] because early prototypes allow for discovery. Early prototypes, which can be revised over time and provide basic functionality to demonstrate the core idea of the product, are less costly and incur less risk than developing a fully functional product. Contrast this with traditional design models where considerable energy goes into planning, designing, and production. Early prototyping is a disciplined process that allows for issues to be encountered and solved early, thereby minimizing disruption upon entering the market. Addressing unforeseen production issues may be costly and, in some instances, cannot be overcome. Instead of order-to-chaos-to-order, early prototyping compresses the process by moving from chaos-to-order more quickly and efficiently [17]. The industry is replete with examples of this in the marketplace. Steve Jobs and Steve Wozniak’s Apple 1 computer in 1975 is only one example of an early prototype that eventually disrupted the market. Similarly, Lonnie Johnson prototyped the Super Soaker squirt gun, which now accounts for one billion in sales.

C. Investigate

Investigate expands the prototype review beyond the initial development team to include designated groups within the company. Avoid protecting the prototype—do not fall in love with the early version. Encourage your investigators to break and expose weaknesses in the design. Suspend emotions and solicit the user’s ideas. Effective innovators use feedback as the raw material for the next iteration. At this early phase, a prototype developed within your team may be an updated presentation, a refined process flow document, a partial simulation, including the market assessment (from the previous phase), or all combined.

Groups outside your department team__may provide truthful, unbiased input. Soliciting input beyond the development team may increase risk if negative feedback is shared throughout the organization. However, positive feedback helps garner support from advocates who may back the prototype towards company adoption. Investigate answers to the questions: What is good? What works? Does it fix a problem? What does not work? Does it satisfy the identified market opportunity? How can it be improved?

D. Network

In contrast to the Investigate stage, Network expands the initial prototype review to include select customers to solve a specific problem. Corporate entrepreneurs must clearly define and blueprint their solution, then secure an intent to purchase before it is officially built. After examining the wreckage of their failures, many innovators trace their ruin back to this step. They spent valuable time and money building a solution for customers or problems that did not exist. Successful innovators must constantly collaborate with customers to define each of the following:

- The actual problem
- A viable solution
- The price range or budget for the solution
- A comprehensive list of decision-makers
- The timing of delivery, and
- A clear statement of which features are necessary

Gathering feedback from real users helps identify issues or areas of improvement that company employees, from a different perspective, may fail to identify. Soliciting input from customers can be assessed as a higher risk. Receiving and incorporating their feedback early in this stage will make the Advocate dimension of the STAR Model easier. In addition, it is vital in the iterative process of hypothesizing and prototyping, investigating, and networking. Like the Investigate phase, in Network stage answers the questions: What is good? What works? Does it fix a problem? What does it not work? How can it be improved?

E. Kreate

Kreate completes the prototype process and solicits company-wide support for its adoption and go-to-market strategy. The development team may be ready to seek approval to proceed or may need additional work. They may need to acquire the support of an expanded set of Advocates or may need to modify the minimum value proposition (MVP). The Kreate stage answers the questions: What is the value of the innovation? When would you need to enable this innovation? How will this innovation be integrated? What is the anticipated disruption associated with integration? How can this innovation be scaled? The MVP sets the stage for the Advocate dimension of the STAR model.

Once the innovation has been sold internally, corporate entrepreneurs must continue the work of external selling. Corporate entrepreneurs must be in constant communication with their “customer evangelist group”. This customer evangelist group should be representative of the total available market (TAM) you intend to capture. These will be the early adopters, think of them as the equivalent to the Key Advocates discussed in the Advocate section. All innovations are created for specific customers to solve specific problems. Corporate entrepreneurs must clearly define and blueprint their solution, then secure an intent to purchase before the solution is officially built. After examining the wreckage of their failures, many entrepreneurs trace their ruin back to this step; they spent valuable time and money building a solution for customers or problems that did not exist. Successful corporate entrepreneurs must constantly collaborate with customers to define each of the following:

- The actual problem
- A viable solution
- Price range or budget for the solution
- All decision-makers
- Timing of delivery, and
- Exactly which features are necessary

All this is accomplished with a tool often referred to as a requirements document. Each section of this nonbinding document is completed before anything is built. Throughout the process information is constantly verified and refined realizing that it is difficult to succinctly define the problem. Once the document is fully completed then executed by the prospective customer, it's then used to build the solution.

1) Application of think. Mr. Conlow and his team from Letric eBikes implemented the Think process to overcome their initial product failure. Conlow enabled a process to create scalable products and solutions that address big problems that propel the company beyond market leadership to market dominance. Using available research data, it was clear to Conlow that although European customers were aggressively purchasing ebikes, the US market was still in its infancy. Existing bicycle companies with dominant market positions and years of experience had not yet embraced the new technology, and as a result, the US market was wide open.

Lectric's first e-bike design was a total failure. However, Conlow knew from the data that the potential US e-bike market was massive. "The problem was not the market; it was our bike. The solution was simple, listen to the customer and re-work the prototype." Letric learned that their customers did not want to spend another $500 to $1500 on a bike rack to transport their new e-bike. The solution was to build a folding ebike that did not require a bike rack. The reduced price point of an ebike without the bike rack was an incredible innovation that was eagerly accepted by multiple market segments and was difficult to replicate by competitors. Additionally, the folding ebike shipped fully assembled in a box that was smaller than the full-size crates used to ship full-size ebikes. The smaller box made is less expensive to ship and minimized the potential for damage during shipment. The company utilized an organic marketing strategy to launch their newly designed ebike. The campaign deliver $4 million in pre-orders within the first 30 days.

VII. Advocate Process

Advocate is the process of gaining approval for new innovation by securing the active support of individuals both inside and outside of the organization. The Advocate process works in concert with Think by securing individual support across the enterprise and its key customers.

Developing advocates to support a new innovation requires that one understands the neuroscience of risk. Knowing that everyone in your organization comes with their own level of risk tolerance is essential to successfully introducing an innovation. The findings of a study by Mueller et al. [18] support the notion of risk tolerance. A Cornell University publication summarized their findings "Why we crave creativity but reject creative ideas," [19]. The summary noted the following four reasons why novel ideas are often scorned.

- Creative ideas are, by definition, novel, and novelty can trigger feelings of uncertainty that make most people uncomfortable.
- People dismiss creative ideas in favor of purely practical ideas.
- Objective evidence shoring up the validity of a creative proposal does not motivate people to accept it.
- Anti-creativity bias is so subtle that people are unaware of it, which can interfere with their ability to recognize a creative idea.

A. Risk

To offset the natural aversion to accepting new ideas, innovations must be introduced as being of exceptionally low risk, intriguing, and viable if they are to gain public support. Incorporated into STAR, a three-part method of language, reports, and stages can help de-risk any innovation by building trust, thus securing buy-in from others.

1) Language. Language is important. Corporate entrepreneurs should approach potential key advocates and other stakeholders using specific low-risk language. New ideas should be presented with a combination of terms that convey safety and appeal. This combination replaces fear with curiosity. Corporate entrepreneurs may introduce a concept by saying something like,"We've been examining a growing market trend that seems to be going unnoticed by our competitors. We've crafted a test that, if successful, could become a very profitable new revenue source, and we could own this market." Notice the language, "test" is a low-risk term. The word "unnoticed" creates intrigue and introduces the idea of timing. The term "we've" implies that there is a group and there is safety in numbers. "Profitable new revenue source" helps to replace fear with curiosity and appeal.

2) Reports. Progress reports should be delivered to key stakeholders. In-person is the preferred method, as visual evaluation of body language and non-verbal cues provides invaluable insight to know if your conversion strategy is working. Updates should include both problem verification and prototype feedback. While email and video conferencing are acceptable, they do not allow for this type of surveillance. These reports follow a simple format that presents test results with a lucid infographic and no confusing information. Updates should be crisp and direct, long updates can become confusing and derail the innovation. A pithy report builds trust and confidence, anything complicated or confusing can be threatening.

Securing the support of advocates is not done all at once; it is enabled carefully in specific stage. It begins with the innovator securing the support of staff within the department and then expands beyond the department to the entire company and eventually includes select customers. This progression is illustrated in the following five stages:

3) Stage 1: Identify Initial Advocates. By its very nature, innovative proposals of any type are often viewed cautiously. The prospect of something new that potentially may change the existing organizational structure, product mix, company strategy, underlying technology, operational processes, employee reward system, or even staffing levels is a red flag for many employees. For an innovation idea to survive beyond its infancy, the innovator must survive!

Carefully selecting individuals to support a new innovation within an existing organization is of paramount importance to minimize innovator risk. By choosing the right team members, organizations can mitigate potential risks and maximize the chances of successful implementation. Firstly, the selected individuals should possess a diverse set of skills and expertise relevant to the innovation's domain, allowing them to tackle various aspects of the project effectively. Secondly, a well-balanced team that combines both seasoned experts and fresh minds can offer a mix of experience and creativity, leading to innovative solutions while avoiding tunnel vision. Thirdly, individuals with a proven track record of adaptability and openness to change are more likely to embrace the inherent risks associated with innovation, fostering a culture of resilience. Lastly, aligning the team's values and commitment to the organization's mission ensures a shared vision, boosting motivation and dedication to overcoming challenges. Careful selection of team members can enhance the innovation process, diminish the burden on individual innovators, and significantly reduce the overall risk, leading to the organization's long-term growth and success.

4) Stage 2: Build the Network. Once the initial advocates (who provide minimal risk) are identified, the innovator starts expanding the network of potential supporters who have a vested interest in the innovation and who can help spread the word about it. This includes a series of key advocates who have the power and position to enable the innovation to move forward. A key advocate is a respected and trusted leader in your organization, one who curries favor with other leaders and the rank and file. Key advocates mitigate risk in the minds of others. Building trust with at least one key advocate is essential to converting supporters and critics into public champions.

In addition, the advocate base must include customers, employees, influencers, industry experts, and other stakeholders. A thoughtful expanded network approach maps inherent risk of soliciting advice for a new innovation based on potential advocates key attributes: positional power, feedback, influence, knowledge, risk aversion, and status. Each attribute plays a vital role in determining the quality and reliability of advice received. Positional power ensures that the advice comes from individuals with the necessary authority and experience to make informed decisions. Feedback provides valuable insights from various perspectives, enhancing the chances of identifying potential pitfalls and opportunities. Influence signifies the potential impact of the advice on the innovation's trajectory, making it crucial to gauge the credibility of the sources. Knowledgeable advisors possess expertise that can significantly improve the innovation's outcome. Risk aversion is crucial to consider, as overly cautious advice might hinder growth, while recklessness could lead to avoidable failures. Finally, understanding the status of potential advocates helps identify biases that might influence their suggestions. By thoughtfully mapping these risk factors, innovators can solicit meaningful information and support to propel innovations towards success.

5) Stage 3: Develop Your Messaging. To create a powerful network of advocates, a clear and compelling message about your innovation is needed. Be specific, emphasize how the innovation provides unparalleled opportunities for the organization to dominate the sector in which it operates.

Targeting your message is of paramount importance when soliciting support for an innovation that may significantly alter the organization’s current environment. The message influences the success and reception of your idea both internally and with prospective customers. Tailoring your message to a specific audience ensures that the innovation’s unique features, benefits, and value proposition are effectively communicated to customers while concurrently communicating how the innovation will enhance the organization market position and brand. By understanding the needs, preferences, and pain points of the target audience, you can craft a message that resonates with them on a personal level, increasing the likelihood of capturing their attention and generating interest. A well-targeted message also aids in establishing a strong brand identity and positioning in the competitive landscape, fostering customer loyalty and advocacy. Moreover, it enables you to focus marketing efforts and resources efficiently, optimizing outreach and maximizing the return on investment. In essence, effective message targeting plays a pivotal role in not only driving initial sales but also fostering long-term relationships with customers, ultimately leading to sustained growth and success for the new product in the market.

6) Stage 4: Engage, Gather Advice and Refine the Offering. Once the innovator(s) has built the network of advocates, it is essential to engage with them regularly. Engage, listen, and refine. Listen to learn. Learn what may not have been known or seen before engaging. Specifically ask for advice; research has shown this to be more effective than requesting feedback [20]. When seeking advice, an innovator solicits the expertise and knowledge from potential advocates gaining insights and suggestions with relevant experience in the industry or domain. The act of seeking advice is constructive as it provides a personal connection between the innovator and the potential advocate. On the other hand, the act of seeking feedback does not result in enhancing the innovator relationship with potential advocates and may result in vague commentary that does not suggest how to improve the innovation.

Additionally, provide potential advocates with the resources, information, and tools they need to spread the word and build momentum. Be mindful to solicit input yourself; do not rely solely on others for their input. Allow advocates to be part of the team and value their opinions and insights. Time spent engaging your advocates transforms them from casual participants to enthusiastic supporters.

7) Stage 5: Monitor Progress and Adjust Strategies. Monitor the progress of your network of advocates and adjust your strategies as needed. Use data to track engagement and optimize your outreach efforts. A decision must be made to go forward decisively, to abandon the innovation, or to start again. A "go" decision moves into the Run stage of STAR. A "no-go" or "start again" decision leads to a reflective debriefing to leverage lessons learned for the next cycle of innovation using the iterative, perpetual process that is the STAR Model. The "Go" decision requires the consent of the final decision maker(s) for formal approval to move to Run.

B. *Advocates in Action*

Mr. Conlow and his Letric team made expert use of advocates and the advocate process to refine their product, identify their ideal customer profile (ICP), and market to their ideal customers. Mr. Conlow and his marketing team identified three different types of key advocates. These included e-bike experts, technical experts, and their prime customer influencer. Once identified, the Letric team expanded this network of key advocates. This network assisted Letric in developing its messaging, for example, Letric's prime customer are recreational vehicle owners or RVer's, the company engaged a social media influencer with thousands of these subscribers to do a product review of their new e-bike. Letric learned from this influencer, and from talking with their other advocates, prospects, and customers, that their messaging should showcase their new e-bike's folding capability. A folding e-bike precluded the extra expense of a bike rack. RV owners viewed Cetric's e-bike as space and expense saver, and the ideal alternative to hauling a car. Letric's expert use of the advocates and the Advocate process as described in the STAR model empowered the company to create and successfully launch their new e-bike and rapidly drive organic sales to over 150,000 units in their first 30 months.

VIII. Run: The Tactical Application of the STAR Model

Run represents a replicable iterative process, meticulously designed to provide a robust framework for venturing into the market at the right time with precisely calibrated resources. In the dynamic landscape of corporate innovation, Run stands as the tactical embodiment of the STAR model; Structure, Think, Advocate, and Run. This model is not a linear procession; rather, its components continuously evolve, interweave, refine, are rigorously tested, and continually evolve and improve in response to the ever-shifting business landscape.

A. *Run: A Replicable Iterative Process*

Run's foundation rests on the steadfast understanding that markets and the conditions for execution are in a constant state of flux. The notion that any go-to-market plan can be perfect is debunked, for such assumptions can lead to disastrous outcomes. Unlike linear processes, Run is an ongoing journey with no fixed conclusion point. With each iteration, it systematically revisits the operational plan, fostering the nimbleness to address unforeseen obstacles and those entirely novel in nature.

B. *The Ongoing Vitality of Structures*

While Structures are designed earlier in the STAR process, their persistent relevance in Run cannot be overstated. These structural underpinnings encompass the fundamental principles and practices that exert a profound influence on organizational outcomes. A meticulously devised innovation reward system, technology deployment to assess and measure innovations, precise staffing levels, selection of innovation staff, strategic aligned marketing and investor communication plans, and unwavering financial support all contribute to the organization's readiness. This readiness transcends the confines of a single innovation; it pertains to the seamless execution of a continuous stream of innovations. A particular emphasis is placed on fostering an open culture that encourages input from all corners of the organization – an invaluable asset during the dynamic Run phase, where the unexpected demands immediate attention.

C. *From Think to Run: Requirements Revalidation*

The requirements document, initially crafted during the Think phase, undergoes a rigorous process of revalidation in Run This document, conceived months prior to the onset of Run, offers a preliminary definition of the problem, a theoretically viable solution, price range parameters, an ideal delivery timeframe, and delineation of the essential product features. In the ever-fluctuating landscape, Run continually recalibrates this ideal solution, ensuring its alignment with the current environment. Specifications are subjected to relentless assessment to guarantee that the innovation not only provides value but also forges a path to market leadership while remaining eminently achievable. The assessment ambitiously extends to encompass the innovation's producibility, pricing, marketing readiness, distribution strategy, and the organization's preparedness to scale the innovation.

D. *Embracing Change and the Quest for Market Leadership*

Human nature inclines towards the illusion of constancy but Run reminds us of the inexorable nature of change. Failing to acknowledge this reality can be perilous, and the belief in an infallible go-to-market plan can prove catastrophic. It is imperative to remain vigilant, continuously scanning the environment for shifts, changes in internal and external advocate support, the sustained presence of organizational financial support, and the competitive landscape. Furthermore, for those resolute in their pursuit of market leadership, the imperative of scaling innovations takes center stage. This endeavor necessitates considerable foresight, unwavering commitment, and strategic investments.

E. Application Market Dynamics Drive Lectric's e-Bikes Success

The triumph of Lectric e-bikes, a prominent player in the electric transportation market, serves as a compelling case study that aligns seamlessly with the principles of the STAR model, with a notable emphasis on scalability. For organizations determined to be a market leader, their innovations must rapidly scale. This requires foresight, commitment and investment as described by Levi Conlow, CEO and Co-founder of Lectric e-Bikes. “At the end of 2020 and early 2021, we began to scale our operations by making big investments and adjustments for how we bill, manufacture, distribute, and warehouse. That investment is starting to pay dividends now.” offered Conlow. Consistent with the STAR model, scalability was paramount if Lectric’s new e-bike was to establish market dominance. Their timing could not have been better, given the spike in U.S. demand for e-bikes. The Covid-19 pandemic triggered a surge in bicycle sales. Starting in July 20211, the twelve-month sales increased for two-wheelers by 65% to $5.3 billion, according to analyst Dirk Sorenson with market researcher NPD Group. “In the past two years, e-bikes grew by a whopping 240%, which made it the third-largest cycling category in terms of sales revenue behind mountain bikes and children’s bikes and ahead of road bikes”, Sorenson said in a recent report [11]. At the close of 2022, Letric was second only to Tesla in the total number of electric transportation units sold and is poised to pass Tesla in 2024. “The team at Lectric accomplished this through a novel approach of design, marketing, distribution, and customer support, which has earned it thousands of highly satisfied, loyal customers” [11]. Bertram Capital partner Ryan Craig said at the time of the company’s VC’s funding announcement.

Lectric’s forward-thinking approach yielded substantial dividends, in perfect harmony with the STAR model's ethos. Lectric’s timing aligned with the surging U.S. demand for e-bikes amid the COVID-19 pandemic, underscored the model's efficacy. Run, as the tactical application of the STAR model, epitomizes a dynamic, adaptive approach to corporate innovation. It underscores the importance of continuous assessment, embraces change as the one constant, and champions scalability.

In summary, Run, as the tactical application of the STAR model, embodies adaptability, continuous assessment, and scalability. It champions an iterative approach where Structure, Think, Advocate, and Run are interwoven and constantly refined. The success of companies like Lectric e-bikes underscores the efficacy of this model in navigating the ever-evolving landscape of corporate innovation.

IX. Getting Started: Assess Innovation in an Existing Organization

To adopt the STAR model, it is essential that senior leadership first assess the organization’s innovation actions and the associated results. Then, this must be differentiated from management and staff perception regarding innovation in the organization. This can be accomplished in two steps. In the first step, senior leadership should ask three questions:

1) How many new innovations or products were announced by your organization over the last five years? (List only)

2) To what extent did these products significantly grow revenue, market penetration, operating margins, or increase operational efficiency? (from the list, indicate the outcome)

3) Did the new innovations help to create an unfair advantage that will make it difficult for other organizations to replicate? (From the list, Yes, No response only)

In the second step, senior leadership should formally assess corporate innovation readiness by administering the STAR survey. Executing the survey will help leadership determine the extent in which management and non-management believe that innovation is crucial for organizational survival, whether there is support to achieve this, and whether they believe their innovation has been successful. Combining the organization’s perception of innovation with the actual results provides the foundation for senior leadership to begin the strategic transformation of the organization to one that continually enables, enhances, and rewards innovation not just by management but is endorsed throughout the organization.

X. Summary

The absence of a corporate innovation model can be a significant obstacle for companies seeking to drive innovation within their organization. Without a defined approach, companies may struggle to differentiate between regular product enhancement and bold new initiatives and subsequently lose the option to pursue bold market leadership opportunities. Without an integrated model to assess new ideas, prototype a promising idea, solicit internal and external feedback, consider scalability, and develop internal and external advocates, the potential to achieve sustained market leadership is problematic. Therefore, to address this need which is supported by a survey of 200 business managers, the authors propose adopting the STAR model, a teachable, replicable of innovation model. The model includes four tenets:

1) Structures: principles and practices that influence organizational outcomes

2) Think: the process that empowers anyone in the organization to envision and propose bold new ideas that can have the potential to deliver market domination.

3) Advocate: process to gain approval for a new innovation by securing the active support of individuals both inside and outside of the organization.

4) Run: a replicable iterative process that provides the framework to go to market at the right time with the right resources.

The current focus in innovation research is primarily focused on new product creation through the vantage point of start-up companies. Agile development and creative thinking models do exist, but these models do not consider staffing, compensation, technology, leadership authority, culture, risk aversion, and lack of innovation support that many existing organizations encounter that start-up companies do not have. New technologies such as AI provide incredible opportunities for competitive advantage in existing organizations. However, technology alone does not create competitive advantage. Skillful, rapid, cost-effective deployment supported by advocates within the organization and with select customers provides competitive advantage. Although adoption of an innovation model does not guarantee that the organization can outperform its competitors and of course new start-ups, it does improve the odds.

REFERENCES

[1] Ries, E. (2011). The lean startup. How today's entrepreneurs use continuous innovation to create radically successful businesses. Crown Business.

[2] Ries, E. (2013). The lean entrepreneur: How to apply lean principles to your startup or corporate innovation. Wiley.

[3] Ries, E. (2015). The leaders' guide: How to use the startup to transform your company and your career.

[4] Furr, N., & O'Keefe, K. (2023) The hybrid start-up. Harvard Business Review March-April 2023.

[5] Rayport, J., & Sola, D. (2023). Entrepreneurship: The overlooked key to a successful scale-up. Harvard Business Review, March-April

[6] Drucker, P. F. (1985). Innovation and entrepreneurship. Harper Business.

[7] Senge, P.M. (1990). The fifth discipline: The art and practice of the learning organization. Doubleday/Currency.

[8] Hackman, J. R. (2002). Leading teams. Harvard Business School Press.

[9] Knezović, E., & Drkić, A. (2020). Innovative work behavior in SMEs: the role of transformational leadership. Employee Relations 43(2), 398-415. https://doi-org.lopes.idm.oclc.org/10.1108/ER-03-2020-0124.

[10] Rikap, C. (2022). Amazon: A Story of accumulation through intellectual rentiership and predation. Competition & Change, 26(3-4), 436-466.

[11] Ohnsmann, A,. (2021). Powering their way into the e-bike boom. Forbes. https://www.forbes.com/sites/alanohnsman/2021/12/01/powering-their-way-into-the-e-bike-boom/?sh=6d7653d2d3db

[12] Schein, E. H. (1999). The corporate culture survival guide. John Wiley & Sons, Inc.

[13] Bratu, E. (2020). The incredible story of skunk works or how to create high-speed projects. [Post]. LinkedIn. https://www.linkedin.com/pulse/incredible-story-skunk-works-how-create-high-speed-projects-bratu/

[14] Hackman, J. R. (2002). Leading teams. Harvard Business School Press.

[15] Rubinstein, M. F., & Firstenberg, I. R. (1999). The minding organization: Bring the future to the present and turn creative ideas into business solutions. John Wiley & Sons.

[16] Kelley, T. (2001). The art of innovation: Lessons in creativity from IDEAO, America's leading design firm. Currency Books.

[17] Rubinstein, M. F., & Firstenberg, I. R. (1999). The minding organization: Bring the future to the present and turn creative ideas into business Solutions. John Wiley & Sons.

[18] Mueller, J. S., Melwani, S., & Goncalo, J. A. (2012). The bias against creativity: Why people desire but reject creative Ideas. Psychological Science, 23(1), 13-17. https://doi.org/10.1177/0956797611421018

[19] Cornell University. "Why we crave creativity but reject creative ideas." ScienceDaily. ScienceDaily, 5 September 2011. www.sciencedaily.com/releases/2011/09/110903142411.htm.

[20] Yoon, J., Blunden, H., Kristal, A., & Whillans, A. (2019, September 20). Why asking for advice is more effective than asking for feedback. Harvard Business Review. https://hbr.org/2019/09/why-asking-for-advice-is-more-effective-than-asking-for-feedback.

Control-Driven Media: A Unifying Model for Consistent, Cross-platform Multimedia Experiences

Ingar M. Arntzen[1], Njal T. Borch[2], Anders Andersen[3]
NORCE Norwegian Research Centre, Tromso, Norway[1] Schibsted, Tromso, Norway[2]
UiT The Arctic University of Norway, Tromso, Norway[3]

***Abstract*—Many media providers offer complementary products on different platforms to target a diverse consumer base. Online sports coverage, for instance, may include professionally produced audio and video channels, as well as Web pages and native apps offering live statistics, maps, data visualizations, social commentary and more. Many consumers also engage in parallel usage, setting up streaming products and interactive interfaces on available screens, laptops and handheld devices. This ability to combine products holds great promise, yet, with no coordination, cross-platform user experiences often appear inconsistent and disconnected. We present *Control-driven Media (CdM)*, an extension of the current media model that adds support for coordination and consistency across interfaces, devices, products, and platforms while remaining compatible with existing services, technologies, and workflows. CdM promotes online media control as an independent resource type in multimedia systems. With control as a driving force, CdM offers a highly flexible model, opening up for further innovations in automation, personalization, multi-device support, collaboration and time-driven visualization. Furthermore, CdM bridges the gap between continuous media and Web/native apps, allowing the combined powers of these platforms to be seamlessly exploited as parts of a single, consistent user experience. Extensive research in time-dependent, multi-device, data-driven media experiences supports CdM. In particular, CdM requires a generic and flexible concept for online, timeline-consistent media control, for which a candidate solution (State Trajectory) has recently been published. This paper makes the case for CdM, bringing the significant potential of this model to the attention of research and industry.**

Keywords*—*Multi-platform; media control; continuous media; data-driven media; interactive media; orchestrated media

I. Introduction

The landscape of media production and *over-the-top (OTT)* delivery holds immense potential. Advanced production tools, automated AI-based technologies, and a wealth of data sources come together, often in multi-step, distributed production chains. On the client-side, highly capable consumer devices offer further opportunities for adaptation, customization, interactivity and data-driven graphics. Concurrently, consumer preferences are evolving; ranging from those who favor the traditional one-size-fits-all broadcast experience, to those seeking multi-device immersion, interactive engagement, customized multi-device setups, personalized narratives, social interactivity, or advanced accessibility features. This diversification poses a significant challenge for media providers aiming to offer advanced, high-quality user experiences to a sizable audience, while simultaneously managing costs and complexity.

Many media providers offer alternative products on different platforms to address needs for richer and more varied experiences. For instance, coverage of major sports events may include live-produced video channels, VoD services, and Web or native apps with support for live feeds, interactive data visualization, and social integration. Many viewers also see these as complimentary offerings and engage in combined usage to further enrich their experiences. This, though, may be less rewarding than perhaps anticipated. Each product must be configured separately, and there is typically no coordination across platforms. Differences in production delays may also lead to inconsistencies, confusion and spoilers. This significantly limits the value of combined usage, possibly driving viewers back to traditional single-product coverage.

We envision a new class of cross-platform media experiences, where independent products and interfaces may be flexibly combined and coordinated to form a single, consistent, user experience. Users can then enjoy a spectrum of experiences, from lean-back entertainment to lean-forward engagement, from a single-device to multi-platform immersion, or from private to social experiences. Furthermore, by opening up for cross-platform usage, media providers can provide advanced and uniquely adapted user experiences cost-effectively by leveraging the combined powers of existing infrastructure and services, while also ensuring quality and brand control for the user experience as a whole.

Unfortunately, issues with cross-platform user experiences are not easily addressed within the current model. Different products are built from independent technology stacks and define entirely separate user experiences. While some solutions exist for coordination and consistency, they are often application-specific or limited to specific technology options, such as data formats, distribution protocols or presentation frameworks. In contrast, we argue that cross-platform coordination must be independent of platform-specific solutions, and that support for cross-platform media experiences should instead be addressed as a fundamental feature of the media model.

To this end, we propose *Control-driven Media (CdM)*, an extension of the current media model with built-in support for consistent, cross-platform user experiences. CdM promotes *control* as a principal resource type in media systems. Through online control sharing, applications can orchestrate connected interfaces and render data sources and media content consistently across platforms. Moreover, as CdM is compatible with existing infrastructure and workflows, adoption can be incremental and cost-efficient. Furthermore, CdM uniquely targets a generic solution for cross-platform consistency, by addressing fundamental limitations of the current media model.

This paper presents CdM and a range of opportunities implied by the model. Section II presents current approaches to cross-platform media experiences, and remaining challenges, leading up to the problem statement in Section III. Section IV outlines the current media model, before the main contribution, the CdM model, is introduced in Section V. Section VI discusses key opportunities of CdM within the context of online sports coverage. Section VII covers key technical challenges. Section VIII includes evaluation and references to implementation and supporting research. Section IX provides a brief discussion before the paper is concluded in Section X.

II. Background

The online world provides unprecedented opportunities for online media, with a wealth of tools and platforms for capturing, editing, distributing, and rendering of data and media content. Moreover, the ongoing AI revolution promises further opportunities for automation, content generation, personalization, and more. To fully exploit this potential, media systems must meet an increasingly complex set of demands. Media providers seek to build their brands through high-quality user experiences and exciting narratives. There might also be a race for advanced features, like impressive graphics, interactive engagement, device adaptation, social integration, and/or multi-device support. This, though, must be balanced with technical and financial considerations, including performance, scalability, complexity and costs. Users, on the other hand, seek media experiences that are relevant, exciting and distraction-free. Beyond this, users are diverse, and preferences will likely diversify further as offerings become more advanced. For example, some users prefer lean-back storytelling, while others want to engage and actively shape the experience for themselves or others. Some prefer generic broadcast coverage, while others prefer a more personalized narrative. Some want experiences to be social, some seek multi-device immersion, and some have specific accessibility requirements.

Supporting diversity is hard, though, and certain demands may even appear conflicting in terms of technology options. In particular, we recognize two technical approaches for audiovisual user experiences; *continuous media* and *data-driven media*. Media production based on continuous media types (e.g. audio and video) supports high-quality storytelling through precise mixing and scheduling of video sources, audio tracks, and graphical elements. This approach can provide user-friendly, lean-back, and action-packed user experiences, yet primarily the same experience for all viewers. In contrast, data-driven media (e.g. Web/native applications) offer dynamic experiences with flexible options for interactivity, visualization, adaptation, personalization and collaboration. However, data-driven media provide limited support for weaving complex time-dependent, lean-back narratives, particularly if they involve multiple data sources and/or different rendering technologies.

To address complex demands from users and media providers, it appears necessary to leverage the combined power of these approaches. This section presents common ways of combining technologies, highlighting the strengths and limitations of each approach.

1) Embedded media: A common way to combine technologies is to embed continuous media within a data-driven interface. For example, audio and video content on the Web platform may be embedded in Web pages using the *HTML5 Media element* [1]. This way, soccer coverage may include both a video stream and a feed of match events. However, despite being part of the same layout, embedded components technically define separate user experiences. For instance, the video pause button might not apply to the match events feed, which will continue to report game developments, even if the video stream is paused. This creates inconsistency, where the two components display data from the same event but from different time-frames. Media providers may address this by introducing additional coordination between components. For instance, on the Web platform, media players and feeds may be controlled by code. This, though, often leads to custom solutions within specific applications and platforms, and the complexity increases with the number of coordinated components.

2) Overlays: A related method is to layer a transparent, data-driven interface on top of a video display. This technique can for instance be used in video production, to burn graphics onto a video stream. Alternatively, graphics can be overlaid in the user interface, using a *z-index* or similar layering concepts supported by the layout system. This provides additional options for individualized graphics and interactivity, but requires that overlay graphics are synchronized with video progression. This can be accomplished by defining cues, hooks, or events on the video timeline. In the Web platform, this is supported by the concept of data tracks [1] integrated with the HTML5 media element. While this provides coordination between continuous and data-driven media, the approach is most useful when the experience is limited to a single video asset within a single interface or layout.

3) HbbTV: HbbTV (Hybrid Broadcast Broadband TV) [2] is a standards initiative and platform allowing HTML-based graphics to be overlaid on broadcast video content. In HbbTV, the focus is on the set-top box or SmartTV, which supports IP-based data access as well as traditional broadcast distribution. By exposing the media clock of the broadcast stream to the HTML5 processing environment, program-specific or even personalized overlays may be presented on the TV display. Broadcasters may define and deploy such overlays as HTML applications associated with a channel or program. Furthermore, the concept of overlays is also extended to companion devices. This means that smartphones may connect to the HbbTV device to access additional contents or interactive capabilities, synchronized with the broadcast clock [3]. However, the adoption of HbbTV has been slow, as it requires developing and deploying HbbTV-enabled consumer devices. Moreover, as the approach hinges on the existence of a physical device, it does not easily generalize to other media domains or beyond the home environment.

4) Low latency streaming: Another way to combine media content and data-driven graphics, is to coordinate the delivery of multiple data streams. For example, an online lottery might want to stream a video from the lottery drawing, and also provide a separate data stream with the winning numbers, for visualization. This requires some coordination between streams, or else the graphics might spoil the suspense of the video, or even make the lottery appear suspect. One approach is to minimize latency in streaming, thereby also limiting the potential misalignment of streams. Global CDN providers such as Akamai [4], Cloudflare [5] and Fastly [6] provide scalable, low-latency streaming solutions based on Dash [7], HLS [8] or WebSocket [9] protocols. Another option is to leverage support for synchronized stream delivery or fixed end-to-end delays. For instance, in IP-based networks, a fixed end-to-end delay may be achieved by transmitting data to a client ahead of time and then scheduling data delivery to the application in reference to a media clock. However, such methods have some limitations. Due to network jitter, ultra-low-latency streaming may be expensive and provide a reduced user experience. Moreover, synchronizing content using the distribution system increases the complexity of the distribution system itself, which goes against current practices of stateless servers to limit cost and improve scalability. In any case, if consistency is achieved through a specific distribution mechanism, it is also limited to that mechanism.

5) Object-based media: In Object-based Media (ObM) [10]–[12], the challenge of combining technologies is already addressed in production. ObM targets increased support for adaptation, personalization and interactivity in broadcasted experiences. The key idea is to represent media as objects, and let client devices be responsible for assembly into rendered presentations. This way, clients may assemble media experiences differently, be sensitive to device capabilities, local context or user preferences, and leverage data-driven technologies for visualization and interactivity. However, the approach requires changes to media formats, with implications for existing workflows and tools. This is speculated to be a key reason for slow adoption [13]. In addition, ObM is still bound to the single-stream broadcast metaphor. This means that objects are packaged and distributed as one asset, even if each viewer will only make use of a subset of objects. Moreover, the approach does not provide any particular support for multi-device media consumption.

6) Leave it to the user: The last approach, and perhaps the standard solution, is to leave the combination of technologies as an exercise for the viewer. *Formula 1 (F1)* online coverage provides an example of this. *F1TV* [14] offers a number of video streams to choose from, including the professionally produced World Feed, onboard vehicle cameras with team radio for that driver, a produced pit lane feed and even a channel with lap times in tabular form. In addition, the *F1 App* [15] provides an interactive race map and numerous data visualizations for detailed race statistics, which also supports manual time shifting. Viewers may then select the most relevant video feeds and visualizations and use multiple devices to follow multiple interfaces in parallel. As such, this approach supports high levels of customization and personalization. At the same time, to realize this potential, viewers must do all the work. For example, synchronization between interfaces is crucial in a fast-paced sport like F1, yet cumbersome to achieve manually. Moreover, users can not easily know which camera feed is most relevant at any time, or when to switch between views. In short, this approach encourages viewers to act as producers, and leaves media providers with little control over the quality of user experiences.

A. Challenges

Combinations of continuous media and data-driven applications have been attempted in different parts of the technology stack, in production, in distribution, and in presentation, yet each method comes with limitations. Still, combining technology platforms remains an attractive prospect. Online F1 coverage, for instance, showcases a significant potential if only some coordination could be provided between products.

In current systems though, there is limited support for such coordination. Fig. 1 *(left)* illustrates four product interfaces (green rectangles), each defining its own user experience (bubble). This leaves the viewer to mediate between separate user experiences, either compensating mentally for any inconsistencies between them, or manually attempting to correct them, for instance by time-shifting or configuring individual interfaces. Neither option is particularly appealing.

Fig. 1 *(right)* illustrates the basic idea of cross-platform media experiences. Here separate product interfaces (green rectangles) contribute to a single, consistent user experience (bubble). Consistency implies (i) that interfaces operate in reference to a shared media clock, and (ii) that user interactivity is not limited to one product, but applies to the experience as a whole. For instance, if the user selects a different F1 driver, this might potentially affect all interfaces in different ways, including overlayed video graphics in F1TV and the race map visualized by the F1 App.

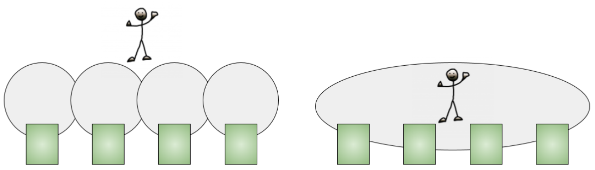


Fig. 1. *(Left)* User engaging with four distinct user experiences (bubbles), defined by four independent products (green rectangles). *(Right)* Instead, the same four products contribute to a single, consistent user experience (user inside the bubble).

To support the notion of cross-platform media experiences, we argue that a solution to coordination is needed, which is not limited to any particular application or platform. We envision an extended media model, with built-in support for coordination. In this model, media products can still be developed for single-platform usage, yet optionally be exploited as parts of larger, consistent, cross-platform media experiences.

III. Problem Statement

Extend the current media model with built-in support for coordination between independent devices, technology types, platforms and product interfaces. The extended model shall

1) support consistent user experiences across platforms *(consistency)*
2) support the combination of continuous and data-driven media *(bridge)*
3) support integration with existing media systems *(compatible)*
4) support high flexibility while limiting complexity *(practical)*

5) support key trends such as automation, adaptation, personalization *(future-proof)*

IV. Current Media Model

Even though *continuous media* and *data-driven applications* represent different technical approaches, we regard them as instances of the same media model. Fig. 2 illustrates this model as a 3-step processing chain (rectangles). The data flow is left-to-right, from data (black) on the left, through assembly (blue) and rendering (green) before reaching the user experience on the right (bubble). User-interactivity (not illustrated) flows in the opposite direction, triggering data updates (black arrow) or control actions (red arrow). Processing steps (data, assembly, render) may be executed within the same process, or be separated by a communication link or a network connection.

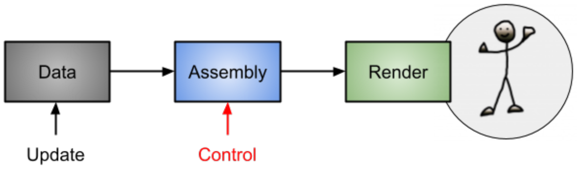


Fig. 2. Conceptual sketch for current media model. The data flow is left-to-right, from data (black) on the left, through assembly (blue) and rendering (green), into the user experience on the right (bubble).

1) Data: Data represents resources for media experiences. Resource types may include common media formats such as audio, video, images, XML/JSON-data, declarative layout or stylesheets. Resource abstractions may include files, streams, databases, or datasets. Resources may be static or dynamic, local or accessed via a network. Access may also be restricted through concepts of ownership and access credentials.

2) Assembly: Assembly (blue) implements a controlled conversion from data sources to render state. This conversion may include processing steps such as selection, filtering, merging, and transforming data from multiple sources. Moreover, assembly is open to dynamic control (red arrow), defining for instance which data sources are selected for rendering, how data is converted into render state, and when. Assembly may also be associated with a media clock, and be expected to update the render state consistently with respect to time progression. This is commonly referred to as playback or sequencing.

3) Render: Render (green) is a function that converts render state into visual display, sound, or other real-world effects. Render components are assumed to be software components, acting as low-latency or fixed-latency proxies for locally connected output devices, such as screens and loudspeakers.

4) User experience: The user experience (bubble) is defined by rendered effects (e.g., audio, visuals, or other). The user may also interact with the media experience, using local input devices such as keyboards, pointing devices and more. User interactivity may result in updates to data sources (black arrow) or control actions (red arrow) targeting the assembly process.

In this model, the assembly step represents the beating heart of the media experience. As either data or control inputs change, the assembly process must continuously reevaluate and adjust the output render state. Over time, this produces a time sequence of render states that defines the progression of the media experience.

A. Continuous Media

Fig. 3 illustrates *over-the-top (OTT)* audio and video production as an instance of the current media model. In this context, data access and media assembly primarily occurs within a production environment. For example, video production may involve a large set of data sources and media contents, mixed together into a single audio or video asset and streamed over a network for playback on viewer devices. Provider-side media assembly may also involve a large production team controlling a number of production parameters, including camera positioning, lighting, sound processing, graphics and visualization, as well as mixing and scheduling. A few aspects of assembly are open to local control by viewers. For instance, media players typically allow viewers to adjust the volume, pause/resume, and select subtitle tracks.

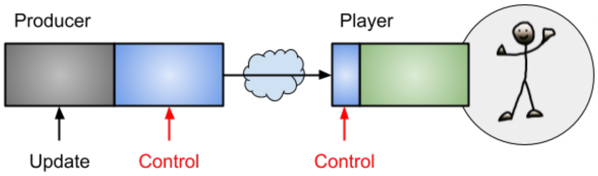


Fig. 3. Data assembled into video or audio streams at the provider side, then distributed for client-side rendering by a media player.

B. Data-driven Applications

Fig. 4 illustrates data-driven Web or native applications as instances of the current media model. In this context, data sources are hosted by servers. Clients connect to fetch or stream data over the network, or receive pushed data. Assembly and rendering can be handled on the client side. For instance, in Web interfaces, clients convert data sources and application state into render state managed by the *Document Object Model (DOM)* [16]. The conversion is defined in application code and may include selection or filtering of data, and also transformations or combinations of data. Moreover, interactive control by the user may trigger changes to the local control state or network-accessible data sources.

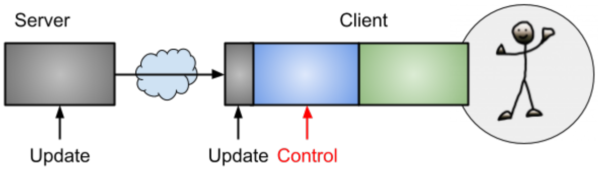


Fig. 4. Web and native applications perform assembly and rendering at the client-side, based on data fetched or streamed from online servers.

V. Control-driven Media

We envision cross-platform coordination as an inherent capability of the media model. Recognizing the central role of *control* in media (see Section IV), we propose a revised media model, where control is redefined from a local signal to an independent, stateful resource, and extended with support for consistent sharing across platforms. This seemingly small change, we argue, comes with profound implications for media systems, applications and user experiences, and also alters the status of control as an object of research, from an application-specific interface feature to a principal system component.

Fig. 5 illustrates the revised media model. Similar to Fig. 2, the model describes a three-step technology stack (Data, Assembly, Render). The difference, though, is that control (red rectangle) is regarded as a special kind of data source and is included in the first step (Data). Like data sources, control sources may be hosted by dedicated online services and accessed by connecting clients.

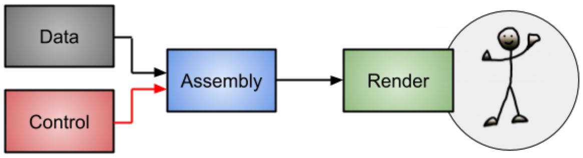


Fig. 5. Control-driven Media, with control represented as a special kind of data source.

Importantly, the revised model does not represent a radical break with the current model. Data sources (black) and render components (green) are unaffected. Assembly (blue) will react to state changes in control sources, instead of live control inputs, but otherwise remain the same. Also, user interactivity (not illustrated) will not target Assembly (blue) directly, but indirectly through modification of data and/or control sources. Beyond this, the model remains the same.

This model is referred to as *Control-driven Media (CdM)*. The name highlights how media production and presentation can be directed through the manipulation of control sources. For example, control sources may define which resources and rendering components are active (e.g. video sources, audio tracks, datasets, layout templates, graphical elements). Control sources may also define values for a variety of interface parameters (e.g. style properties, playback offsets, audio levels, viewport positions). By manipulating the value of such control sources, media experiences will change in predictable ways. As such, control sources represent aspects of application behavior or appearance which are explicitly opened up for external control. The following are defining properties of Control-driven Media.

1) Control is defined as a stateful, shareable resource, local or network accessible.
2) Control is defined in reference to a timeline.
3) Control defines the experience.
4) Assembly is an independent step.

1) Control is a stateful, shareable resource, local or network accessible: Control sources may be local objects. However, with control represented as online resources, considerable opportunities arise with respect to sharing and exchange of control. In particular, control may be distributed in real time between any connected clients across various interfaces and platforms. This may allow media providers to implement production control across a distributed production chain and also across consumer devices. Effectively, this makes consumer devices an integral part of the production infrastructure, blurring the distinction between production and presentation. Online control may also provide a valuable integration point for automated processes, for instance allowing cloud-based AI agents to remote control user experiences in accordance with user preferences. Moreover, online control can be shared between media providers and viewers, opening up for increased collaboration. For instance, when the viewer requests more detailed graphics, the production system can access the online control state for this viewer and implement appropriate changes through modification of relevant control sources.

Control as a stateful resource opens up for a variety of control patterns. Online controls may be private, public, or limited to groups. Access restrictions may also discriminate between roles such as owner, editor or viewer. Online control can support multiple patterns for control exchange (1:1, 1:N, M:1 or M:N) and support both one-way (asymmetric) and multi-way (symmetric) control relations.

2) Control is defined in reference to a timeline: With control as a network-accessible resource, latency is no longer negligible. This is particularly problematic as control signals are often time-dependent. Control actions might refer to a specific offset on a media timeline and/or describe time-dependent transitions. When transferring control signals over a network, such temporal relations must be preserved. Control may also be shared between processes operating in different time-frames. For instance, on-demand consumption requires time-shifted replay of previously recorded control sequences. Moreover, in scenarios with real-time sharing, small skews may be introduced to mask network jitter and avoid buffering issues. To support this, we assert that control must support timeline-consistent sharing between processes. Timeline consistency implies that a control signal can be captured and serialized in reference to a media clock, and reproduced correctly according to a different media clock.

Consistency with a timeline is also a defining characteristic of continuous media. This means that control sources may be regarded as media objects in their own right, and also be described using terminology traditionally reserved for continuous media types. So, like video, control may be captured, recorded, distributed, time-shifted, rewinded, edited, and played back.

3) Control defines the experience: While CdM changes the nature of control, it maintains established distinctions between data and control. Data sources tend to be raw material for an experience, whereas control sources define a particular realization. There might also be a certain asymmetry between the two entities, where data sources may represent large and stable datasets, whereas controls are more lightweight and dynamic. As such, control provides a basis for variation, where different experiences can be produced, without necessarily modifying the data. By opening up for production and distribution of control as a separate and independent resource type, this pattern can be exploited directly in consumer interfaces, providing personalized user experiences as unique permutations of shared data sources and personalized control sources.

> CdM implies a state-based, reactive approach to control. Developers define what is considered control state in a particular application, how control state can be modified, what sharing scope and access restrictions are appropriate, and how control is coupled with application logic (i.e. assembly and interface components). Interface components typically initiate control actions, whereas assembly processes react to changes in control state. If control sources are shared online, this decoupling between producers and consumers of control may apply in the global scope.

4) Assembly is an independent step: In the current media model (see Fig. 2), local control signals dictate that assembly is co-located with the controller, be it the media provider in a studio, or the user in an interactive interface. This forces a choice between provider-side and client-side assembly, with no clear middle ground. With control as an independent resource, assembly instead becomes an independent processing step, with no particular restrictions concerning location. This means that the assembly function can be shifted between locations without modification, for instance between a physical studio and consumer devices, even if rendering technologies may be different for these locations. Moreover, assembly can be split into logical steps and assigned to different nodes in a production chain, such as cloud-hosted production services, edge nodes, and consumer devices (see Fig. 6). Even though nodes may operate in different time-frames, and are not necessarily directed by the same control parameters, the entire chain can still be controlled through the same mechanism.

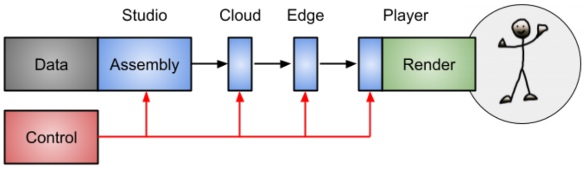


Fig. 6. A stepwise production chain, from physical production studio to consumer device, through cloud-based production and CDN/edge services. Control is time-shifted for each step to allow time for data transfer.

A. Cross-platform Media Experiences

Control-driven Media (CdM) provides a new and highly flexible foundation for cross-platform media experiences. Fig. 7 illustrates three media products $\{P1, P2, P3\}$ hosted by different platforms. Collectively, these products are supported by assembly functions $\{A1, A2, A3\}$ (blue), control service $\{C\}$ (red) and data services $\{D1, D2, D3, D4\}$ (black). Dependencies for each product may be deduced by traversing the graph in reverse, starting from rendering components. For instance, product $\{P2\}$ is defined by render components $\{R2, R3\}$ depending on $\{A2, A3, C, D2, D3, D4\}$.

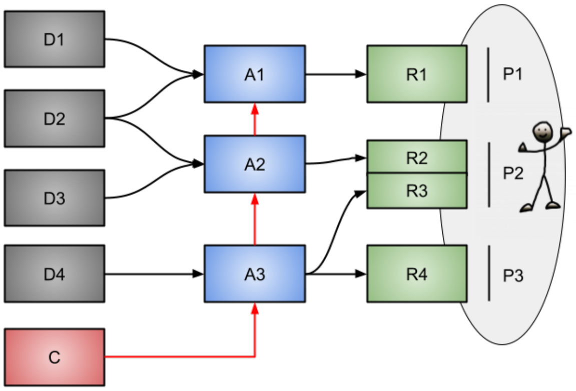


Fig. 7. Cross-platform media experience with three separate products supported by independent technology stacks.

While products $\{P1, P2, P3\}$ may be developed and used independently, they can also be combined to form a larger cross-platform media experience. In principle, this comes down to the scope of control resources, i.e. whether control resources are defined to be exclusive for each product or shared across multiple products. Moreover, with a loose coupling between control resources and application components, aspects of control sharing may easily be changed between applications, interfaces, or even dynamically during a session. Importantly, this means that CdM can support cross-platform media experiences, without requiring architectural changes or significant changes to existing backend systems or render components.

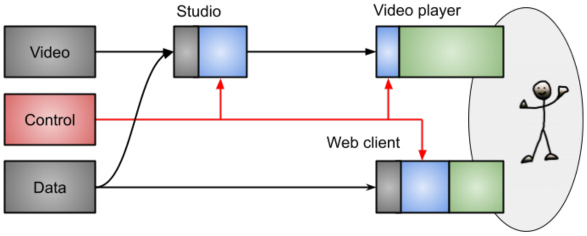


Fig. 8. Cross-platform media experience with video and data-driven visualization.

Fig. 8 illustrates a combined production chain with video content and Web visualization. For instance, in the context of Formula1 racing, *Video* (top, left) could be camera feeds from the track, and *Data* (bottom, left) could be a live dataset with lap times. In the *Studio* (top, middle), a producer may then add data-driven graphics to the video content, for instance switching between a *lower third* with the most recent lap time for the featured driver and a larger table with lap times for different drivers. The *Web client* (bottom, right) may present similar, HTML-based graphics, from the same data source. Furthermore, control sources can be shared between different steps in the production chain. This way, actions taken by the studio producer (e.g. activating or deactivating a graphics element) could also be used to direct Web clients, though time-shifted to match the time-frame of the *Video player* (top, right). The resulting user experience could then be a consistent narrative presented across very different technologies.

VI. Applications

Control-driven Media (CdM) provides ample opportunities for innovation. This section details specific examples in the context of *Formula1 (F1)* racing in order to highlight some possibilities. Still, the themes discussed are likely relevant for a broad range of media applications, and also beyond the world of online sports.

A. Formula 1 Online Coverage

Over recent years, Formula 1 has strengthened its online presence, adding the F1TV [14] streaming service and the F1 App [15] to their classical broadcast offering via TV rights deals. In total, F1 caters to a worldwide audience of motorsport enthusiasts, with an estimated viewership of about 70 million per race in the 2023 season [17]. Ambitions are clearly stated at the F1 site [18]: “Every F1 session live and on-demand.” F1 coverage offers a variety of interfaces into race events, based on vast amounts of data and content streams captured by sensors, cameras and microphones – placed inside vehicles and around the racing track. Products include premixed video channels, specific camera angles or audio tracks, maps with live driver tracking, timetables, interactive visualizations of sensor data and analytics, as well as edited highlights. F1 fans enrich their own experiences by switching between different interfaces, or using multiple interfaces in parallel. In particular, by opening multiple instances of F1TV (5 simultaneous devices are allowed per account) and F1 App on different devices, more screen estate is available for a more immersive experience (see Fig. 9).

Fig. 9. Formula1 online offerings. Image credits formula1.com

B. Case: Adaptive Graphics

Graphics are an essential part of the F1 experiences, amplifying and illustrating important trends and sudden developments from an otherwise overwhelming data corpus (estimated to over a TB of data per race, just from the cars [19]). With adaptive graphics even more opportunities arise. For example, graphics could adapt to custom aspect ratio and personal preferences (e.g. “novice”, “regular”, “engineer”). This is not feasible with traditional video production.

In CdM, graphics can be rendered on the client-side, as video overlays or standalone, data-driven components, while at the same time being controlled from a centralized production system. Client-side graphics can be sensitive to user interactivity and local context (e.g. available display area, GPU support, power-level). Control-driven graphics may also be precisely aligned with video content and connected to live data sources. For instance, during strategically important periods, specific graphics could monitor relevant details for a specific strategic battle. Such graphics already exist, but are only shown sporadically in the World Feed, and normally only for top-team cars involved in a close battle. Furthermore, time-shifted graphics may display up-to-date information, and also remain open for interactive exploration.

C. Case: Immersive Multi-device

F1 is a data-intensive sport, and many F1 enthusiasts are immersing themselves with different views across multiple devices. A successful multi-device setup provides more screen estate to the user experience, and also allows capabilities of different device types to be exploited, for instance by placing video contents on large screens, and interactive features on handheld devices with touch displays. In current offerings, all configuration, synchronization and selections must be done manually by the user.

In CdM, multi-device usage could be simplified by automated configuration. For instance, in response to a pit stop period, video sources and F1 App visualizations could be switched to present relevant data, and personalized HTML-based overlay graphics could be presented consistently with both video content and app visualizations. Control sources may also be used for real-time interactivity between interfaces. For instance, driver selection in the F1 App could affect graphics and content selection across multiple devices. Moreover, control state could also define a mapping between interface components and devices, opening up for automated reconfiguration as devices join or leave the experience [20]. This could for instance be controlled by AI-based agents trained on manual configuration patterns or predefined policies.

D. Case: Lean-back, Lean-forward

F1 advertises a highly customizable product, where viewers can shape their own experience by switching between F1TV streams and interacting with F1 App visualizations. While such interactive engagement is clearly attractive, it may also be taxing in the long run, and even the most forward-leaning F1 enthusiasts may want to lean-back at times, for instance during highly exciting race sequences. In current solutions though, this typically implies a set back to generic broadcast coverage, e.g. the World Feed.

With CdM, the distinction between lean-back and lean-forward coverage can be softened. For instance, by actively controlling aspects of data-driven interfaces, media providers may exploit traditionally lean-forward technologies as part of a produced narrative. Shifting between lean-back and lean-forward modes can also occur seamlessly within the user experience. For instance, switching from lean-back coverage to interactive exploration may simply be a matter of switching control sources, from official provider controls to private controls.

E. Case: Personalization

F1 caters to a highly diverse audience, and many users might appreciate specialized coverage based on preferences such as team affiliation, nationality, language, proficiency level (novice, regular, enthusiast), or accessibility requirements. Individual adaptations, though, are costly in the context of video production.

In CdM, user experiences may to a larger extent be assembled from smaller building blocks (i.e. control and data sources). This provides multiple opportunities for variation. Interfaces may be set up with different versions of data and control sources, and assembly logic may combine them in different ways. This way, CdM may provide the appearance of unique adaptation, even from a modest number control parameters. Personal preferences may also be reconfigured with immediate effects. For instance, changing the proficiency level from “regular” to “enthusiast” may trigger a cascade of changes with low-level control resources, effectively transforming the experience.

F. Case: Social and 3’rd Party Integration

By opening up for collaboration and 3’rd party contributions, media experiences could become both richer and more social. For instance, graphics could indicate the presence of friends watching the same event, and allow viewers to join sessions, or receive reactions from friends who watched earlier. A group of friends could choose to watch the latest F1 race together online, taking turns bringing up interesting stats in the F1 App for all to see and discuss. This notion of collaborative production could also extend to 3’rd party service providers. For instance, a national group for Ferrari fans could provide its members with exclusive contents through the official F1 platform. With a closed production system though, such integrations are often complex and also problematic with respect to content ownership.

In CdM, social and 3’rd party integration can be addressed through control sharing. For instance, by sharing control sources between friends, race events can be presented in synchrony, ensuring that all react to race events at the same time. It would also be possible to follow the experience of a friend or actively direct some aspects of the presentation for a group. Moreover, 3’rd party content sources could be integrated directly into client-facing interfaces, thereby avoiding integration with backend systems.

G. Case: Value-added Time-shifted Coverage

F1 is a live sport, and most F1 viewers prefer to follow live coverage. Still, F1 races are hosted at venues around the world and across many different time zones, making time-shifted consumption a more convenient option in some cases. Time-shifted coverage could also provide other benefits. For instance, production errors could be corrected, and results from time-consuming analysis could be added. A lot of information is also held back during the race itself for strategic reasons. Currently though, time-shifted offerings are limited to replaying the original video content and edited highlights.

In CdM, time-shifted coverage is not limited to video, but may also include replays of data-driven interfaces. The latter implies that time-shifted coverage can remain interactive and open to adaptation. Moreover, control sources can be modified at any time after production, for the benefit of time-shifted viewers. For instance, post-race analysis could optimize control state for the selection and scheduling of video sources (i.e. camera angles) to perfectly capture passes and race incidents as hindsight gives perfect information. This may also apply to live viewers, as live experiences are also time-shifted to some degree, for instance due to delays in production chains for video content.

VII. Technical Challenges

Control-driven Media introduces some new technical challenges. A key challenge concerns the redefinition of *control*, from local signal to online resource. Another important challenge concerns increased complexity in *assembly*, where potentially a large number of data and control sources need to be consistently converted into render state.

A. Control

In CdM, control is a stateful, network accessible resource, defined in reference to a timeline (see Fig. 10).

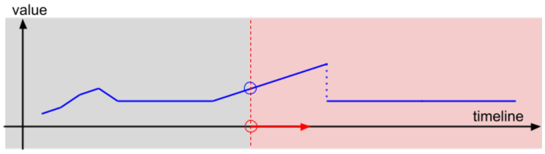


Fig. 10. A control parameter defined in reference to a timeline. The control value (blue) increases and decreases gradually, remains static for a while, increases again, before it is abruptly set to a lower value. By time-shifting the media clock (red ring), past values of the control parameter may be played back again. Illustration adapted from [21]

Challenge 1 - Generic control: The ability to share control is a fundamental property of CdM. Moreover, control sources must support a variety of control patterns to be shared across different time-frames and across a variety of products and application types. To allow such flexible usage, CdM will benefit from a common abstraction for control. At the same time, though, applications are different, and developers must also be able to specify unique control relations and custom control logic. The challenge, then, is to provide a generic and flexible mechanism for control sharing, to be used across very different applications, while at the same time opening up for application-specific extensions or adaptations.

Challenge 2 - State representation: In order to be shared across a network, a serialized format must be defined for control state. One challenge is to define a representation which supports various types of control, including discrete and dynamic state changes. Another challenge concerns the representation of dynamic control signals (e.g. transitions or pointer drags), which must be reproduced with high fidelity for a pleasing user experience, while also minimizing network traffic and latency. Solutions might seek to downsample, approximate and compress dynamic signals ahead of distribution, and correspondingly decompress or upsample signals on receipt. Transitions can be expressed as deterministic mathematical functions, making them particularly effective in terms of network bandwidth.

Challenge 3 - Consistent state sharing: A protocol must also be defined for efficient distribution of control state and changes. Clients observing an online control source need to maintain a consistent view of control state. This may for instance be achieved by receiving the initial state on connect, followed by notifications for subsequent state changes. In a centralized architecture, global ordering for state updates can be used to ensure eventual consistency for observing clients. Different consistency models are possible for control state. A decentralized architecture might also be possible, as long as the consistency model matches application requirements.

Challenge 4 - Low latency: Control sharing over a network implies additional delays for distributing state changes. For instance, in a centralized architecture, update requests must be transmitted from client to a server, before change notifications can be multicast back to observing clients. Internet delays are generally too high for interactive applications, where immediate feedback is expected for interactive control operations. However, this can be addressed by implementing control changes speculatively locally, ahead of Internet distribution, thus removing delay for the local interface. Importantly though, speculative changes may lead to consistency issues, and must occasionally be rolled back if the local ordering of state changes is not consistent with global ordering. In addition, low latency distribution is important for real-time control sharing, and collaborative scenarios in particular. General principles apply for efficient implementation of low latency distribution.

Challenge 5 - Timeline consistent control: Finally, online sharing of control state must also support timeline consistency. This means that the control state must be serialized in reference to a media clock (capture), and correctly reproduced in reference to a different media clock (playback). For example, a live control signal originating from a pointing device may be encoded as a sequence of states, each referring to a segment on the timeline. In playback, these state changes must be repeated at the correct time, including both discrete and dynamic control changes. Reproduction of control state must also be sensitive to changes to the media clock, such as pause, resume or time-shift. Timeline consistency across devices or products requires that devices and products have access to a common media clock.

B. Assembly

In control-driven media, assembly implements a conversion from data sources and control sources, into state prepared for rendering. The objective is to encapsulate application-specific logic and provide simpler resource abstractions matching the requirements of render components. This way, rendering components can be efficient and stateless data sinks, as illustrated in Fig. 11.

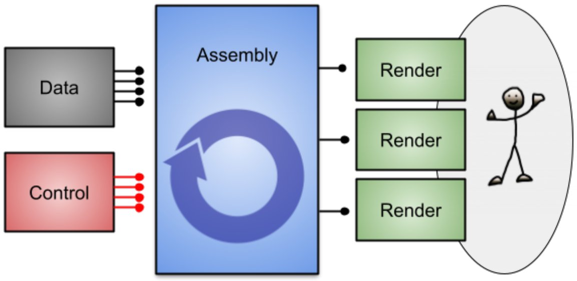


Fig. 11. Assembly converting four data sources and four control sources into three virtual sources of rendering state.

Challenge 6 - Masking heterogeneity: In CdM, assembly (blue) is represented as an independent step, decoupled from data and control sources (black, red) and rendering components (green). By encapsulating application-specific functionality in the assembly step, render components may remain simple, generic, and possibly stateless. This provides opportunities for the reuse of render components across different versions of assembly. Similarly, data and control sources may be reused across different assembly components, or dynamically replaced by alternative sources. To unlock this flexibility, masking heterogeneity will be a key challenge. For instance, by introducing a uniform resource abstraction, assembly functionality may work across input sources with different API or data formats. Similarly, if assembly exports render state in a uniform way, render components may more easily be reused across different assembly components.

Challenge 7 - Timeline consistent render state: The assembly step may also be required to produce render state in accordance with a media clock. While timeline consistency has already been defined for control sources, it may additionally apply to sources of timed data, such as media tracks, logs or timed event sequences. For instance, given a subtitle track, the correct subtitle must be activated and deactivated in render state, during playback. Moreover, if the media clock is altered (e.g. pause, rewind, resume), or if the data source itself is modified (e.g. subtitle edit), render state must be modified accordingly. While this is feasible today using specific tools or ad-hoc solutions, in CdM, timeline consistency is regarded as an integral aspect of state management. The challenge then is to define a practical programming model for time-dependent state assembly, including appropriate concepts and tools.

Challenge 8 - Complexity: The assembly step implements an application-specific conversion from multiple control sources and data sources into active render state. This may involve a variety of standard processing operations such as merging, layering, filtering, aggregation, selection, transformation, etc. There may also be dependencies between sources, constraints, or even conflicts. For instance, a control source specifying more advanced graphics may have to be ignored if the screen is too small, or if power is running low. Moreover, assembly must be well-behaved for a vast set of permutations of control and data sources. This increases code complexity and the burden on developers. The challenge, then, is to allow application-specific and dependable assembly functions to be implemented while limiting the complexity. For instance, framework support for assembly could support advanced processing graphs built from simple, pre-defined processing operations, each open to application-specific parametrization and customization.

VIII. Evaluation

Control and assembly are key challenges in Control-driven Media (CdM). With respect to control, State Trajectory [21] may be a candidate solution. State Trajectory is a unifying concept for interactivity and control in data-driven applications, with built-in support for online sharing and timeline-consistent playback. State Trajectory is supported by implementation and addresses challenges (1-5) in Section VII-A. Concepts and tools for assembly challenges (6-8) in Section VII-B is work in progress. This section presents prior research supporting CdM, and then demonstrates that CdM addresses the problem statement set out in Section III.

A. Prior Research

Control-driven Media (CdM) builds on extensive research over the last 10+ years, in topics such as distributed media synchronization, real-time data sharing, multi-device media experiences, and timeline-consistent, data-driven visualization. Research results include reference implementations for key concepts and services, technical evaluation, and prototype applications.

Addressing the need for timeline-consistent rendering in data-driven media experiences, the Timing Object [22] was introduced in 2015 as a generic concept for media clocks and timeline control. Timing objects are stateful resources that can be shared, observed, and manipulated in application code. Research in media synchronization explored the limitation of audio and video synchronization on the Web platform, demonstrating that echoless synchronization was possible on high-end smartphones as early as 2015 [23]. Sequencing tools for dynamic datasets [24] demonstrated precise synchronization of time-dependent data. Media State Vector [25] demonstrated a scalable solution for online media clocks, with global availability and precision down to a few milliseconds. Effectively, this extended the scope of Timing Objects from a single interface to online multi-device applications.

Together, these concepts define a media model for timeline-consistent, data-driven, multi-device, media experiences [26]. Timingsrc [27] provides reference implementations for time control, media synchronization and sequencing tools for the Web platform. These concepts, along with practical solutions for real-time data sharing, have consistently demonstrated their value as building blocks in data-driven, multi-device media applications.

CdM extends this model by providing a more generalized concept for control. In particular, control in CdM is not limited to media clocks, but can in principle apply to any parameter of a media experience. State trajectories [21] may represent application variables of different types (e.g. int, float, string, object, collection). Moreover, State trajectories track value changes in reference to a timeline, supporting discrete as well as time-dependent (e.g. deterministic, dynamic) changes. State trajectory has been evaluated through implementation, demonstrating a lightweight footprint (CPU, bandwidth, storage), real-time sharing, timeline-consistent playback and a high potential for scaling.

Additionally, CdM extends the scope of assembly. Prior research in data sequencing [24], exclusively addressed timeline consistency for timed data. In CdM, timeline consistency is seen as part of a larger challenge - application state management. It follows that challenges regarding timeline-consistency must be addressed alongside other software concerns, such as modularity, flexibility, composition, customization, and uniformity (masking heterogeneity). To assist developers in reaching these goals, ongoing research targets framework support for assembly in CdM, and development of new and appropriate programming concepts which encapsulate support for timeline-consistency.

B. Addressing the Problem Statement

1) Support consistent user experiences across platforms: CdM provides consistent user experiences across platforms by (i) representing control as an independent system component, open to integration in user-facing products as well as production systems, (ii) opening up for consistent sharing of control between interfaces, and (iii) supporting timeline-consistent assembly of render state.

2) Support the combination of continuous and data-driven media: CdM allows control state to be shared between media players and data-driven interfaces (e.g. overlays, secondary devices), as a basis for consistent presentation and coordination, without introducing any additional limitations. Moreover, in CdM, control sources may be integrated with production systems for continuous media, or time-shifted for direct usage in interactive interfaces on consumer devices.

3) Support integration with existing media systems: CdM does not represent a radical break with existing systems, but rather adds control as a new and independent system component. As such, CdM remains compatible with existing data backends and rendering technologies. Moreover, the added cost of online control is likely modest, as online controls can be massively scaled by dedicated, cloud-hosted services, and do not introduce significant overhead in client interfaces [21]. A stepwise integration path is possible, starting with select render components in specific products.

4) Support high flexibility while limiting complexity: CdM maximizes flexibility by decoupling assembly from control and data sources as well as rendering components. At the same time, CdM provides powerful, generic concepts for control and assembly, encapsulating significant complexity. This combination of decoupling and unified resource representation also makes CdM a practical model for developers, with opportunities for modularity, code reuse, composition, specialization and dynamic recombination.

5) Support key trends, such as automation, personalization, collaboration: In CdM, media experiences are explicitly opened up for external control, through online control sources. This allows automated production processes to monitor live developments, and also direct a narrative through complex control sequences. CdM media experiences are highly flexible, and can support personalization of data, control sources and assembly logic. Moreover, CdM supports collaboration through real-time interactions with control and data, and also collaboration across time-frames, by time-shifting control or data.

IX. Discussion

1) Quality, costs and scalability: Quality, costs, and scalability are important measures for media providers. Importantly, Control-driven Media (CdM) does not significantly alter these measures. Quality, costs and scalability will still be largely determined by existing infrastructure for the production, distribution and rendering of data sources and media content. In comparison, control sources are shown to be lightweight entities (CPU, bandwidth, storage), and may likely be hosted cheaply at a massive scale by dedicated services. On the other hand, CdM may offer improved quality of service, for instance, through provider-driven interface control and consistency in presentation. Such features may be offered to premium subscribers and motivate users to log in to sites that would otherwise be open. CdM may also support high levels of personalization at scale while limiting costs. This could be possible by implementing unique adaptations as part of media assembly on the client side, thus ensuring that consumer devices carry much of the costs. CdM uniquely supports this by offering a common mechanism for production control across the entire production chain, including client-side interfaces.

2) Automation and AI: AI-assisted automation is a hot topic in media production, with opportunities for cheaper production of high-quality content sources. However, automation of the producer role may represent an even larger potential. For instance, AI-based agents could engage in storytelling, weaving together narratives from content sources, datasets, graphical layers and interactive visualizations, adapted to individual preferences, device capabilities or other localized contexts. Essentially, this would correspond to a shift from text-based narratives of first-generation *large language models (LLM)* to true multimodal narration. CdM is well positioned to support this shift. Control sources provide a natural integration point, allowing AI-based agents to monitor control state from consumers, and express narratives through the manipulation of control sources. Moreover, AI-agents do not have to be hosted on consumer devices but may remote control AI-driven experiences from a cloud-hosted production system. Furthermore, AI-based agents may learn directly from live interactivity or be trained by time-shifted replays of historical sessions.

3) Online multiplayer games: CdM bears some similarity to online multiplayer games, where gameplay is driven in real-time by joysticks or fast-paced, pointer-driven interaction. Virtual 3D games also demonstrate a striking potential for client-side rendering, where hardware acceleration (GPU) is exploited to provide rich, smooth and responsive graphics, from a shared data model and a modest number of control parameters (e.g. player position, movement, orientation). However, in gaming platforms such as *Unity and Unreal* [28], [29], control mechanisms are an integral aspect of platform design and are highly optimized with respect to game logic and competitive fairness. In contrast, CdM offers technical solutions for control which may be reused in different applications, or serve as a bridge between platforms. While CdM indeed supports real-time sharing of dynamic control state, support for consistent time-shifting and playback of control may be even more valuable. For instance, in the context of live esports, CdM controls could be used to broadcast control state from competitive gameplay, opening up for slightly time-shifted rendering directly in game engines. Compared to live video distribution, this might offer more flexibility and higher resolution, while also reducing delays and distribution costs.

4) Media orchestration: CdM may also be compared to solutions for command-driven media orchestration. For instance, MIDI [30] is a classic solution for orchestration of equipment related to musical or theatrical performances. MIDI allows command messages to be broadcast within a group of devices, originally restricted to a low-latency, cabled setup. By distributing commands regarding tempo, pitch, notes, volume etc., different instruments may be directed to play in synchrony. MIDI may be used for live collaborative sessions as well as time-shifted performances based on scripted or pre-recorded command sequences. In contrast, CdM models control as a stateful, time-dependent and generic resource, as opposed to transient, application-specific commands. This way, CdM may extend the scope of media orchestration from local to global network, from specific application domain to general usage, and also encapsulate complexity related to time-shifting.

5) A Unifying media model: Finally, CdM may be regarded as a unified model for online media, encapsulating characteristics of classical approaches, such as continuous media, data-driven, interactive media, and command-driven media orchestration. Following this, CdM helps bridge common technical divides, such as one-way broadcast vs multi-way collaboration, single vs multi-device presentation, lean-back storytelling vs lean-forward engagement, or generic coverage vs individual adaptations. With CdM these divisions can largely be reinterpreted as variations in control patterns and sharing scopes. As such, CdM may offer a unifying approach to online media. Moreover, with a more flexible media model, media applications can avoid limitations from early design choices and easily be extended in new directions, if needed.

X. Conclusion

Vast opportunities in media technologies and automation, paired with growing viewer diversity, fuel a demand for increasingly advanced and varied media experiences. To meet the demands of both viewers and media providers, there is an increasing need to leverage complementary technologies, platforms and interfaces as part of a single, consistent user experience. While this challenge is typically addressed within the context of specific applications or platforms, this work shows that a generic solution is possible.

This paper presented *Control-driven Media (CdM)*, an extension of the current media model with built-in support for consistent, cross-platform user experiences. CdM introduces online control as a generic and independent resource type in media systems, thereby opening up for media experiences to be consistently controlled and coordinated across different device types, user interfaces, media products and technology platforms. CdM may also be seen as a generalization over existing approaches; Like continuous media, CdM supports media experiences which are *provider-driven* and *time-driven*. Like Web and native applications, CdM supports experiences which are *data-driven* and *user-driven*. Moreover, CdM is compatible with existing service backends and frontend technologies.

To further demonstrate the flexibility of CdM, the paper also presented a number of opportunities for innovation, relevant for key challenges in the media domain, such as AI-based automation, personalization, collaboration and multi-device support. The evaluation referenced an extensive backlog of supporting research, emphasizing that solutions already exist for key technical challenges.

Finally, by tackling the problem statement from Section III, CdM is confirmed as a practical and forward-looking solution for cross-platform media experiences, with the flexibility needed to address increasingly complex demands from users and media providers.

Acknowledgment

This work was supported by the Research Council of Norway and MediaFutures (Grant number 323302 and 309339).

References

[1] W3C. (2024) HTML Living Standard Media. [Online]. Available: https://html.spec.whatwg.org/multipage/media.html

[2] (2009) Hbbtv. hybrid broadcast broadband tv. HbbTV. [Online]. Available: https://www.hbbtv.org/

[3] (2015) Dvb-css. companion screens and streams. etsi ts 103 286-1 technical specification. DVB. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/103200_103299/10328601/01.01.01_60/ts_10328601v010101p.pdf

[4] (1999) Akamai content delivery network. Akamai Technologies, Inc. [Online]. Available: https://www.akamai.com/

[5] (2009) Cloudflare content delivery network (cdn). [Online]. Available: https://www.cloudflare.com/cdn/

[6] (2011) Fastly content delivery network. [Online]. Available: https://www.fastly.com/

[7] ISO, "Dynamic adaptive streaming over HTTP (DASH)," International Organization for Standardization, Standard, Aug 2022. [Online]. Available: https://www.iso.org/standard/83314.html

[8] R. Pantos and W. May, "Http Live Streaming," Internet Requests for Comments, IETF, RFC 8216, August 2017. [Online]. Available: https://tools.ietf.org/html/rfc8216

[9] I. Fette and M. Alexey., "The WebSocket Protocol," Internet Requests for Comments, IETF, RFC 6455, Dec 2011. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6455

[10] (2015) Object-Based Media. BBC Research & Development (R&D). [Online]. Available: https://www.bbc.co.uk/rd/object-based-media

[11] M. Armstrong, M. Brooks, A. Churnside, M. Evans, F. Melchior, and M. Shotton, "Object-based broadcasting-curation, responsiveness and user experience," in *IBC2014 Conference*. IET Digital Library, 2014.

[12] M. Evans, T. Ferne, Z. Watson, F. Melchior, M. Brooks, P. Stenton, I. Forrester, and C. Baume, "Creating object-based experiences in the real world," *SMPTE Motion Imaging Journal*, vol. 126, no. 6, pp. 1–7, 2017.

[13] (2021, Sep) Object-based Media report. The Office of Communications (Ofcom). [Online]. Available: https://www.ofcom.org.uk/research-and-data/technology/general/object-based-media

[14] (2018) Formula 1 TV (F1TV). [Online]. Available: https://f1tv.formula1.com/

[15] (2018) Formula 1 Mobile App (F1 App). [Online]. Available: https://www.formula1.com/en/subscribe/download-the-official-F1--app.html

[16] "Document Object Model (DOM)," Tech. Rep., Feb 2024. [Online]. Available: https://dom.spec.whatwg.org/

[17] C. Brittle. (2023, 11) F1 2023 season review: Steady tv viewership, attendances rising, and the andretti problem. Blackbook Motorsport. Accessed: 2024-03-11. [Online]. Available: https://www.blackbookmotorsport.com/features/f1-2023-season-review-tv-viewership-attendance-andretti/

[18] (2022) Formula1. stream f1 live, your way. Accessed: 2024-03-11. [Online]. Available: https://www.formula1.com/en/subscribe-to-f1-tv.html

[19] (2024) Feature: Data and electronics in f1 explained. AMG Petronas Formula One Team. Accessed: 2024-03-11. [Online]. Available: https://www.mercedesamgf1.com/news/feature-data-and-electronics-in-f1-explained

[20] M. Zorrilla, N. Borch, F. Daoust, A. Erk, J. Flórez, and A. Lafuente, "A web-based distributed architecture for multi-device adaptation in media applications," *Personal and Ubiquitous Computing*, vol. 19, pp. 803–820, 2015.

[21] I. M. Arntzen, N. T. Borch, and A. Andersen, "State Trajectory. A Unifying Approach to Interactivity with Real-Time Sharing and Playback Support," in *Proceedings of the Future Technologies Conference (FTC) 2023, Volume 2*, K. Arai, Ed. Cham: Springer Nature Switzerland, 2023, pp. 1–20.

[22] I. M. Arntzen, F. Daoust, and N. T. Borch, "Timing Object; Draft community group report," Tech. Rep., Nov 2015. [Online]. Available: http://webtiming.github.io/timingobject/

[23] N. T. Borch and I. M. Arntzen, "Mediasync Report 2015: Evaluating timed playback of HTML5 media," Norut Northern Research Institute, Tech. Rep. 28, Dec 2015. [Online]. Available: https://hdl.handle.net/11250/2711974

[24] I. M. Arntzen and N. T. Borch, "Data-independent Sequencing with the Timing Object: A JavaScript Sequencer for Single-device and Multi-device Web Media," in *Proceedings of the 7th International Conference on Multimedia Systems*, ser. MMSys '16. New York, NY, USA: ACM, 2016, pp. 24:1–24:10.

[25] I. M. Arntzen, N. T. Borch, and C. P. Needham, "The Media State Vector: A Unifying Concept for Multi-device Media Navigation," in *Proceedings of the 5th Workshop on Mobile Video*, ser. MoVid '13. New York, NY, USA: ACM, 2013, pp. 61–66.

[26] I. M. Arntzen, N. T. Borch, and F. Daoust, "Media Synchronization on the Web," *MediaSync: Handbook on Multimedia Synchronization*, pp. 475–504, 2018.

[27] I. M. Arntzen. (2015) Timingsrc. Multi-device Timing for Web. [Online]. Available: https://webtiming.github.io/timingsrc

[28] (2005) Unity real-time development platform. Unity Technologies. [Online]. Available: https://unity.com/

[29] (1998) Unreal engine. Epic Games. [Online]. Available: https://www.unrealengine.com/

[30] (1985) Midi. musical instrument digital interface. [Online]. Available: https://midi.org/

AI-Assisted Academic Writing: A Comparative Study of Student-Crafted and ChatGPT-Enhanced Critiques in Ubiquitous Computing

Edward R Sykes
School of Computer Science-College of Engineering and Physical Sciences,
University of Guelph, Guelph, Ontario, Canada, N1G 2W1

***Abstract*—This study examines the impact of Large Language Models (LLMs), such as ChatGPT, on the development of academic critique skills among fourth-year Computer Science undergraduates enrolled in a Ubiquitous Computing course. The research systematically evaluates the differences between student-authored critiques and those revised with the aid of ChatGPT, utilizing established readability metrics such as the Flesch-Kincaid Grade Level, Flesch Reading Ease, and Gunning Fog Index. The findings highlight the potential of AI to enhance readability and analytical depth, while also revealing challenges related to dependency, academic integrity, and algorithmic bias. These results extend implications across learning sciences, pedagogy, and educational technology, providing actionable insights into leveraging AI to augment traditional learning methods and enhance critical thinking and personalized education.**

Keywords*—*AI in higher education; AI in academic writing; readability metrics; LLM; ethical considerations

I. Introduction

Generative AI, encompassing technologies such as Large Language Models (LLMs) like ChatGPT, and advanced image generation tools, is profoundly reshaping various facets of the digital and real world. As educators and researchers, it is essential to stay abreast of these advancements to enhance our teaching methods and pedagogical tools, thereby enriching both our own and our students' skillsets.

This research delves into the specific application of LLMs, focusing on ChatGPT, to assess its impact on developing academic critique skills among Computer Science undergraduates enrolled in a fourth-year Ubiquitous Computing course. The core objective of this study is to evaluate and discern the differences between student-authored critiques and those augmented by ChatGPT's assistance.

Through this investigation, we aim to highlight the potential of LLMs to bolster students' critical thinking and writing capabilities. This paper seeks to provide valuable insights into how AI tools can be seamlessly integrated within educational frameworks. Additionally, we explore the transformative role of AI-driven tools in supporting student learning and enhancing the personalized and engaging nature of academic critique processes.

The remainder of this paper is structured as follows: The next section offers a background on Generative AI, focusing on the Transformer architecture, ChatGPT, its applications in education, and readability metrics. In Section III, we outline the methods employed to investigate the effectiveness of ChatGPT as a collaborative tool for supporting technical critique writing. The findings from our quantitative and qualitative analysis are presented in Section IV. Section V reflects on these findings within a broader context, and Section VI concludes the paper while highlighting potential future directions for this line of research.

II. Background

The rapid evolution of Generative AI has profoundly transformed technological capabilities, significantly influencing societal interactions, business processes, and educational methodologies. This section is divided into three main parts. The first part delves into the key technologies that have driven this transformation, highlighting their impact and implications. The second part presents an overview of current applications of Generative AI in the context of education, exploring how these innovations are being integrated into teaching and learning environments. The last section presents an overview of various readability metrics commonly used in the assessment of text.

A. Foundations of Generative AI

Before 2017, Natural Language Processing (NLP) relied heavily on architectures like Constitutional Neural Networks, Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory networks (LSTMs) [1]. These architectures were adept at processing sequences and were widely used for various NLP tasks. However, they often struggled with long-range dependencies and were computationally intensive due to their sequential nature, limiting their scalability and performance on larger datasets [1].

NLP took a major leap forward in 2017 with the introduction of the Transformer model by Vaswani et al. [2]. It marked a paradigm shift in sequence modeling, emphasizing the importance of self-attention mechanisms [2]. Essential for developing models that require a complex understanding of sequence data, such as those used in real-time language translation and interactive conversational agents, the Transformer architecture has become a core component of many state-of-the-art AI systems, influencing advancements across numerous fields including healthcare, finance, and autonomous vehicles.

Following the Transformer's success, Google AI's BERT (Bidirectional Encoder Representations from Transformers)

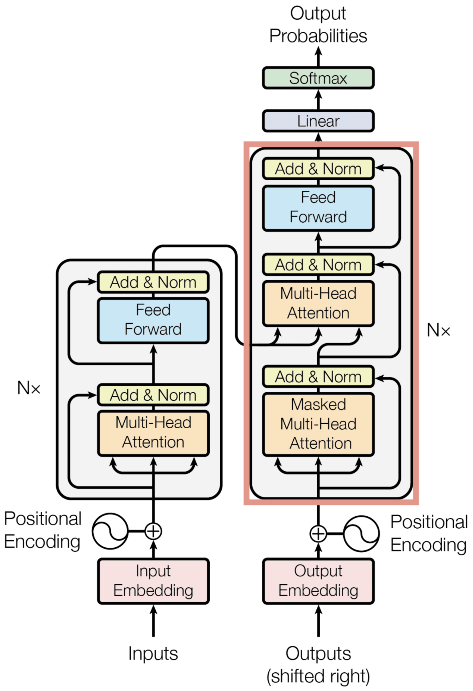


Fig. 1. The Transformer – model architecture [2].

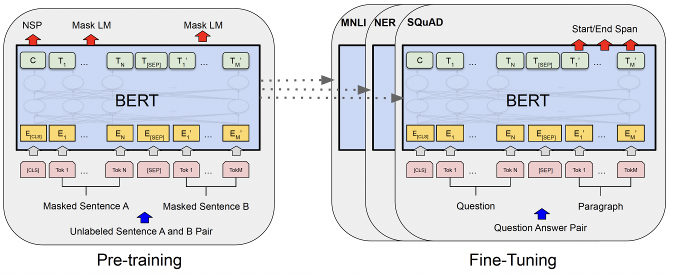


Fig. 2. BERT architectural model [3].

emerged as a significant advancement, leveraging the Transformer's architecture (see Fig. 1) to enhance language understanding [3]. BERT revolutionized natural language understanding by employing a transformer-based mechanism that processes words in the context of the entire sentence, rather than in isolation [3]. This breakthrough has led to substantial improvements in a range of language processing tasks, including translation, text summarization, and sentiment analysis. BERT's architecture leverages a stack of transformer blocks that feature two key components: multi-head self-attention mechanisms and fully connected feed-forward networks. This architecture enables the model to capture complex word relationships and contextual nuances across different parts of the text, facilitating more effective learning and prediction capabilities compared to previous models. BERT is also trained on a large corpus of text in a self-supervised manner using two tasks: masked language modeling and next sentence prediction, which help improve its language understanding. Fig. 2 presents the BERT architectural model, illustrating its deep neural network structure and attention mechanisms that contribute to its powerful performance.

Building on BERT's architecture, RoBERTa (Robustly Optimized BERT Approach) by Facebook AI modified key training methodologies to optimize performance [4]. Specifically, RoBERTa is trained with dynamic masking, full-sentences without NSPloss, large mini-batches, and a larger byte-level BPE [4]. Additionally, RoBERTa includes two other important components that were under-emphasized in the BERT architecture, namely: (1) the data used for pretraining, and (2) the number of training passes through the data. The RoBERTa model is trained on more data, with larger batches and longer sequences. As a result, RoBERTa offers a significant improvement in its language understanding capabilities, pushing the boundaries of what AI can comprehend and respond to, thereby setting new standards for model robustness and accuracy in language tasks [4].

Since 2019, Hugging Face has emerged as a pivotal player in democratizing AI through the development and maintenance of the Transformers library, which provides a vast collection of pre-trained models for a variety of NLP tasks [5]. This open-source library has enabled researchers and practitioners to easily implement cutting-edge models, fostering innovation and accelerating the adoption of AI technologies across industries. Currently, there are over 200 different transformer models available from Hugging Face: https://huggingface.co/docs/transformers/index.

The Generative Pre-trained Transformer (GPT) series, developed by OpenAI, includes notable releases such as GPT-2 in 2019, GPT-3.5 in 2022, and GPT-4 in 2023. Each iteration has progressively expanded the scale and capabilities of language models [6], [7]. With each version, there has been a significant leap in sophistication; for instance, GPT-4 is capable of producing text that closely mimics human writing across a wide range of genres and styles. These advancements underscore the creative potential of AI in generating coherent and contextually relevant text. However, they also bring to light critical ethical considerations, including concerns over misinformation, copyright issues, and the autonomy of AI-driven content creation.

Fig. 3 presents the architecture of the Generative Pre-trained Transformer. The GPT architecture is built upon the foundation of the Transformer model, which utilizes layers of self-attention mechanisms to process text [5]. At the core of the GPT architecture, as illustrated in the figure, is a series of Transformer blocks stacked on top of each other. Each block contains two main sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network.

Input embeddings, which convert tokens (words or pieces of words) into vectors of numbers, are first modified by positional encodings to retain the sequence order of the input text. These embeddings then pass through the Transformer blocks, where the self-attention mechanism allows the model to weigh the importance of different words relative to each other, regardless of their position in the text. This is followed by normalization and feed-forward layers that help in refining the representation with nonlinear transformations.

Each attention head in the multi-head attention layer computes an attention score, which represents how much focus to place on other parts of the input sentence as each word

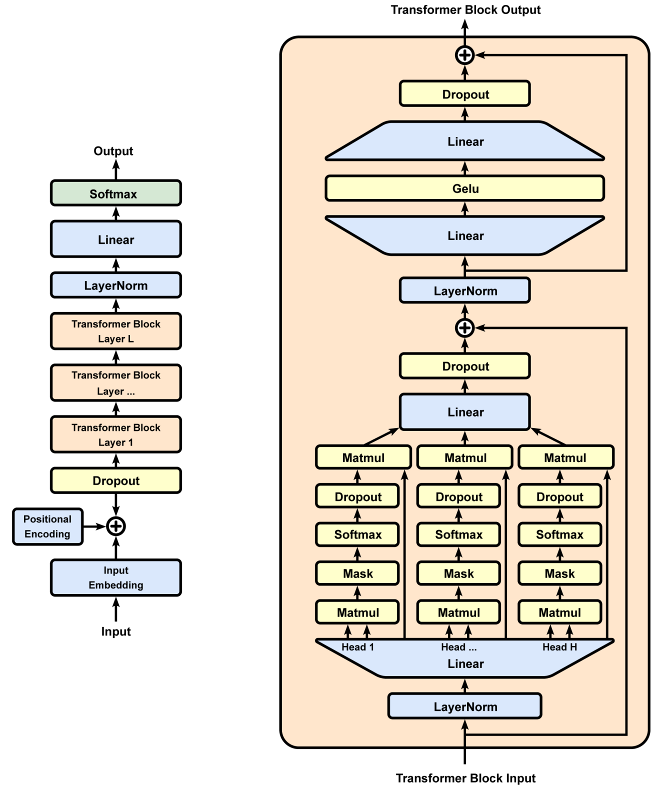


Fig. 3. The GPT Architecture [7].

is processed. The attention outputs are then combined and linearly transformed into the expected dimensions. Dropout layers are incorporated to prevent overfitting by randomly omitting subsets of features during training. The entire process within a Transformer block is designed to be differentiable so that it can be efficiently trained using gradient descent-based optimization.

This architecture enables GPT to generate text by predicting the next word in a sentence based on the words that came before it, learning to generate coherent and contextually relevant language over time. This capability makes GPT highly effective for a range of applications from automated content generation to sophisticated conversation simulations.

In 2020 Facebook AI Research published a paper on Retrieval-Augmented Generation (RAG) which introduces an approach that combines the generative capabilities of models like GPT with the retrieval of factual information during the generation process [8]. This methodology enhances the model's ability to produce relevant and accurate responses by dynamically retrieving context from a vast corpus of data, representing a significant advancement in efforts to bridge the gap between human-like understanding and AI output. Some examples of RAG models include:

1) Call centre agent support: Call centre agents require extensive knowledge of potentially hundreds of products and services, as well as commonly occurring product issues and their resolution. RAG solutions could assist agents in quickly finding answers to client requests.

2) Customer chatbots: RAG is strong enabler for creating customer-facing chatbots to answer frequently asked questions. Combining the natural language abilities of LLMs and the enterprise-specific responses of RAG can deliver a compelling, conversational customer experience.

3) Support / helpdesk: Similar to call centre agents, IT operations and support personnel require deep knowledge of the configuration of complex systems deployments along with knowledge of common and previously seen issues and their resolution. RAG solutions could assist support personnel with quickly finding relevant answers to reported problems and observed issues.

In the context of education, RAG offers transformative potential for educational environments by enhancing personalization, efficiency, and interaction in learning processes [9]. By combining the generative capabilities of models like GPT with dynamic information retrieval, RAG can create tailored educational content, support research and writing, and enhance interactive tutoring systems [8]. RAG's ability to pull relevant data from extensive knowledge bases allows it to generate accurate, context-specific content, making it ideal for developing personalized learning materials and dynamic assessment tools. Additionally, its application in question-answering systems can provide students with precise and informative responses, thereby fostering a deeper understanding of the subject matter [8].

However, the integration of RAG into educational tools must be approached with caution, addressing potential challenges such as ensuring the accuracy and reliability of content, mitigating data biases, and upholding stringent privacy standards. Continuous collaboration with educators, careful dataset curation, and rigorous testing are essential to leverage RAG's capabilities effectively while maintaining educational integrity and compliance [9], [10].

B. ChatGPT in Education

ChatGPT, has been progressively integrated into educational contexts, showcasing potential across various teaching and learning activities. This section explores its primary applications and the emerging studies surrounding its efficacy and challenges in educational settings.

1) Automated content and assessment tools: ChatGPT excels in generating and customizing educational content such as quizzes, reading materials, and assignments. It also offers potential in preliminary assessments by providing feedback on written assignments, which can be particularly beneficial in managing large classes [11].

2) Tutoring and support: As an interactive tutor, ChatGPT responds to student inquiries, assists with homework, and explains complex concepts, offering personalized support outside traditional classroom settings. This 24/7 availability can significantly enhance student learning, especially in subjects requiring frequent practice or clarification, such as languages and sciences [12].

3) Enhancing engagement and language learning: In discussions, ChatGPT can stimulate engagement by posing challenging questions or introducing diverse viewpoints. For language learners, it serves as an invaluable practice tool, facilitating conversation in various languages, which helps improve linguistic fluency and cultural awareness [13].

4) Challenges and ethical considerations: Despite its benefits, the deployment of ChatGPT raises concerns regarding reliability, potential biases, and academic integrity. Misinformation, inherent biases from training data, and the potential for students to misuse essay-writing capabilities require careful consideration and regulation. Ensuring that these tools are used to complement traditional educational methods rather than replace them is crucial for maintaining the quality and integrity of education [14].

The integration of these advanced Generative AI technologies in educational settings offers unprecedented opportunities for enhancing teaching and learning. Educators are leveraging these tools to develop more engaging and interactive learning environments, tailor educational content to individual needs, and foster critical thinking skills [9], [10]. However, this integration also poses challenges, including ensuring the ethical use of AI in classrooms, protecting student data privacy, and maintaining academic integrity.

C. Readability Metrics

The following set of established readability metrics are important in the assessment of text difficulty, engagement, and appropriateness for educational content.

1) Readability metrics used in educational assessments: Understanding the readability of educational content is crucial for tailoring materials to appropriate comprehension levels. This study employs several established readability metrics to evaluate the text complexity and accessibility of student-generated critiques. Below is a detailed explanation of each metric and its significance:

a) Flesch-kincaid grade level: Estimates the U.S. school grade level needed to understand the text. Lower scores indicate easier readability. A score of 12.0, for instance, suggests that the text is suitable for twelfth graders or equivalent [15]. The Flesch-Kincaid grade level is calculated with the following formula:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59 \tag{1}$$

b) Flesch reading ease: Evaluates text simplicity based on the average sentence length and the average number of syllables per word. Scores range from 0 to 100, with higher scores indicating easier readability. For example, texts scoring between 60 and 70 are considered suitable for standard reading, while a score in the range of 10.0 - 30.0 is ranked at the ‘College graduate’ level, typically very difficult to read and best understood by university graduates [16]. The Flesch Reading Ease score is calculated as follows:

$$206.835-1.015\left(\frac{\text{total words}}{\text{total sentences}}\right)-84.6\left(\frac{\text{total syllables}}{\text{total words}}\right) \tag{2}$$

c) Dale-chall readability score: Uses a list of familiar words to assess the grade level. Higher scores indicate more challenging text. A score of 9.0-10.0 suggests that the text is best understood by college-level readers [17]. The Dale-Chall readability score is calculated with the following formula:

$$0.1579\left(\frac{\text{difficult words}}{\text{words}}\times 100\right)+0.0496\left(\frac{\text{words}}{\text{sentences}}\right) \tag{3}$$

d) Automated Readability Index (ARI): Uses characters per word and words per sentence to estimate the grade level required for comprehension. A score of 13.0 indicates that the text is suitable for college freshmen [18]. The ARI score is calculated:

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right) \tag{4}$$

e) Coleman liau index: Estimates the U.S. school grade level necessary to understand the text, based on characters per word and words per sentence. A score of 11.0-12.0 indicates high school senior level complexity [19]. The Coleman–Liau index is calculated with the following formula:

$$CLI = 0.0588\cdot L - 0.296\cdot S - 15.8 \tag{5}$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

f) Gunning fog index: Estimates the years of formal education needed to understand the text. A score of 16.0 suggests college graduate level difficulty [20]. The Index is calculated as:

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right)+100\left(\frac{\text{complex words}}{\text{words}}\right)\right] \tag{6}$$

g) Linsear write formula: Calculates the U.S. grade level based on sentence length and the number of easy or difficult words. A score of 8.0 means the text is suitable for eighth graders [21]. (See [21] for the algorithm to compute the Linsear Write value.)

h) SMOG index (Simple measure of gobbledygook): Estimates the years of education needed to understand a text based on the number of polysyllabic words. A score of 17.0 implies graduate-level readability [22]. SMOG is calculated using this formula:

$$1.043\sqrt{\text{num of polysyllables}\times\frac{30}{\text{num of sentences}}}+3.1291 \tag{7}$$

i) SPACHE score: Specifically designed for early readers, assessing sentence length and word familiarity. A score suitable for first to third graders would typically be below 4.0 [23]. This instrument is not suitable for upper level undergraduate or graduate students.

These readability metrics collectively provide insights into the accessibility of written content, ensuring that educational materials are appropriately challenging yet understandable for the intended audience.

2) Additional readability and quality metrics: Beyond traditional readability metrics like the Flesch Reading Ease and Linsear Write Scores, other computational assessments such as the Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Metric for Evaluation of Translation with Explicit Ordering (METEOR) provide a broader evaluation of text quality, especially in contexts involving generative AI. BLEU measures the precision of generated text against reference texts by comparing the overlap of phrases and their order, making it ideal for assessing translation accuracy and content generation tasks in educational tools. This metric is widely used in machine translation to quantify how closely machine-generated text resembles human-like translations. Similarly, ROUGE is crucial for evaluating the coverage and recall of summaries produced by AI, ensuring that essential content is not omitted. It compares the extent to which the generated summaries capture the content present in a set of reference summaries, which is particularly useful in the evaluation of text summarization systems. Meanwhile, METEOR enhances evaluation by incorporating synonymy and stemming, alongside exact word matching, providing a balanced measure of fluency and intent preservation in generated text. Unlike BLEU, METEOR is designed to align more closely with human judgment by accounting for the flexibility in language use. Collectively, these metrics help in assessing the suitability of AI-generated educational content, aligning it with pedagogical goals and learner needs. Their application ensures that educational tools powered by AI not only generate content that is factually accurate but also presented in a manner that is understandable and engaging for students.

In summary, both BLEU and METEOR are traditionally utilized in machine translation to evaluate the alignment of translated text against one or more reference texts. These metrics quantify the extent of word and phrase overlap in the machine-generated text relative to the reference texts. Meanwhile, ROUGE assesses the quality of summaries by measuring the overlap between a generated summary and reference summaries, focusing particularly on the recall of essential content. However, in this study, there are no *reference texts* available that would typically be required for these metrics to function effectively. Consequently, traditional readability metrics, which do not rely on reference texts, are deemed most appropriate for evaluating the critiques created by the students.

In the next section, we present the methodology supporting our investigation into the potential of LLMs to bolster students' critical thinking and writing capabilities.

III. Methodology

This comprehensive study was implemented during the Fall 2023 semester with fourth-year undergraduate students enrolled in the Ubiquitous Computing (UbiCom) course in the Computer Science department[1]. The methodology included multiple components designed to rigorously evaluate the effectiveness of AI-assisted academic critiques.

A. Participants

Participants were recruited through convenience sampling through advertisements throughout the Computer Science Club, course postings, and other mediums available at the education institution. There were a total of 22 participants involved in this study all in their final year of study in their Honours Bachelor of Computer Science baccalaureate degree. There were 5 females and 17 males; the minimum age was 21, the mean was 25, and the maximum age was 31.

B. Detailed Process of Critique Assignment

Each week, students were assigned two academic papers centred around pivotal UbiCom topics such as Smart Homes, Smart Cities, IoT, and Wearable Technology. These topics were chosen to ensure that students were exposed to diverse applications and theoretical advancements within the field of Ubiquitous Computing. The selection process involved curating papers that varied in complexity and scope, providing a robust testbed for critique development. Seminal papers were selected, such as, Mark Weiser's 'The computer for the 21st century' in 1999 [24]. Mark Weiser is often referred to as the father of ubiquitous computing [7].

> *... Now we are in the personal computing era, person and machine staring uneasily at each other across the desktop. Next comes ubiquitous computing, or the age of calm technology, when technology recedes into the background of our lives.* —Mark Weiser

Other papers provided a survey of the state-of-the-art in a specific area (e.g., Smart Cities). For example, 'Systematic literature review of context-awareness applications supported by smart cities' infrastructures' [25].

C. Baseline and AI-Assisted Critique Process

Initially, critiques were written individually, and independently, without the assistance of any Generative AI assistance (e.g., ChatGPT, Bing, etc.) to establish a baseline for each student's analytical and writing abilities. Students were then asked to write critiques with the assistance of ChatGPT, following a structured well-defined methodology. Students were also given a 1-hour training session on effective prompt engineering and how to objectively assess ChatGPT's responses. This session aimed to empower students with the skills needed to elicit detailed and specific feedback from ChatGPT, enhancing their ability to refine their arguments and writing clarity.

D. Continuous Feedback and Iteration

A critical component of the methodology was the continuous feedback mechanism. After each AI-assisted critique, students received personalized feedback from both the course instructor and the AI, highlighting areas of improvement and success. This feedback loop was essential for guiding students' development over the semester and for refining the use of AI in the critique process. The students were asked to elicit feedback from ChatGPT at most 3 times. Please see the Appendix for the full details of the assignment.

E. Comprehensive Evaluation Metrics

The assignments were evaluated using a combination of quantitative and qualitative metrics, as described below.

1) Quantitative metrics: The core metrics were the readability tests. A suite of these tests were applied to each critique, providing a multifaceted view of how readability evolved with AI integration. This included advanced readability formulas that assessed not only text difficulty but also engagement and grade-appropriateness of the content.

[1] This research was approved by Sheridan's Research Ethics Board No. 2023-10-001-022.

Additional metrics collected and analyzed included the critiques' length, structure, and complexity changes from baseline to AI-assisted versions. The specific readability metrics computed were: Flesch-Kincaid Score, Flesch Reading Ease, Dale Chall Readability Score, ARI Score, Coleman Liau Index Score, Gunning Fog Score, Linsear Write Score, and SMOG Score.

a) Statistical analysis: Statistical analyses were performed on the readability data collected. Standard descriptive statistics were computed. Furthermore, we aimed to determine if there were any statistical differences over time in the readability scores, both between and within the groups: 'student-authored only' and 'student-and-ChatGPT-co-authored' critiques. These comparisons were made by examining the weekly critiques for most of the term. Due to the non-parametric nature of the data, one-way repeated measures ANOVAs using the Friedman test were computed.

2) *Qualitative metrics:* Evaluations were further deepened by analyzing the critiques for argumentative depth, logical coherence, and the use of evidence, which were scored using a rubric developed specifically for this course.

a) Educational outcomes monitoring: Beyond the critique process, the study monitored broader educational outcomes, such as student engagement, perceived ease of completing assignments, and overall satisfaction with the learning process. Surveys and interviews were conducted at the beginning, middle, and end of the semester to capture students' attitudes towards the use of AI in their learning process.

b) Ethical considerations and bias monitoring: Given the use of AI in educational settings, the study also addressed ethical considerations, particularly concerning the dependence on technology and the potential for AI to introduce biases into the students' work. Measures were put in place to monitor and mitigate any adverse effects, ensuring that the AI's integration was both responsible and beneficial.

IV. Findings

This section presents the qualitative and quantitative findings from this study.

A. Quantitative Findings

The following section provides an overview of the key findings from using ChatGPT to support co-authoring of academic critiques based on the readability scores for 'student authored' vs. 'student and ChatGPT co-authored.' Table I presents a comprehensive report of the readability analysis including standard descriptive statistics for 'student authored' critiques for the entire term (i.e., weeks 1-8). Each entry represents the group mean for that specific week for the respective readability metric.

It can be seen that the readability generally decreased as the term went on. This was particularly evident for the Flesch Reading Ease which started at 34.74 and declined to 24.32 at the end of the term (week 8). When referencing the mean, the readability levels across the entire group and the term, were: Flesch-Kincaid Score: 14.01 (college/university level), Flesch Reading Score: 30.24 (college/college graduate), Dale Chall: 11.38 (graduate level), ARI: 14.72 (college level), Coleman-Liau: 14.86 (graduate level), Gunning Fog: 16.21 (college senior), Linsear: 15.44 (college senior), and SMOG: 15.39 (undergraduate).

Table II presents the readability analysis including standard descriptive statistics for 'student and ChatGPT co-authored' critiques for the term.

A similar pattern emerged as in Table I. The Flesch Reading Ease started at 26.88 and declined to 15.50 by week 8. Using the mean of the readability scores across the group yielded the following results: Flesch-Kincaid Score: 14.89 (college/university level), Flesch Reading Score: 23.46 (college graduate), Dale Chall: 11.97 (graduate level), ARI: 15.65 (college level), Coleman-Liau: 16.17 (graduate level), Gunning Fog: 17.16 (college senior), Linsear: 15.87 (college senior), and SMOG: 16.00 (undergraduate).

Fig. 4 presents the weekly 'student authored' readability score analysis over the term. The most obvious pattern is the Flesch Reading score which shows a general decreasing trend throughout the term. The other readability scores were relatively consistent throughout the term.

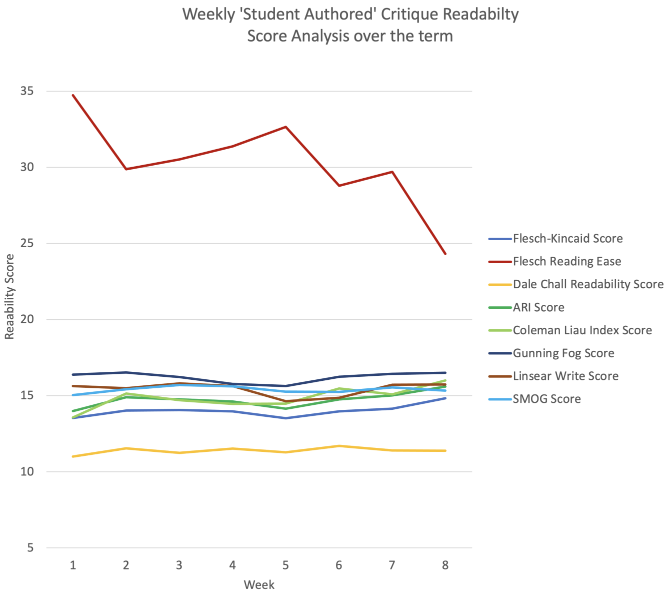


Fig. 4. weekly 'student authored' critique readability score analysis over the term.

Fig. 5 displays the weekly readability score analysis for critiques co-authored by students and ChatGPT over the term. Similar to the trends observed in Fig. 4, the most notable pattern is the general decline in the Flesch Reading Ease score throughout the term. Other readability metrics remained relatively stable over the period.

Statistical Analysis and ANOVA Results Prior to performing the ANOVAs, we verified the necessary assumptions to ensure the appropriateness of the statistical models. These assumptions included the independence of observations, normality of the data distributions, and homogeneity of variances. The Shapiro-Wilk test was used to confirm normality [26], and Levene's test was applied to assess the homogeneity of variances [27].

TABLE I. Comprehensive Readability Metrics with Standard Descriptive Statistics for Student Authored Critiques (weeks 1-8)

Week	Flesch-Kincaid Score	Flesch Reading Ease	Dale Chall Readability Score	ARI Score	Coleman Liau Index Score	Gunning Fog Score	Linsear Write Score	SMOG Score
1	13.54	34.74	11.00	13.98	13.57	16.37	15.63	15.04
2	14.02	29.87	11.53	14.89	15.14	16.52	15.49	15.41
3	14.06	30.52	11.25	14.76	14.70	16.23	15.80	15.71
4	13.97	31.38	11.51	14.62	14.45	15.77	15.63	15.61
5	13.52	32.65	11.28	14.14	14.47	15.63	14.63	15.26
6	13.97	28.79	11.70	14.75	15.47	16.24	14.85	15.24
7	14.15	29.69	11.40	15.03	15.09	16.44	15.72	15.53
8	14.83	24.32	11.39	15.59	15.99	16.51	15.73	15.34
Min	13.52	24.32	11.00	13.98	13.57	15.63	14.63	15.04
Max	14.83	34.74	11.70	15.59	15.99	16.52	15.80	15.71
Mean	14.01	30.24	11.38	14.72	14.86	16.21	15.44	15.39
Std Dev	0.41	3.05	0.21	0.50	0.74	0.34	0.44	0.22

TABLE II. Comprehensive Readability Metrics with Standard Descriptive Statistics for Student and ChatGPT co-Authored Critiques (weeks 1-8)

Week	Flesch-Kincaid Score	Flesch Reading Ease	Dale Chall Readability Score	ARI Score	Coleman Liau Index Score	Gunning Fog Score	Linsear Write Score	SMOG Score
1	14.65	26.88	11.65	15.26	15.12	17.44	16.33	16.40
2	15.02	22.79	12.06	15.88	16.38	17.65	16.18	16.14
3	15.26	20.96	12.18	16.07	16.64	17.48	16.24	16.59
4	14.90	24.95	12.00	15.62	15.62	16.82	16.29	16.29
5	14.01	27.51	11.75	14.75	15.77	16.06	14.44	15.39
6	14.52	24.61	12.09	15.23	16.12	16.96	15.16	15.59
7	14.71	24.49	11.93	15.54	16.11	16.91	15.73	15.96
8	16.06	15.50	12.09	16.85	17.56	18.00	16.55	15.65
Min	14.01	15.50	11.65	14.75	15.12	16.06	14.44	15.39
Max	16.06	27.51	12.18	16.85	17.56	18.00	16.55	16.59
Mean	14.89	23.46	11.97	15.65	16.17	17.16	15.87	16.00
Std Dev	0.60	3.83	0.18	0.63	0.73	0.60	0.72	0.43

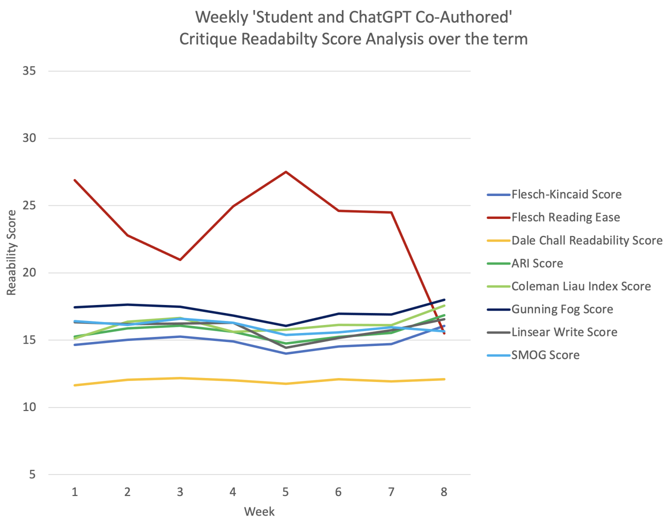


Fig. 5. Weekly 'student and chatGPT Co-Authored' critique readabilty score analysis over the term.

After confirming these assumptions, ANOVAs were conducted for each readability metric to determine if there were statistically significant differences between critiques authored solely by students and those co-authored with ChatGPT. The results are as follows:

- *Flesch-Kincaid Score*: Significant difference; $F(1, 14) = 11.974$, $p = 0.003$.
- *Flesch Reading Ease*: Significant difference; $F(1, 14) = 15.356$, $p = 0.0015$.
- *Dale Chall Score*: Significant difference; $F(1, 14) = 34.994$, $p < 0.0001$.
- *ARI Score*: Significant difference; $F(1, 14) = 10.558$, $p = 0.005$.
- *Coleman Liau Index*: Significant difference; $F(1, 14) = 12.609$, $p = 0.003$.
- *Gunning Fog Score*: Significant difference; $F(1, 14) = 15.076$, $p = 0.001$.
- *Linsear Write Score*: No significant difference; $F(1, 14) = 2.050$, $p = 0.174$.
- *SMOG Score*: Significant difference; $F(1, 14) = 12.707$, $p = 0.003$.

The results indicate that for all metrics except the Linsear Write Score, there were statistically significant differences between the student-authored and ChatGPT co-authored critiques over the eight weeks. These findings suggest that the integration of AI like ChatGPT in the writing process significantly influences the readability and possibly the quality of student critiques.

Interestingly, the decrease in readability scores over time, especially in measures like the Flesch Reading Ease, might reflect a transition towards more complex academic language. This trend could be attributed to the students' exposure to high-quality academic literature and their advancement in understanding and synthesizing complex concepts. It appears that as students engaged with seminal works and sophisticated material, their ability to emulate academic rigour in their own writing improved, leading to the production of text that, while potentially more challenging for lay readers, aligns more closely with fourth-year undergraduate and graduate-level standards. This evolution in writing style underscores the effectiveness of AI tools in fostering higher-order cognitive skills, including critical analysis and academic writing prowess.

1) Individual student performance observations: Deeper investigations on specific participants yielded some interesting results. The following examples illustrate these findings. Participant #23: Flesch Reading Ease: In one critique, the score improved from a very low 4.728 to a higher 10.666. This is a significant improvement in readability, suggesting that the co-authoring process made the document more readable. Gunning Fog Score: In the same document, the score changed from 19.783 to 18.669 with the assistance from ChatGPT. This indicates a reduction in sentence complexity, contributing to better readability.

Participant #22: Flesch Reading Ease: For one critique, the score dropped slightly from 36.667 to 32.726. While this is a decrease, it's still within a range that suggests good readability, possibly indicating a more balanced approach to complexity and readability in the co-authored version.

Participant #3: Flesch Reading Ease: Improved from 23.316 to 27.886 in one document, indicating an increase in readability. Linsear Write Score: Decreased from 13.115 to 12.538, which points towards simpler sentence structure.

B. Qualitative Observations

In several cases, the co-authored documents show an improvement in readability scores, suggesting that ChatGPT can help in making complex academic content more accessible.

Conversely, there are instances where the complexity of the documents increased, possibly reflecting a deeper level of analysis or more advanced vocabulary and sentence structures due to the academic nature of the critiques. Some documents show a balance between maintaining academic rigour and ensuring readability, which is crucial in educational settings.

V. Discussion

The results of this study underscore significant implications for integrating AI, specifically Large Language Models like ChatGPT, into educational settings. The data reveal a discernible trend of fluctuating readability scores throughout the semester, suggesting that while AI tools can enhance the readability of academic critiques, their effectiveness may vary based on the complexity of the assignments and the adaptability of students.

A. Interpretation of Quantitative Findings

The quantitative analysis shows that readability scores, such as the Flesch Reading Ease and Flesch-Kincaid Grade Level, generally declined as the semester progressed. This trend could reflect the increasing complexity of the topics covered in the assignments, requiring a higher cognitive load from students, which might impact their writing clarity when attempting to articulate complex ideas.

Additionally, the decline in readability might also be attributed to a natural evolution in students' writing abilities. Exposure to complex texts is crucial in academic settings as it significantly correlates with improved analytical skills and academic writing proficiency, as discussed in the study by Graesser et al. [28].

This perspective is further supported by cognitive load theory. Engaging with sophisticated texts can enhance cognitive and analytical skills in higher education students, supporting the notion that challenging materials promote academic rigour [29]. The methodology of this study, through its iterative approach, encouraged students to use ChatGPT wisely by reflecting on the AI's suggestions and revising their critiques accordingly. This approach may have impacted the development of cognitive skills such as critical thinking and academic writing.

Furthermore, a study by O'Sullivan et al. (2020) demonstrated the impact of AI tools on learning, suggesting that tools like ChatGPT can foster critical thinking and academic writing skills [30].

Moreover, the statistical analysis employing ANOVA revealed significant differences in readability metrics between critiques authored independently by students and those co-authored with ChatGPT. This result highlights AI's potential to distill complex articles into more accessible academic language, thereby enhancing the accessibility of the critiques. However, it also emphasizes the need for careful integration of these tools to preserve content depth and ensure analytical rigour.

A potential concern with the observed decrease in readability scores could be an over-reliance on AI tools, which might lead to less critical engagement with the material and result in more convoluted expressions in student writings. However, this issue was not observed by the instructor during class work, class discussions, or in the grading of the students' critiques.

B. Qualitative Insights and Student Engagement

Qualitatively, the data indicated significant variation in individual student experiences, with some students demonstrating marked improvements in writing clarity and others showing increased complexity in their expression. This variance suggests a need for tailored approaches in AI integration that consider the individual profiles and needs of students.

C. Challenges and Ethical Considerations

The study also highlights several challenges and ethical considerations. The potential dependency on AI tools raises concerns about the ability of students to develop independent critical thinking skills. Balancing the use of AI for educational benefits while ensuring that students remain the primary agents in their learning processes is crucial. Moreover, ethical issues related to data privacy, bias in AI algorithms, and the authenticity of student work require ongoing attention and the development of robust regulatory frameworks.

D. Educational Implications and Future Directions

This research contributes to the ongoing discourse on the role of AI in education by providing empirical evidence of its benefits and limitations. Future research could explore the long-term impacts of AI integration on student learning outcomes through longitudinal studies. Additionally, investigating diverse educational settings and varied student demographics could help generalize the findings and tailor AI educational tools more effectively.

Overall, AI offers substantial opportunities for enhancing educational practices, but its integration must be managed judiciously to complement traditional learning methods and support the holistic development of students' critical and analytical skills.

VI. Conclusion

This study explored the impact of integrating ChatGPT into academic critique writing, highlighting notable improvements in the writing quality and analytical depth across various metrics. The results affirm the potential of AI tools to not only enhance traditional educational methods but also to personalize and enrich learning experiences for both educators and students.

However, the study's limitations, including its small sample size and the demographic homogeneity of the participants, caution against broad generalizations of these findings. The short duration of the study also limits insights into the long-term effects of AI integration in educational settings. Future research should, therefore, expand these investigations across more diverse educational contexts and over longer periods to better understand the enduring impacts of AI on learning outcomes. It is also essential to address the ethical challenges associated with AI use in education, such as ensuring data privacy and managing the risk of academic dishonesty, through the development of robust ethical frameworks and guidelines.

In an effort to support ongoing research and foster an open academic environment, we have made the following resources available:

- our Jupyter notebook that utilizes Python for reading Microsoft Word documents (i.e., critiques) and computing readability metrics is available at https://bitbucket.org/ed_sykes_team/workspace/overview/. This notebook employs various Python libraries including *os*, *csv*, *aspose.words* (available at https://products.aspose.com/words/), *re*, the *Readability* framework (available at https://pypi.org/project/readability/), and *statistics* to facilitate analysis.
- An Excel workbook containing anonymized data from weeks 1-8, complete with standard descriptive statistics and complete ANOVA statistical analysis results, can be accessed here: https://bitbucket.org/ed_sykes_team/workspace/overview/.

A. Future Research

As AI continues to integrate into educational landscapes, it is imperative for educators, researchers, and policymakers to adopt a balanced approach to its use. This approach should aim to ensure that AI complements traditional teaching methodologies, enhancing educational outcomes while safeguarding the integrity and rigour of academic processes.

The sample size in this study was very small, comprising only 22 participants. Future work should include larger and more diverse samples to cross-validate the findings. Additionally, this study spanned only one semester; future research should consider extending the duration to several semesters or even academic years to assess the long-term effects on student learning outcomes. Conducting such longitudinal studies would yield a more comprehensive dataset.

Lastly, while this paper briefly addressed ethical questions about AI in education, further research is needed to elaborate on measures that could minimize potential biases and ensure academic integrity. Such efforts would pave the way for establishing practical guidelines for educators.

Acknowledgment

Special thanks to all the participants whose cooperation was essential for this research.

References

[1] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*, 2020.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[9] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An argument for AI in Education*. Pearson Education, 2016.

[10] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence In Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, 2019.

[11] U. Schmid, L. Goertz, and S. Radomski, "The impact of artificial intelligence on learning, teaching, and education," *Publications Office of the European Union*, 2019. [Online]. Available: https://data.europa.eu/doi/10.2760/382730

[12] V. Kumar, "The role of ai-powered feedback in education: A scalable approach to enhancing human learning," *Educational Technology Research and Development*, vol. 68, pp. 2939–2960, 2020.

[13] B. Smith and M. Johnson, "Harnessing artificial intelligence to enhance language learning: A review," *Language Learning & Technology*, vol. 25, no. 1, pp. 34–52, 2021. [Online]. Available: https://www.lltjournal.org/item/3098

[14] S. A. Jones and M. Buntins, "Ethical implications of ai in education: A review," *Journal of Research on Technology in Education*, vol. 54, no. 3, pp. 409–424, 2022.

[15] P. D. Farr JN, Jenkins JJ, "Simplification of flesch reading ease formula," *Journal of Applied Psychology*, vol. 35, no. 5, p. 333–337, 1951.

[16] R. Flesch, "A new readability yardstick," *J Appl Psychol*, vol. 32, no. 3, pp. 221–33, 1948, flesch, r Journal Article United States 1948/06/01 J Appl Psychol. 1948 Jun;32(3):221-33. doi: 10.1037/h0057532.

[17] E. Dale and J. S. Chall, *A Formula for Predicting Readability: Instructions*. Educational Research Bulletin, 1948.

[18] R. Senter and E. Smith, "Automated readability index," *AMRL-TR*, vol. 1, pp. 1–14, 1967.

[19] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, vol. 60, p. 283–284, 1975.

[20] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill, 1952, p. 36–37.

[21] G. R. Klare, "Assessing readability," *Reading Research Quarterly*, vol. 10, no. 1, pp. 62–102, 1974, linsear readability. [Online]. Available: http://www.jstor.org/stable/747086

[22] G. H. McLaughlin, "Smog grading—a new readability formula," *Journal of Reading*, vol. 12, no. 8, pp. 639–646, 1969.

[23] G. Spache, "A new readability formula for primary-grade reading materials," *The Elementary School Journal*, vol. 53, no. 7, pp. 410–413, 1953. [Online]. Available: http://www.jstor.org/stable/998915

[24] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, p. 3–11, jul 1999. [Online]. Available: https://doi.org/10.1145/329124.329126

[25] N. Pacheco Rocha, A. Dias, G. Santinha, M. Rodrigues, C. Rodrigues, A. Queirós, R. Bastardo, and J. Pavão, "Systematic literature review of context-awareness applications supported by smart cities' infrastructures," *SN Applied Sciences*, vol. 4, no. 4, p. 90, 2022. [Online]. Available: https://doi.org/10.1007/s42452-022-04979-0

[26] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[27] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, Ed. Stanford University Press, 1960, pp. 278–292.

[28] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-metrix: Providing multilevel analyses of text characteristics," *Educational Researcher*, vol. 40, no. 5, pp. 223–234, 2011.

[29] J. S. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books, 1995.

[30] I. O'Sullivan and L.-A. Perryman, "Collaborating with ai: Exploring networked learning perspectives," *Journal of Interactive Media in Education*, vol. 2020, no. 1, 2020.

Appendix

Weekly Assignment Overview

- ✓ Every week you need to read the required readings (selected articles)
- ✓ Write two 1-2 page critiques on any 2 of the selected articles for each topic
- ✓ Use MS word; double spaced, times new roman, font size 12
- ✓ Note: A sample critique document is provided on the course website.

Critique Paper Structure:

Title Page

A critique of "<title of paper> by <author(s)>"

By: <your name, student #, course code, etc.>

Summary: The section should summarize the purpose, methodology, and main findings of the paper.

Critique: This section should critically examine the paper, identify the pros and cons, limitations, and achievements. Your critique should be objective and factual. There are many sites that can guide you in writing a critique (e.g., University of Calgary, University of Michigan, etc.)

Questions: List 2-4 questions that you have about the paper. The intention is that you can raise these questions during the Discussion period after the Ubicom Topic Presenters give their talk.

Create two versions for each of your critiques:

1) *Student Authored Version:* Without any writing assistance, create a critique of the academic paper by yourself; and
2) *ChatGPT Co-authored version:* Using ChatGPT, ask it to review your critique and provide suggestions for improvement. Reflect on the prompt engineering sessions you had with the instructor and use these suggestions to revise your critque based on your judgement. Repeat this process up to 3 times.

Submission Instructions:

1) Submit these two versions to the course website using the following naming convention for your files:
   'Name_Week_X_Critique_1_Student_Authored_Only.docx'
   'Name_Week_X_Critique_1_Student_ChatGPT_co_authored.docx'

   'Name_Week_X_Critique_2_Student_Authored_Only.docx'
   'Name_Week_X_Critique_2_Student_ChatGPT_co_authored.docx'

2) Print out a hardcopy of critiques and bring them to class so that you can refer to it during the discussion portion of the class. At the end of the class, hand them to the professor.

A Natural Language Processing Model for the Development of an Italian-Language Chatbot for Public Administration

Antonio Piizzi[1], Donatello Vavallo[2], Gaetano Lazzo[3], Saverio Dimola[4], Elvira Zazzera[5]
Tempo S. R. L., Bari, Italy[1, 2, 3, 4]
Kad3 S. R. L., Fasano, Italy[5]

***Abstract*—Natural Language Processing models (NLP) are used in chatbots to understand user input, interpret its meaning, and generate conversational responses to provide immediate and consistent assistance. This reduces problem-solving time and staff workload and increases user satisfaction. There are both rule-based chatbots, which use decision trees and are programmed to answer specific questions, and self-learning chatbots, which can handle more complex conversations through continuous learning about data and user interactions. However, only a few chatbots have been developed specifically for the Italian language. *T*he development of chatbots for Public Administration (PA) in the Italian language presents unique challenges, particularly in creating models that can accurately understand and respond to user queries based on complex, context-specific documents. This paper proposes a novel natural language processing (NLP) model tailored to the Italian language, designed to support the development of an advanced Question Answering (QA) chatbot for PA. The core of the proposed model is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, enhanced with an encoder/decoder module and a highway network module to improve the filtering and processing of input text. The principal aim of this research is to address the gap in Italian-language NLP models by providing a robust solution capable of handling the intricacies of the Italian language within the context of PA. The model is trained and evaluated using the Italian version of the Stanford Question Answering Dataset (SQuAD-IT). Experimental results demonstrate that the proposed model outperforms existing models such as BIDAF in terms of F1-score and Exact Match (EM), indicating its superior ability to provide precise and accurate answers. The comparative analysis highlights a significant performance improvement, with the proposed model achieving an F1-score of 59.41% and an EM of 46.24%, compared to 49.35% and 38.43%, respectively, for BIDAF. The findings suggest that the proposed model offers substantial benefits in terms of accuracy and efficiency for PA applications.**

Keywords—Natural Language Processing; chatbot; BERT; transformer; Italian language

I. Introduction

A chatbot is an application that uses various techniques to understand user input, interpret its meaning, and generate responses in a conversational manner based on the user input. Implementing a chatbot as a support tool offers several advantages for both the end users and the businesses [1]. First, chatbots can provide immediate assistance to users by helping to reduce the time it takes to solve problems. Second, chatbots can provide 24-hour support, even outside normal business hours, with shorter resolution times and higher user satisfaction. In addition, chatbots can be used to identify common problems and trends through analysis of user questions, enabling the company to proactively address these issues and reduce the number of support requests. It is also cost-effective as it helps reduce the workload of IT staff and the need for additional staff or overtime. A chatbot also offers benefits to the business because it helps reduce the workload of IT staff and the need for additional staff or overtime. Chatbots can be classified into two main variants: rule-based and self-learning. Rule-based chatbots are developed to answer specific questions or perform actions based on predefined rules and logic because the responses are written in advance and correspond to a set of predefined questions or commands. However, these chatbots are limited to understanding complex natural languages and are unable to adapt to new situations. Therefore, chatbots based on Artificial Intelligence (self-learning chatbots) have been developed to overcome these issues. Self-learning chatbots learn from data and user interactions, gradually improving their ability to respond accurately. They can manage more complex conversations and adapt to new scenarios and user requests. However, they require initial training and are more complex to develop than rule-based chatbots. Complexity is to ensure the best response to each and every unpredictable user input. In the context of public administration (PA), a self-learning chatbot should be capable of understanding and interpreting documents provided by users as input in order to generate appropriate responses. For example, such a chatbot could handle requests that involve reading and interpreting official documents, offering precise and relevant answers. A chatbot developed with these requirements would represent a great advantage in terms of time as well as efficiency for public administration staff.

Despite the widespread use of chatbots in various languages, there is a significant gap in the development of advanced NLP models specifically designed for the Italian language, particularly in the context of Public Administration (PA). Most existing chatbots are either rule-based, which limits their ability to handle complex and varied user inputs, or they are designed for languages like English, for which extensive datasets and pre-trained models are available. However, Italian-language chatbots remain underdeveloped due to the lack of comprehensive datasets and specialized models that can accurately process and understand Italian text, especially within the specific and formal context of PA.

This research seeks to address this gap by proposing a novel NLP model based on the BERT architecture, tailored specifically for the Italian language. The model is designed to overcome the limitations of existing approaches by incorporating an encoder/decoder module and a highway network module to enhance the processing of Italian text. By training and evaluating the model on the SQuAD-IT dataset, this study aims to provide a robust solution that significantly improves the accuracy and efficiency of Italian-language chatbots in PA contexts, thereby filling a critical void in current NLP research. In particular, the proposed work presents an NLP model for the development of a QA self-learning chatbot able to read any document and provide relevant and specific responses to users. The proposed model architecture is based on the BERT [2] model architecture, with the addition of an encoder/decoder module and a highway network module to improve the filtering of the input text. Specifically, the encoder module takes as input a sequence of tokens and transforms them into a dense vector representation that captures the semantic and syntactic information of the text. The decoder module takes the vector representation provided by the encoder as input and generates a new sequence of text. The highway network filters irrelevant information before processing the last dense layers. Moreover, the proposed model is trained on the SQuAD-IT dataset to develop a chatbot specifically for the Italian language. Only a few works have used the Italian version of the SQuAD dataset, and the proposed work is intended to present an efficient and suitable architecture for developing an Italian-specific chatbot. The main contributions of the proposed work are summarized below.

1) A customized BERT model architecture with the addition of an encoder/decoder module and a highway network module has been proposed. The proposed model is developed to be integrated into an Italian-specific chatbot to improve the work of PA staff by reducing time and errors in processing and understanding documents.

2) The proposed model is trained and tested on the Italian version of the SQuAD dataset to evaluate the model's ability to process the Italian language.

3) The results of the experiments conducted on the SQuAD-IT dataset show that the proposed model has a good ability to provide exactly the expected answers. Moreover, a comparative analysis shows that the proposed model outperforms compared to other NLP models, such as BIDAF.

II. Related Work

Different types of self-learning chatbots have been developed based on the deep-learning models used. Three macro-categories can be identified: chatbots based on Convolutional Neural Networks (CNNs), chatbots based on Recurrent Neural Networks (RNNs), and chatbots based on hybrid models. CNNs are mainly used for pattern recognition in text data, such as sentences or paragraphs. Subsequently, sentence similarities between question-answer pairs are used to assess relevance and rank all candidate answers. In study [3] a CNN for learning an optimal representation of question-and-answer sentences has been proposed. The CNN encodes the correspondences between words to better acquire interactions between questions and answers, resulting in a significant increase in accuracy. The proposed CNN consists of two distributional sentence models based on convolutional neural networks (ConvNets) that map question-and-answer sentences into their distributional vectors, which are then used to learn their semantic similarity. In study [4], a model that considers both similarities and differences between question-and-answer by decomposing and composing lexical semantics on sentences has been proposed. Given a pair of sentences, the model represents each word as a low-dimensionality vector and computes a semantic correspondence vector for each word based on all the words in the other sentence. Then, based on the semantic correspondence vector, each word vector is decomposed into two components: similar and dissimilar. The similar components of all words are used to represent the similar parts of the sentence pair and the dissimilar components of each word are used to model the dissimilar parts explicitly. Subsequently, the similar and dissimilar components are composed into a single feature vector that is used to predict sentence similarity. In study [5], a CNN with a Siamese structure with two sub-networks processing the question and the candidate response has been proposed. The input is a sequence of words each of which is translated into its corresponding distributional vector, producing a matrix of sentences. Convolutional feature maps are applied to this matrix of sentences, followed by ReLU activation and simple max-pooling to achieve a representation vector for the query and candidate response. CNNs can have difficulty capturing long-term dependencies in text. Therefore, they are combined with RNNs or transformers to improve performance.

RNNs can process data sequences of variable length, making them ideal for text. In research [6], a bilateral multi-perspective matching (BiMPM) model has been proposed. The proposed model encodes the sentences through a bidirectional Long Short-Term memory Network (BiLSTM). The matching between encoded sentences is aggregated through a layer BiLSTM into a matching vector of fixed length. It is used in the last fully connected layer of the proposed model to make a decision. The approach in study [7] extends a long short-term memory (LSTM) proposed as a holographic dual LSTM (HDLSTM). HDLSTM is a unified architecture for both deep sentence modeling and semantic matching. RNNs are less efficient than CNNs and transformers in terms of parallelization and training speed. Therefore, hybrid models that combine the advantages of each model have been developed. RNNs have been combined with CNN [8], with attention mechanisms [9] or with transformer [10], [11]. Minjoon Seo et al. [12] have proposed a Bi-Directional Attention Flow network (BIDAF) that uses a bi-directional attention mechanism to obtain a query-sensitive context representation. The attention layer is not used to summarize the context paragraph into a vector of fixed size, but attention is computed for each time step, and the expected vector in each time step, together with the representations of the previous layers, can flow through the next modeling layer.

The introduction of the transformers in NLP models has led to advantages in efficiency, ability to capture long-range relationships, parallelization, and quality of representations. The transformer architecture allows higher parallelization during training and inference because it does not require sequential computation as in RNNs. This significantly reduces training time compared with RNNs. Therefore, using transformers represents one of the best choices to develop self-learning chatbots.

The proposed model is based on BERT architecture that use transformer to construct deep and bidirectional representations of words so as to capture complex contextual relationships. Moreover, the proposed model implements additional modules in the standard BERT architecture to increase the model's capabilities in providing correct answers.

III. Materials and Methods

The proposed model is based on an extended version of the BERT model by adding an encoder/decoder module and a highway network module. The proposed model is illustrated in Fig. 1. The proposed model consists of four main modules.

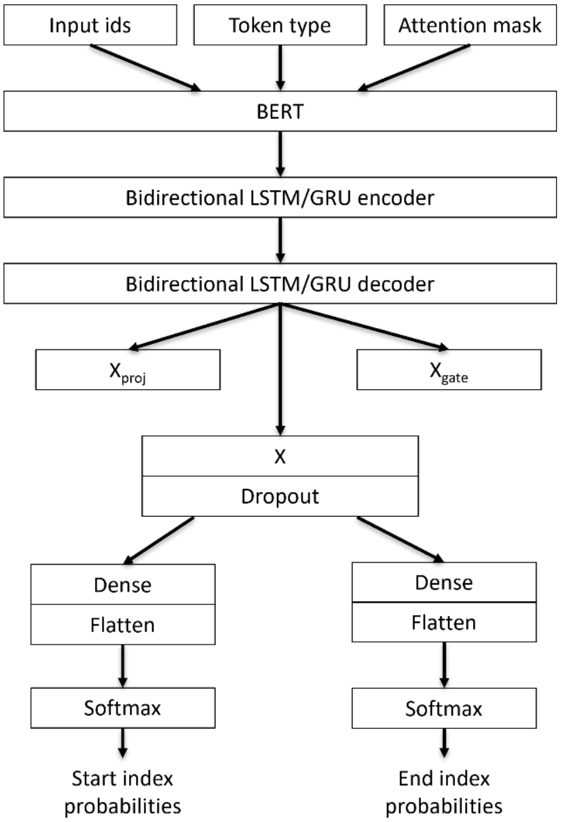


Fig. 1. Proposed model architecture.

A. Input Module

The first module consists of BERT architecture to encode the input into a vector representation that is then processed by the subsequent structures. BERT architecture can be represented as a multilayer bidirectional transformer encoder. BERT's pre-training is based on two different unsupervised tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, a portion of the words in the text are masked, and the model must predict them. In NSP, the model must determine whether two sentences appear consecutive in the original text. This pre-training makes the BERT model scalable (fine-tuned) for different tasks, such as QA. BERT takes as input the combinations of the question and context as a single embedded sequence. The input embeddings are the sum of the token embeddings and segment embeddings. Specifically, token embeddings represent the encoding of the question into an embedding vector, and segment embeddings represent vectors indicating the segment to which each token corresponds. The segment embeddings are used to distinguish between question and context in the input text. Let $x = [x_1, x_2, \dots, x_n]$ represents a sequence of input, and e_i represents an embedded achieved combining the token and segment embeddings for each x_i . The sequences of embeddings $E = [e1, e_2, \dots, e_n]$ is the input of the BERT module. The module BERT processes the embedding sequence through L transformer layers to obtain the output sequences $H^L = [h_1^L, h_2^L, \dots, h_n^L]$ where h_i^L is the hidden representation of $x_1$at the L level.

B. Encoder / Decoder Module

The encoder/decoder module consists of two sequential BiLSTM layers. The introduction of this module better captures the context and temporal sequence of words, thus improving the model's overall performance in understanding. Specifically, the BERT output $H^L = [h_1^L, h_2^L, \dots, h_n^L]$ is taken as input to encoder/decoder module to produce a new sequence of hidden representations $H^{BiLSTM} = \left[h_1^{BiLSTM}, h_2^{BiLSTM}, \dots, h_n^{BiLSTM}\right]$.

C. Filter Module

The highway network module aims to filter out irrelevant information before processing the last dense layers. Highway Network transformations are based on a linear combination between the non-linear transformation of the input and the original input following a gate function. The output of the Highway network module is defined in Eq. (1), where ∘ represents the element-by-element multiplication.

$$y_i = T\left(h_i^{BiLSTM}\right) \circ S\left(h_i^{BiLSTM}\right) + \left(1 - T\left(h_i^{BiLSTM}\right)\right) \circ h_i^{BiLSTM} \quad (1)$$

The linear transformation $T(\cdot)$ is defined by Eq. (2), where W_T and b_T are the weights and the bias of the gate function, respectively, and σ represents the sigmoid function.

$$T\left(h_i^{BiLSTM}\right) = \sigma\left(W_T h_i^{BiLSTM} + b_T\right) \quad (2)$$

The non-linear transformation $S(\cdot)$ is defined in Eq. (3), where W_S and b_S are the weights and the bias of the non-linear transformation, respectively, and $ReLU$ is the activation function.

$$S\left(h_i^{BiLSTM}\right) = ReLU\left(W_S h_i^{BiLSTM} + b_S\right) \quad (3)$$

D. Output Module

The output module consists of two fully connected layers with softmax activation function. The output module predicts the start and end positions of the response within the context following Eq. (4) and Eq. (5), respectively.

$$P_{start}(i) = softmax(W_{start}y_i + b_{start}) \quad (4)$$

where, W_{start} and b_{start} represent the weights and bias of the fully connected layer for the prediction of the start token.

$$P_{end}(i) = softmax(W_{end}y_i + b_{end}) \quad (5)$$

where, W_{end} and b_{end} represent the weights and bias of the fully connected layer for the prediction of the end token.

IV. Dataset

SQuAD-IT [13] dataset is a translated and adapted Italian version of the popular SQuAD dataset [14] SQuAD dataset has been developed to evaluate NLP models in English. The main advantage of this dataset is that it is realistic because humans crowdsourced it manually. It includes 536 English Wikipedia articles with more than 100,000 related question-answer pairs. Each crowd-worker was asked to answer up to five questions about a Wikipedia passage, highlighting the answer in the passage. Each question is associated with a segment of text within the article (context), and the answer is a subset of the context. Each question had several specific answers provided by different people. In SQuAD-IT, the texts, questions, and answers in SQuAD have been translated into Italian. The translation aims to maintain the same structure and content as the original dataset but ensures that the sentences are natural and grammatically correct in Italian. SQuAD-IT has been developed to evaluate NLP models that are not limited to English. The use of the SQuAD-IT dataset for the experiments enables the evaluation of the proposed model in understanding the Italian language to develop an Italian-specific chatbot.

V. Experimental Evaluation

Experiments have been conducted to evaluate the proposed model's performance in providing the correct answer in the QA task. The performance has been evaluated in terms of F1-score and Exact Match (EM). F1-score is the harmonic mean of accuracy and recall rate. In other words, the F1-score measures the overlap between the words in the predicted answer and the corresponding words in the correct answer (ground truth). The EM computes the percentage of correct answers generated by the model compared to the ground truth. F1-score and EM have been computed following Eq. (6) and Eq. (7), respectively.

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (6)$$

$$EM = \frac{\sum_{i=1}^{N} I(pred_i = truth_i)}{N} \quad (7)$$

In Eq. (7), N represents the number of test examples, $pred_i$ and $truth_i$ represent the predicted answer and the correct answer, respectively, and $I(\cdot)$ is a function that returns 1 if the condition is truth or 0 otherwise.

To evaluate the proposed model, the SQuAD-IT dataset has been divided following the train-test split ratio of 80:20, and the proposed model has been trained following the parameters detailed in Table I. The results of the proposed model in terms of F1-score and EM are shown in Table II. The proposed model achieves good performance in providing exactly the correct answer, even considering that it is obtained on an Italian dataset explored by very few studies in the literature. The F1-score of 59.41% indicates a good balance between precision and recall, and an EM of 46.25% indicates that almost half of the answers provided by the model are perfectly accurate. An EM score lower than the F1-score means that the model often approximates correct answers but does not provide correct answers. EM is a very rigorous metric because it measures the percentage of answers that are exactly the same as the correct answers.

TABLE I. Training Parameters for the Proposed Bert-Based Model

Parameter	Value
Encoding dimension	128
Decoding dimension	64
Loss	Sparse Categorical Cross-Entropy
Optimizer	Adam
Batch size	8
Learning rate	5e-5
Number of epochs	6
Dropout	False

TABLE II. Performance Results of the Proposed Bert-Based Model on the Squad-It Dataset

Metric	Score (%)
F1-score	59.4106
EM	46.2544

Moreover, a comparative analysis has been conducted to evaluate the performance of the proposed model compared to the BIDAF model. To a fair comparison, the BIDAF model is trained and tested with the same train-test split ratio used for the proposed model. Table III details the training model parameters used for BIDAF. The comparative analysis shows that the proposed model achieves an improvement of 10.06% in F1-score and 7.81% in EM compared to BIDAF, as shown in Table IV. In other words, the proposed model is capable of providing significantly more correct answers compared to the BIDAF model.

TABLE III. Training Bidaf Model Parameters

Parameter	Value
Loss	Sparse Categorical Cross-Entropy
Optimizer	Adam
Batch size	10
Learning rate	5e-4
Number of epochs	10
Dropout	0.2

TABLE IV. Comparative Analysis

Model	F1-score	EM
Proposed	**59.4106**	**46.2544**
BIDAF	49.3504	38.4313

VI. Discussion

The development of an Italian-language chatbot tailored for Public Administration represents a significant step forward in the application of NLP models to non-English languages. Throughout this study, it became clear that the unique linguistic and contextual challenges of the Italian language, especially in formal and legal settings, demand specialized models that go beyond generic NLP solutions. Our personal insight is that the integration of an encoder/decoder module and a highway network module into the BERT architecture not only enhances the model's ability to process complex input but also addresses the specific needs of the Italian language, which is often syntactically richer and more flexible than English. This study has demonstrated the potential of these modifications to improve the accuracy and relevance of responses in a chatbot context, which is crucial for the efficiency of Public Administration tasks.

Looking towards the future, several areas offer promising opportunities for further exploration. Firstly, the proposed model could benefit from ensembling strategies, where multiple models are combined to refine the accuracy and robustness of the responses. This could help mitigate the limitations observed in exact match accuracy, especially in cases where the model approximates but does not fully capture the correct answers. Additionally, expanding the dataset beyond the SQuAD-IT to include domain-specific data from Public Administration could enhance the model's contextual understanding and make it even more effective in real-world applications. Another avenue worth exploring is the application of transfer learning techniques, where the model could be pre-trained on broader Italian-language datasets and then fine-tuned on Public Administration-specific data. This approach could further improve the model's performance in specialized contexts. Finally, considering the rapid advancements in NLP, future work could also involve adapting the proposed model to more recent architectures like GPT, which might offer even greater capabilities in terms of natural language understanding and generation.

VII. Conclusion

An extended architecture of the BERT model specific to understanding the Italian language has been proposed. The proposed model introduced an encoder/decoder module and a Highway network module before the fully-connected layers into the BERT architecture to improve the ability of the model to capture the context and temporal sequence and to filter out irrelevant information. The proposed model performs well in providing the exact expected answers. Comparative analysis shows that the proposed model outperforms the BIDAF model, providing almost 8% more correct answers. Moreover, the proposed model is one of the first models developed specifically to be able to process Italian language texts, as it was tested with the Italian version of the popular SQuAD dataset. The proposed model represents the ground for developing an Italian-specific chatbot for the PAs. Therefore, the proposed model, implemented in a chatbot, could make the work of PAs more efficient in terms of time and workload reduction. In future, ensembling strategies could be used by combining responses from multiple models to improve overall accuracy and the number of incorrect answers.

Acknowledgment

This research was funded by Regione Puglia - Progetto NOVA PA - Contabilità ed E-procurement per una P.A. Digitale - POR PUGLIA FESR 2014-2020 - Grant number: SJA8705.

References

[1] M. Pislaru, C. S. Vlad, L. Ivascu, e I. I. Mircea, «Citizen-Centric Governance: Enhancing Citizen Engagement through Artificial Intelligence Tools», Sustainability, vol. 16, fasc. 7, p. 2686, 2024.

[2] J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». arXiv, 2018.

[3] A. Severyn e A. Moschitti, «Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks». arXiv, 2016.

[4] Z. Wang, H. Mi, e A. Ittycheriah, «Sentence Similarity Learning by Lexical Decomposition and Composition». arXiv, 2016.

[5] R. Sequiera et al., «Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering». arXiv, 2017.

[6] Z. Wang, W. Hamza, e R. Florian, «Bilateral Multi-Perspective Matching for Natural Language Sentences». arXiv, 2017.

[7] Y. Tay, M. C. Phan, L. A. Tuan, e S. C. Hui, «Learning to Rank Question Answer Pairs with Holographic Dual LSTM Architecture», in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku Tokyo Japan, 2017, pp. 695–704.

[8] M. M. A. Zaman e S. Z. Mishu, «Convolutional recurrent neural network for question answering», in 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, 2017, pp. 1–6.

[9] W. Wang, N. Yang, F. Wei, B. Chang, e M. Zhou, «Gated Self-Matching Networks for Reading Comprehension and Question Answering», in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 2017, pp. 189–198.

[10] T. Shao, Y. Guo, H. Chen, e Z. Hao, «Transformer-Based Neural Network for Answer Selection in Question Answering», IEEE Access, vol. 7, pp. 26146–26156, 2019.

[11] M. Kamyab, G. Liu, e M. Adjeisah, «Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis», Appl. Sci., vol. 11, fasc. 23, p. 11255, nov. 2021.

[12] M. Seo, A. Kembhavi, A. Farhadi, e H. Hajishirzi, «Bidirectional Attention Flow for Machine Comprehension». arXiv, 2016.

[13] D. Croce, A. Zelenanska, e R. Basili, «Neural Learning for Question Answering in Italian», in AI*IA 2018 – Advances in Artificial Intelligence, vol. 11298, C. Ghidini, B. Magnini, A. Passerini, e P. Traverso, A c. di Cham: Springer International Publishing, 2018, pp. 389–402.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, e P. Liang, «SQuAD: 100,000+ Questions for Machine Comprehension of Text». arXiv, 2016.

Deep Neural Network and Human-Computer Interaction Technology in the Field of Art Design

Lan Guo, Lisha Luo, Weiquan Fan*
College of Fine Arts, Jingchu University of Technology, Jingmen 448000, Hubei, China

***Abstract*—Traditional art design is usually based on the designer's intuitive creativity. Limited by individual experience, knowledge and imagination, it is difficult to create more abundant and higher quality works, and the workload is huge, which limits the production efficiency of artworks. Through deep neural networks and human-computer interaction technology, the quality of art design can be improved; the workload and cost of designers can be reduced, and more artistic inspiration and tools can be provided to designers. The main contribution of this paper is to propose the use of a Cycle Generative Adversarial Network (Cycle GAN) to realize the automatic conversion of text to image and provide an immersive art experience through human-computer interaction technology such as virtual reality. In addition, the target audience of this paper is art designers and researchers of human-computer interaction technology, aiming to help them break through the traditional creation mode and lead art design to diversification and avant-garde. The content loss rate of character image conversion in Cycle GAN was reduced by 74.5% compared with that of human image conversion. The average peak signal-to-noise ratio of figure images generated by Cycle GAN was 57.9% higher than that of figure images generated by the artificial method. The character images generated by Cycle GAN reduce content loss and are more realistic. Deep neural networks and human-computer interaction technology can promote the development and progress of art design, break the traditional creative mode and bondage, and lead art to be more diversified and avant-garde.**

***Keywords*—Deep neural network; human-computer interaction; Cycle Generative Adversarial Networks; art design; image generation**

I. Introduction

Traditional art design often relies on manual production and processing, which requires a longer time and a large amount of labor. This limits the production efficiency and quantity of artworks, making it difficult to meet large-scale market demand. Deep neural networks can automatically generate images, providing artists with ideas and more design options. The sound interaction and motion interaction in the human-computer interaction technology collect the designer's voice input and movement analysis, and deepen the interaction with the art work. Deep neural networks and human-computer interaction technologies provide artists with new ways to create and extend the boundaries of their imagination and creativity. Artists can express their creativity and ideas directly through natural language, without being limited by technology or tools. They can use the language and expression they are familiar with to translate their creative inspiration into images or design works. Although existing research has explored the application of artificial intelligence in art design, most studies have not analyzed in depth the specific advantages of deep learning techniques in improving the quality and design efficiency of artworks, nor have they systematically compared the practicability and effectiveness of traditional methods with emerging technologies.

The deep learning neural network analyzed in this paper is a kind of machine learning technology that mimics human brain neural networks. It can realize automatic learning and processing of massive data by constructing complex networks. This method has been well applied in computer vision and natural language processing. In terms of art design, the use of deep learning neural networks can allow computers to learn different art styles, skills and elements, and then achieve the purpose of artificial intelligence design. In this paper, Cycle GAN is used to automatically generate images. Cycle GAN is a method to understand and process the semantics and grammar of natural language, and corresponding pictures can be automatically generated according to the description of artists. Deep Convolutional Inverse Graphics Networks (DC-IGN) are also used to transform image styles in combination with time or space. Through sound interaction and body interaction in human-computer interaction, artists can interact with their ideas and create the works they want as long as they express them in a natural way. In the existing literature, although some studies have explored the impact of artificial intelligence and digital tools on artistic creation, this paper verifies the actual effectiveness of deep learning in artistic design through empirical research, especially in the innovative application of automatic text-to-image conversion and image style conversion. Further, through detailed technical comparative analysis, this paper demonstrates the remarkable effects of Cycle GAN in reducing content loss and improving realism in art image generation, providing a new creation tool and methodology for art designers.

This paper aims to explore the application of deep neural networks and human-computer interaction technology in the field of art design, especially how to improve the quality and efficiency of art creation through these technologies, while reducing the workload and cost of designers. The goal of the research is to evaluate the performance of Cycle Generative Adversarial Networks (Cycle GANS) in automatic text-to-image conversion, and combine interactive technologies such as virtual reality to enhance the display of art works and the immersive experience of the audience, in order to promote the development of art design in a more diversified and avant-garde direction.

Main contributions:

1) Deep neural network improves the quality of artistic creation: This paper proposes to use deep neural network technology, especially Cycle GAN, to realize the automatic conversion of text to image, so as to assist art designers to create more abundant and higher quality art works, and reduce the labor intensity of designers.

2) Human-computer interaction technology enhances art experience: This paper discusses how human-computer interaction technology, including virtual reality and multimodal interaction, provides a more immersive and interactive experience for art design, and makes the display and viewing of art works more modern and personalized.

3) Significant improvement of art design efficiency and economy: Through practical data and case analysis, this paper shows that the application of deep neural network and human-computer interaction technology in art design can significantly improve design efficiency, reduce costs, and increase the market competitiveness of art works and audience participation.

This paper first sets the research background in the introduction part, and summarizes the research gaps of deep neural network and human-computer interaction technology in art design. Then, the related work section reviews the current status and challenges of the application of artificial intelligence in art design. The main part of this paper is divided into eight core sections: In Section III, the application of deep neural network in image generation and intelligent creation tools is discussed; secondly, Section IV analyzes how human-computer interaction technology promotes art design and display. Finally, the practical effects of these technologies in improving design efficiency, reducing costs, enhancing user experience and economic benefits are evaluated in Section V. In Section VI, the cost-effectiveness of traditional methods and the proposed methods are further compared, and the impact of human-computer interaction design on user satisfaction is deeply analyzed in Section VII. The conclusion in Section VIII summarizes the research results, emphasizes the importance of technological innovation in the field of art design, and puts forward the direction of future research. All citations are listed in the References section, which provides academic support for the research.

II. Related Work

The field of art design often relies on traditional artificial design. The works designed are affected by many factors, and the design time is relatively long. The emergence of the digital age has provided new means for art design. Changsheng WANG's research showed that AI art creation workflows, such as AI image straight-out method, AI-assisted drawing method and ControlNet precision drawing method, effectively improved the creation efficiency and accuracy, and promoted the innovative expression of digital art [1]. Anantrasirichai explored the application of AI technology in the creative industries, analyzed the use of AI in content creation, information analysis, content enhancement, information extraction, and data compression, and noted that AI's potential as a tool to enhance human creativity was greater than its ability as an independent creator [2]. Through practical workshops for Finnish pre-service craft teachers, Vartiainen explored the potential benefits and challenges of text-image generation AI in art design, such as algorithm-based bias and copyright issues, and analyzed the complex relationship between creative production and generative AI [3]. Wang, Xiaolong designed an intelligent management system of art exhibition based on IoT and AI, which had a short response time and could meet the requirements of practical applications. At the same time, the multi-touch system based on BPNN was developed, and the gesture recognition accuracy was high, providing an efficient solution for the field of interactive art [4]. Cetinic comprehensively reviewed the application of AI in art analysis and creation, including tasks such as art digitization, classification and retrieval, as well as the practice and theory of AI in art creation, and looked forward to the future development of AI in art understanding and creation [5]. Mayo emphasized the importance of pre-service teacher education in integrating AI for artistic creation, proposing the need for structured curricula and technical support to train educators and students while addressing biases and ethical issues in the use of AI to drive innovation in art education [6]. Li Sixian discussed the intelligent development of digital media art through interdisciplinary research methods, analyzed the integration of technology, communication and art, as well as the problems and solutions in the wave of intelligence, and aimed to explore innovative strategies for digital media art [7]. In a critical review of the existing literature, it is found that although the application of artificial intelligence in the field of art and design has been extensively explored, the existing theoretical framework has shortcomings in integrating deep learning with human interaction techniques. For example, while some studies demonstrate AI-based image generation techniques, they often do not consider in depth how these techniques integrate with the creative thinking of designers. In addition, although text-to-image generation methods are theoretically innovative, their application and integration in the actual design process have not been fully explored. At the same time, the existing literature also lacks sufficient depth in exploring how human-computer interaction technology enhances artistic experience.

Deep neural networks and human-computer interaction technology bring new ideas to art design. Nie Zexian discussed the application of artificial intelligence and human-computer interaction technology in art design, through machine learning algorithms and visual interaction technology, and analyzed audience responses to improve the expression of artworks. The design of artistic visual communication systems was proposed, and innovation and development in the art field were promoted [8]. Huang Lihua discussed the application of artificial intelligence technology in visual design, proposed the construction of intelligent design systems to improve design efficiency and quality, and explored the man-machine collaboration model to promote innovation and development in the field of art design through deep neural networks and multi-domain expert systems [9]. Liu Lina studied the methods of context awareness and machine learning to enhance user interaction experience in mobile systems, proposed the design principles of interactive display based on intangible cultural heritage, and discussed the application of augmented reality in folk art appreciation, providing new ideas and practical cases for art design [10]. Guo Qiongqiong studied the application of virtual reality technology in product design, and enhanced user experience through dynamic simulation and haptic feedback. She established product models in combination with computer-aided design, and discussed the implementation mechanism of haptic feedback in virtual environment, providing innovative interactive means and design ideas for art design [11]. Li Xiong proposed an intelligent assisted concept sketch design framework based on deep learning, including GAN and style transfer network for sketch generation and rendering, which verified its ability to quickly generate innovative sketches and style transformations through experiments, and developed an intelligent sketch design generator to reduce the threshold for designers to use AI and improve design efficiency and innovation [12]. ChatGPT's performance as a general-purpose AI model on emotional computing tasks highlighted the broad applicability and robustness of deep learning techniques in diverse fields such as art and design without the need for task-specific training [13]. Although existing research has made progress in applying AI to art and design, they have generally failed to delve into the specific advantages of deep learning techniques in improving the quality of art works and the efficiency of design. In addition, existing approaches have limitations in adapting to diverse design needs and delivering personalized art experiences. Through empirical analysis, this study aims to make up for these deficiencies, verify the innovative application of deep learning technology in art design, and explore how to enhance the display of art works and audience experience through human-computer interaction technology.

III. Deep Neural Network in Art Design

A. Image Generation

Deep neural networks can learn and simulate a large amount of image data, to achieve automatic image generation. This provides an entirely new means of creation for art designers, enabling them to better express their creativity and ideas.

At present, the mainstream text-generating image methods need complete text-image data pairs in the training process, which leads to the current methods being mostly trained on some specific data sets, and it is difficult to apply to other scenarios. In recent years, there have been some unsupervised field transitions. Cycle GAN is an image transformation model based on generative adversarial networks that converts text into images without the need for pairs of training data.

Cycle GAN was originally a model for the image domain, but has since been extended to text-image conversion tasks as well. The overall structure of text-image conversion is shown in Fig. 1.

As shown in Fig. 1, text is first encoded by a text encoder into a specific hidden space. By converting text features into vector representations, the encoded text vectors become part of the input generator, which is responsible for converting the text vectors into real images.

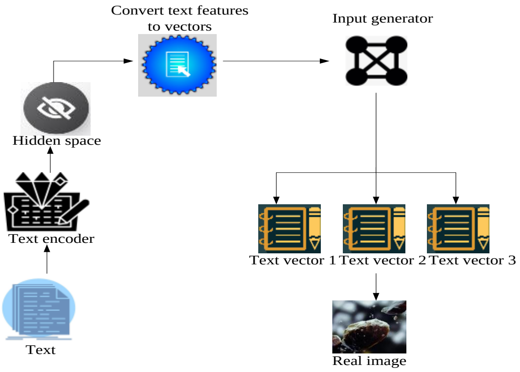


Fig. 1. Overall structure of text-image conversion.

The text is encoded using a pre-trained network to get a semantic embedded representation of the text:

$$w = B_i \cdot L(t) \tag{1}$$

$L(t)$ is the length of text; B_i represents text features at the word level. In order to enhance the diversity of text description, conditional enhancement network is used to make full use of different texts with the same semantics to enhance the semantic information of text vectors. For the obtained text features, it is taken as the input of the network, and the result of data enhancement $F_c(s)$ and noise are taken together as the input f_0 of the first stage image generator:

$$f_0 = F_0\left(z, F_c(s)\right) \tag{2}$$

In the text loop, the cross-entropy loss is used to calculate the consistency between the original text p_t and the reconstructed text s_t. The cross entropy loss L_{cycle} of the text loop is as follows:

$$L_{cycle} = -\sum_{t=0}^{J-1} \log p_t(s_t) \tag{3}$$

For different attribute information, the encoding method is different. For class information, the variational inference method is used to encode. The encoder inputs a given class of information and a noise vector sampled from a standard Gaussian distribution to make variational inferences about hidden variables. Supposing that the posterior distribution of class information follows a multivariate Gaussian, the formula is:

$$L_{cycle}(G, F) = \mathrm{L}_1 \cdot G(x) \cdot \mathrm{F(y)} \tag{4}$$

G and F are generator network and discriminator network respectively; x and y represent distribution obedience in two standard Gaussian distributions respectively.

The attribute vector is transformed into the same dimension as the text vector through a learnable linear layer, and then a neural network is used to jointly learn the text information and the attribute information to obtain the fused conditional vector. Through adversarial training, the generator network can continuously improve the authenticity of the generated image, making it more close to the real image distribution of the target domain or the original domain:

$$L_{GAN}\left(D_M, D_N\right) = sim\left(1-D_M\right)\cdot\left(1-D_N\right) \tag{5}$$

D_M and D_N represent the target domain and the original domain respectively, and the cross-entropy loss function is used to constrain the training of generator network and discriminator network.

After the fused code is transformed into the image space through the network, the attention mechanism is used to introduce the text features at the word level to better constrain the image generation. In order to maintain coherence between text and image, cyclic consistency loss is introduced into text-image conversion module. This loss encourages the generator network to retain as much content and style information as possible in the input image:

$$L_{D_i} = \log D_i\left(x_i\right) - \left[\log\left(1-D_i\left(x_i\right)\right)\right] \tag{6}$$

D_i is the weight hyperparameter, which is used to balance the ratio between the individual subloss functions. By optimizing the total loss function, the parameters of generator and discriminator network can be updated by gradient descent to optimize the performance of the network.

In this paper, text with 10, 20, 30, 40 and 50 characters is selected to analyze the success rate of text generation image of artificial method and Cycle GAN, as shown in Fig. 2. In Fig. 2, the horizontal coordinate represents the characters and the vertical coordinate represents the success rate.

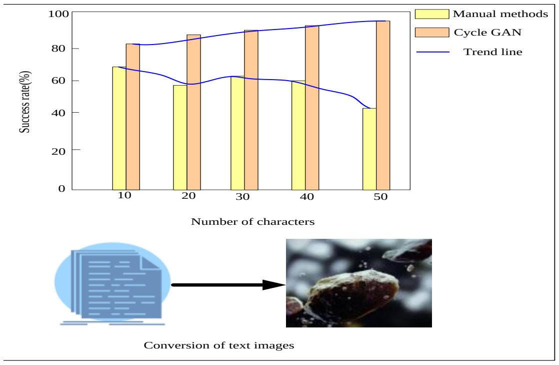


Fig. 2. Success rate of text generation image based on artificial method and Cycle GAN.

As shown in Fig. 2, it can be seen that the success rate of text-image generation in Cycle GAN is always higher than the success rate of text-image generation in artificial methods, which means that the success probability of text-image generation in Cycle GAN is greater.

B. Intelligent Authoring Tools

The wide application of deep neural network enables it not only to be applied in traditional machine learning tasks, but also as an intelligent creation tool to provide better human-computer interactive creation experience and help in the field of art [14-15].

Generative Adversarial Networks (GAN) are composed of a generator network and a discriminator network, which generate images through adversarial training. Image generation and judgment based on generative adversarial network are shown in Fig. 3.

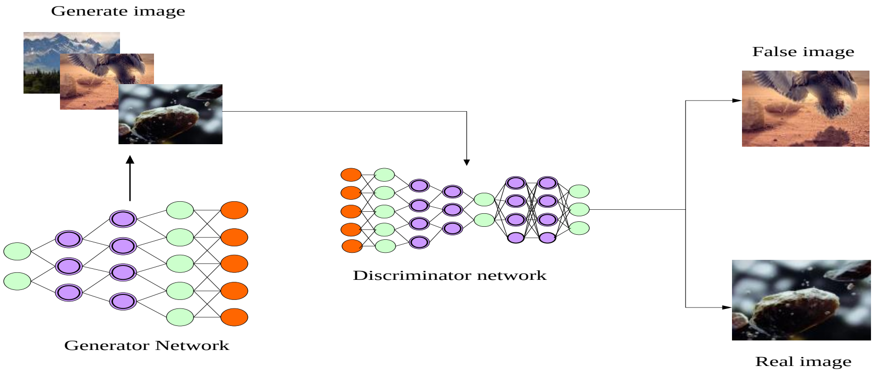


Fig. 3. Image generation and judgment based on generative adversarial network.

As shown in Fig. 3, the generator network is responsible for generating the image, while the discriminator network is responsible for determining whether the generated result is realistic. Through continuous adversarial learning, the generator network gradually improves the quality of generation. The goal of the generator is to produce images that are realistic enough to fool the discriminator so that it cannot tell whether the resulting image is real or not.

The generator's loss function can be implemented by maximizing the probability that the generated image is judged to be true by the discriminator. The specific formula is as follows:

$$\mathrm{L\{G\}} = \xi\left[\mathrm{-logD(G(z))}\right] \tag{7}$$

ξ is random noise sampled from a prior noise distribution. The goal of the discriminator is to judge the difference between the generated image and the real image. The specific formula of its loss function is as follows:

$$\mathrm{L\{F\}} = \mathrm{logD(h)\text{-}log}\left[\mathrm{1\text{-}D}\left(G\left(z\right)\right)\right] \tag{8}$$

L{F} is the loss function of the discriminator and D(h) is the real image. Through repeated iterations, the generator gradually learns how to generate an image that matches the natural language description, while the discriminator gradually learns how to accurately distinguish between the generated image and the real image. This adversarial training process allows the generator to output images that match the natural language description.

In this process, the antagonistic process of training the generator and discriminator enables the generator to generate an image or design that matches the natural language description. Through iterative training, the generator gradually learns how to generate works of art based on natural language, while the discriminator gradually learns how to accurately distinguish between generated works and real works. This adversarial training process enables the generator to output an image or design that matches the natural language description [16-17].

C. Improvement of Creation Efficiency

With deep neural networks, artists can turn their ideas into creative works more quickly. Compared with the traditional artificial design or painting process, the use of intelligent creation tools can greatly shorten the creation time and reduce the workload of modification and adjustment [18-19].

As for the current method of text image generation, the quality of the final generated image depends on the quality of the initial generated image, so there are problems in image generation. Deep neural networks can be used to improve the quality of the initial generated image, and different generation tasks can be performed in different generation stages to improve the quality of the generated image. Deep neural network can fully learn the features of different levels of text generated images, so it has good performance.

The time taken by artificial method and intelligent creation of an inset, animation, trademark, person, build, landscape and animal image are counted respectively. The time taken by artificial method and intelligent creation is shown in Table I.

As shown in Table I, when the type of image needs to be created is determined, the creation time of the artificial method is measured in minutes, while the intelligent creation is measured in seconds. In several image types, the longest intelligent creation time is 32.20 seconds, while the shortest artificial creation time is 2.02 minutes.

D. Image Style Conversion

In the mainstream image style transformation technology, after many repeated operations to achieve the desired effect, some features of the main image are distorted. The most widely recognized is DC-IGN, which takes the full connection of traditional neural networks to replace the convolution operation. It combines time or space, reduces network free parameters, reduces training complexity, and is well used in image style conversion. By using the idea of generative adversarial network and the structure of convolutional inverse graph network, DC-IGN realizes the transformation of image style.

DC-IGN introduces adaptive instance normalization. As long as it inputs a content and a style information by adjusting the input content to match the variance and mean of the input style, it can effectively merge the content of the main and style diagram, and output a new image. It can be calculated using the mean square error as follows:

$$h_t = \mathrm{M}_{\mathrm{SE}}\left(\alpha, \beta\right) \tag{9}$$

Among them, the content feature α of the generated image is extracted from the generated image, and the content feature β of the target image is extracted from the target image.

The style extraction network uses a similar convolutional neural network to take the style image as input and select the feature output of different layers as the style representation of the image. These features reflect the texture, color, and stylistic features of the image. Style loss is used to measure the difference in style characteristics between the generated image and the target image. It is measured by calculating the M_{SE} between the generated image and the target image in a specific layer, and the formula is as follows:

$$r_t = \mathrm{M}_{\mathrm{SE}}\left(\mathrm{Gram}\left(\alpha\right), \mathrm{Gram}\left(\beta\right)\right) \tag{10}$$

$\mathrm{Gram}\left(\alpha\right)$ and $\mathrm{Gram}\left(\beta\right)$ are Gram matrices that generate the stylistic features of the image and the target image at a particular layer, respectively. In order to train a generator in DC-IGN, it needs to define a generator loss function, which consists of content loss and style loss, designed to minimize the goal.

TABLE I. TIMING OF MANUAL METHODS AND INTELLIGENT AUTHORING

Type	Manual Method (Minutes)	Intelligent Creation (s)
Inset	2.44	21.84
Animation	2.02	32.20
Trademark	4.57	23.04
Person	2.05	19.39
Build	2.77	24.96
Landscape	2.61	22.51
Animal	2.38	27.83

DC-IGN can encode not only the content of the image, but also the style information of the image. The style and content of image can be separated, which also deepens the understanding of image processing. It has a broad application prospect and can be used for image processing, video processing and as an auxiliary tool for style design.

IV. Human-Computer Interaction Technology in Art Design

A. Human-Computer Collaboration Design

Human-computer collaborative design refers to the cooperation and interaction between designers and computers to improve design efficiency and quality. By making full use of the computing and processing power of computers, designers can complete design tasks more conveniently and quickly, and obtain better design results [20-21]. In human-computer collaborative design, there are many technologies and tools that can realize the cooperation and interaction between designers and computers, among which sound interaction and somatosensory interaction are some of the common implementation methods [22-23].

1) Sound interaction: Sound interaction technology is a simulation system based on human hearing and understanding systems. When the sound sensor converts the sound signal into an electrical signal, the electrical signal can trigger the circuit in the interactive system to work. In art viewing, when the audience's voice is perceived by the interactive system, it can produce corresponding feedback output after being processed by the pre-set system. The sound interaction technology based on sound sensors can also be subdivided into volume-led interaction systems and speech recognition.

In volume-dominated interactions, the system only needs to sense the absolute volume of urine and faeces at a specific location to respond accordingly [24-25]. However, in the speech recognition interaction, the system also needs to distinguish the content of the sound. If the sound emitted in a specific area is meaningless, it can be judged by the system as an invalid sound and cannot give feedback. Only when the sound is carrying information and is successfully recognized by the system can it produce corresponding feedback based on the content of the sound? For example, the "Ascending the River at Qingming Festival" in the Sound Museum is shown in Fig. 4.

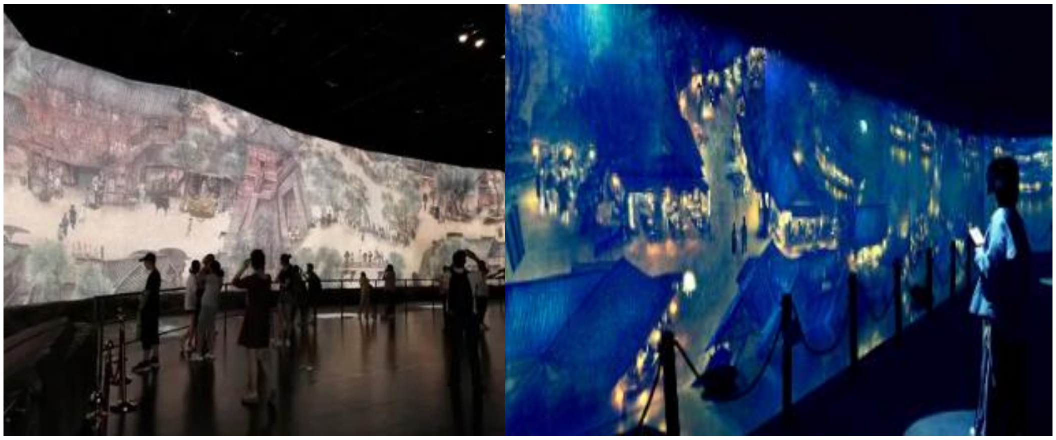


Fig. 4. Heard "Ascending the river at qingming festival" art exhibition.

As shown in Fig. 4, the art exhibition of "Riverside Scene at Qingming Festival" uses holographic projection and holographic sound technology, and through the independently developed soundscape interaction system, the outdoor exhibition restores the life status of folk people in the Song Dynasty, and people interact with the scroll in sound and painting while walking. Through multi-dimensional immersive experience, people can listen to the sound art that has come to modern society through thousands of years.

By simulating the communication between people, voice interaction enables designers to directly interact with computers in a natural way, such as voice input, which greatly reduces the learning cost and improves the work efficiency of designers [26]. Voice input can realize the voice interaction between the designer and the machine, providing a more convenient and natural way of operation. Through a microphone or other audio device, the designer's voice input is collected, and these voice signals are used as input data for subsequent processing.

2) Somatosensory interaction technology: The somatosensory interaction technique is a technique to track, record and dynamically capture human motion trajectory in three-dimensional space by using the mixed method of optical passive and inertial motion measurement. Somatosensory interaction includes gesture interaction and body interaction. Due to the diversity and ambiguity of gesture and body behavior, various combinations can be used to input a variety of information in interaction, as shown in Fig. 5:

As shown in Fig. 5, the somatosensory interaction technology usually relies on infrared sensors, cameras, smart wearable devices and other technologies to achieve the input behaviour in the interaction.

Since people usually use body movements to complete communication with other natural persons in social activities, somatosensory interaction is a simulation of human natural communication behavior, which has the advantage of being easy to use and understand.

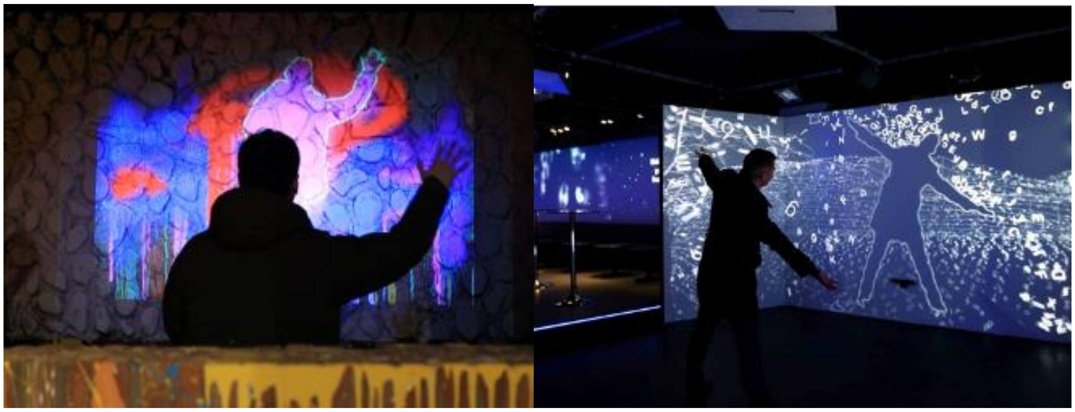


Fig. 5. Somatosensory interactive projection.

B. Art Display

Human-computer interaction technology can combine art display with user interaction. For example, digital artworks can be displayed in a virtual exhibition hall, where the audience can experience and interact through technologies such as virtual reality, increasing the interest and interactivity of the works.

1) Virtual reality: Virtual reality is an information display technology that uses three-digit graphics generation technology, multi-sensor interaction technology and high-resolution display technology to generate a three-dimensional realistic virtual environment [27]. The environment simulated by virtual reality technology is very similar to that in the real world, which is difficult to distinguish, and people can have a sense of immersion in the experience process. The current virtual reality technology has been able to deliver hearing, vision, smell, touch, taste and other feelings, which is a comprehensive simulation system and has brought a huge revolution to the output of interaction design [28].

The Virtual Reality Painting tool allows users to wear a virtual reality headset and handle and then paint in a virtual environment in the form of an aerial drawing. Users can draw lines, shapes and colors in three-dimensional space with the handle, and create three-dimensional works of art by touching and rotating them. This virtual reality painting tool provides an innovative art creation tool that enables artists to paint in completely new ways, creating artwork with three-dimensional and dynamic effects, as shown in Fig. 6.

As shown in Fig. 6, through the immersive experience and interactive nature of virtual reality, the audience can interact more deeply with the artwork and freely explore and create art.

This virtual reality painting tool can also combine traditional painting with technology, providing more possibilities and innovative potential for artistic creation. Virtual reality technology provides the audience with an immersive experience, allowing them to feel the authenticity and emotion of the artwork. It provides artists with new creative tools and ways of expression to create more free and innovative works of art. Virtual reality can break through the limitations of time and space, bring the audience into different art scenes and art history, and expand the dimension of art creation and viewing.

2) Augmented reality: Augmented reality is a technology that integrates virtual information with scenes in the real world [29]. This technology uses multimedia, three-dimensional modeling, intelligent interaction, sensors and other technical means to apply virtual information produced by computers in the real world, which is complementary to the real world, so it is called augmented reality [30-31].

In the art field, augmented reality technology can be used to combine art museum exhibitions with virtual reality. Through the camera of a mobile phone or tablet computer, the audience can interact with these virtual elements by watching, hearing and touching, which entices the understanding and experience of artworks [32-33]. The Museum of augmented reality Art is shown in Fig. 7.

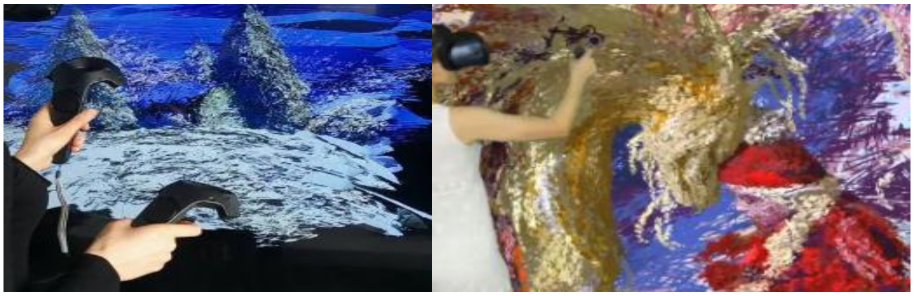


Fig. 6. Virtual reality painting.

Fig. 7. Museum of augmented reality art.

As shown in Fig. 7, the viewer can point the device at the artwork in the museum. Virtual elements related to the artwork would then appear on the screen, such as 3D models, animations, audio, etc.

This augmented reality art museum application expands the form and content of art exhibitions to a certain extent, providing viewers with a richer visual and sensory experience. At the same time, through interaction and participation, the audience can have deeper communication and understanding with the works of art, so as to enhance the sense of participation and the depth of art appreciation [34-35].

Visitors themselves feel and experience art themes through audio-visual touch and even smell, which further broadens the way participants receive information and expands the communication function of art design [36-37]. In contemporary art design, the interactive design running augmented reality technology can better create an immersive display atmosphere and improve the display effect [38-39].

V. Design Effect of Deep Neural Network and Human-Computer Interaction Technology

A. *Image Generation Effect*

The image generated by artificial methods is often a one-of-a-kind work that is not easy to reuse and scale. If multiple similar images need to be generated, the drawing or design work needs to be repeated, which further reduces efficiency and feasibility.

The resulting image fidelity is evaluated using the peak signal-to-noise ratio (PSNR). The sharper image PSNR is between 30 decibel (dB) and 40dB. The higher the PSNR, the more realistic and clear the image. Eight images of people, buildings, landscapes and animals are generated by manual method and Cycle GAN respectively. Compared with the peak signal-to-noise ratio, the average peak signal-to-noise ratio of images generated by different methods is shown in Fig. 8. The horizontal axis represents people, build, landscape and animal; the vertical axis represents the peak signal-to-noise ratio; the unit is dB.

As shown in Fig. 8, the average peak signal-to-noise ratio of person, build, landscape and animal images generated by artificial methods is 23.73dB, 20.58dB, 20.99dB and 21.10dB respectively on the left side of Fig. 8. On the right side of Fig. 8, the average peak signal-to-noise ratio of person, build, landscape and animal images generated by Cycle GAN is 37.48dB, 32.86dB, 33.65dB and 38.04dB respectively.

The average peak signal-to-noise ratio of figure images generated by Cycle GAN is 57.9% higher than that of figure images generated by artificial method ($\frac{37.48\text{-}23.73}{23.73}*100\% = 57.9\%$).

Each layer of Cycle GAN can extract the features of each level of the image, and through the effective combination of the features of each level, it has higher complexity and higher quality. In contrast, in the manual method, it is necessary to undergo many processes of processing and debugging in order to obtain a complete image.

B. *Style Conversion Quality*

Artificial methods often need to manually select features to express the content and style of the image, while deep neural networks can superimpose multiple neurons layer on layer to form a richer and more abstract image feature expression. Because deep neural networks can process any size and any kind of image, artificial methods often have strict constraints on the size and category of images. Therefore, using a deep neural network for image style conversion can adapt to more applications and have wider applicability.

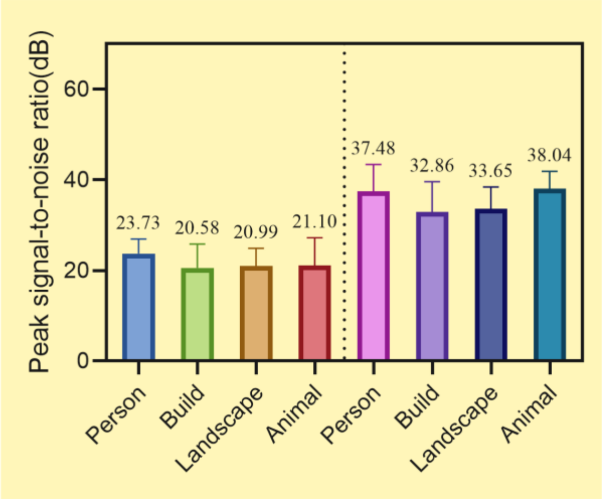


Fig. 8. Average peak signal-to-noise ratio of images generated by different methods.

The content loss rate is used to evaluate the content difference between the converted image and the original image. It can evaluate the style conversion quality of the image. The higher the content loss rate, the worse the style conversion quality. Besides, eight images of people, buildings, landscapes and animals are converted by manual method and Cycle GAN respectively, and their content loss rate is calculated.

The content loss rate of manually converted images is shown in Table II.

As shown in Table II, the average content loss rates of the eight images of person, build, landscapes and animals converted by manual methods are 5.68%, 5.36%, 5.67% and 5.62%, respectively.

The content loss rate of Cycle GAN converted images is shown in Table III.

As shown in Table III, the average content loss rate of each of the eight images of person, build, landscapes and animals converted by Cycle GAN is 1.45%, 1.60%, 1.17% and 1.53%, respectively.

Compared with manual method, the content loss rate of character image conversion in Cycle GAN increases by -74.5%($\frac{1.45\text{-}5.68}{5.68} = \text{-}74.5\%$), that is, the content loss rate of character image conversion is reduced by 74.5%.

Cycle GAN can learn the more complex correlation between image content and style from massive data, and realize rapid transfer of image style through optimization and acceleration technology. Manual methods often take a long time, but Cycle GAN can achieve rapid style change in the case of real-time and interactive.

C. Reduce Design Costs

The cost includes human resource cost, material cost, production cost, equipment cost and site cost. The cost of design work for manual method design and combined method (deep neural network and human-computer interaction technology) is shown in Fig. 9. The horizontal coordinate represents human resources, materials, manufacture, equipment and site; the vertical coordinate represents cost; the unit is yuan.

The left side of Fig. 9 shows the human resource cost, material cost, manufacture cost, equipment cost and site cost of the manual method design work, with the highest cost exceeding 1000 yuan. The right side of Fig. 9 shows the human resource cost, material cost, manufacture cost, equipment cost and site cost of the combined method design work. The maximum cost is less than 1,000 yuan.

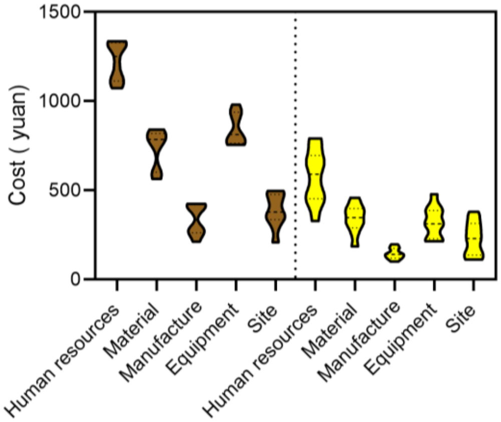


Fig. 9. Cost of manual method design and combined method design work.

TABLE II. CONTENT LOSS RATE (%) OF MANUALLY CONVERTED IMAGES

Image Sequence Number and Average	Person	Build	Landscape	Animal
1	6.88	6.18	5.25	6.19
2	5.34	4.00	5.36	5.84
3	4.50	4.52	6.37	6.93
4	5.79	6.32	6.78	4.81
5	5.47	6.49	6.00	4.65
6	6.89	6.80	5.48	6.04
7	5.89	4.25	5.26	4.38
8	4.70	4.30	4.89	6.14
Average	5.68	5.36	5.67	5.62

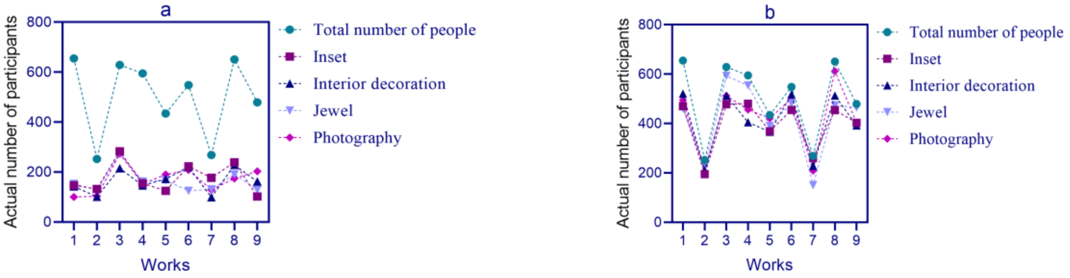
TABLE III. CONTENT LOSS RATE OF CYCLE GAN CONVERTED IMAGES (%)

Image Sequence Number and Average	Person	Build	Landscape	Animal
1	2.48	2.41	2.01	0.62
2	2.37	1.08	0.98	1.81
3	1.16	1.32	2.32	1.26
4	0.76	2.05	0.62	1.06
5	1.13	0.91	0.67	2.19
6	0.53	1.97	0.51	1.20
7	1.89	0.72	1.62	2.03
8	1.26	2.36	0.61	2.03
Average	1.45	1.60	1.17	1.53

Manual methods often take a long time, through the design, production, touching up and other processes, in order to ensure the quality and effect of the work. If the production cycle is too long, it not only causes the rise of production costs, but also misses the changing market demand.

D. Engagement

The display of artworks based on human-computer interaction can bring real experience to visitors no matter in sensory experience, behavior mode, or use environment [40]. This paper selects nine works of illustration, interior decoration, jewelry and photography designed by manual method and combines method respectively to investigate the audience's participation, as shown in Fig. 10. The horizontal coordinate represents the works, and the vertical coordinate represents the actual number of participants.



(a). Audience participation in the artificial design of the work.

(b) Audience participation in the design of the combined method.

Fig. 10. Audience participation in the design of works by manual methods and combined methods.

Fig. 10(a) shows the participation of the audience in the inset, interior decoration, jewelry, and photography by manual methods. It can be seen that the actual number of participants is much less than the total number, and the highest number of participants is less than 400.

Fig. 10(b) shows the audience participation in inset, interior decoration, jewelry, and photography designed based on the combination method. It can be seen that the difference between the actual number of participants and the total number of participants is not very far, and the highest number of participants exceeds 400.

Human-computer interaction emphasizes "people-oriented", providing the computer with the functions of touch, vision, hearing and other aspects, so that users can interact with people through gestures, expressions, eyes, sounds and other ways. The multi-dimensional input and multi-output of computer have greatly increased the frequency band of communication between human and computer [41-42]. With the electronization of information resources, the diversification of information presentation methods, and the continuous improvement of new application requirements for various industries, art design, as the carrier of human spiritual civilization, cultural tradition and scientific knowledge, is gradually becoming the spiritual food of human beings.

E. Economic Benefits

The appearance of human-computer interaction technology has brought great convenience and efficiency to art design. This paper uses virtual reality or augmented reality technology, and art designers can quickly preview and adjust the design effect in virtual reality, thus saving a lot of time and cost. Eight works are randomly selected and the economic benefits they brought are calculated. The income generated by manual method design and combined method design is shown in Table IV (the difference in income in Table IV is the income generated by combined method design minus the income generated by manual method design):

TABLE IV. INCOME FROM MANUAL METHOD DESIGN AND COMBINED METHOD DESIGN (TEN THOUSAND YUAN)

Works	Manual methods	Combining method	Differential income
1	6.01	17.48	11.47
2	8.16	20.72	12.56
3	5.25	15.04	9.79
4	8.33	19.02	10.69
5	8.78	23.51	14.73
6	9.40	24.83	15.43
7	6.45	21.81	15.36
8	8.88	16.17	7.29

As shown in Table IV, the economic income brought by manual method design is less than 100,000 yuan, while the economic income brought by combined method design is more than 100,000 yuan. It can be seen that the economic income brought by the combined method design is higher.

This paper highlights the advantages of the adopted method through in-depth comparison with previous studies. Through quantitative analysis, this study demonstrates the significant effect of Cycle GAN in reducing the content loss rate in automatic text-to-image conversion, as well as the high fidelity achieved in image style conversion. In addition, compared with traditional design methods, this research method has obvious advantages in improving the quality and creation efficiency of artworks, while reducing the production cost and enhancing the market competitiveness of artworks. These comparisons not only validate the validity of this research method but also demonstrate its potential to drive innovation and progress in the field of art design.

VI. DISCUSSION

Fig. 9 shows the overall higher cost of designing work using manual methods. In the manual method, designers, engravers, painters, artisans, etc., have to put a lot of energy into completing a work. It not only requires a high technical level, but also requires them to spend a lot of time on each piece of work. In traditional art design, the materials used are generally very expensive, such as canvas, paint, wood, metal and so on. Such materials are prone to loss and waste in the production process, thus increasing the production cost.

Interactive technology is a kind of technology that appears with the computers and artificial intelligence devices in the information age. Therefore, in everyday use, interactive technologies are often designed to have a technological, futuristic look and style. However, in art design, the fundamental goal of interactive technology is to realize the transmission of artistic ideas. Therefore, in the interactive design of art, attention should be paid to the consistency of the interactive equipment with the displayed theme in terms of appearance, audition style, etc., to avoid conflict or incompatibility with the design theme.

Table IV shows that the economic income brought by artificial method design is not as high as that brought by combined method design. The appearance of human-computer interaction technology has brought great convenience and efficiency to art design. By utilizing virtual reality or augmented reality technology, art designers can quickly preview and adjust design effects in virtual reality, thus saving a lot of time and cost.

The works of human-computer interaction design pay more attention to bringing good experience to users, that is, the indicators based on users' emotions, such as satisfaction. It also has a certain subjectivity, emphasizing that users can be immersed in interactive experiences while obtaining information about artworks. Human-computer interaction plays an important role in realizing human value and satisfying human spiritual needs. In the process of art design, users manipulate some interactive elements on the digital media terminal to get corresponding visual, auditory and tactile feedback. If such feedback is consistent with the user's own knowledge, skills or values, it causes a feeling of pleasure in the "emotional center" of the user's brain, thus satisfying the user's deep understanding of the value and connotation of the work.

The research results show that the application of deep neural network and human-computer interaction technology in the field of art design can significantly improve the quality and efficiency of design works. Compared with traditional design methods, these techniques reduce the content loss rate and improve the realism and clarity of images. In addition, they help to reduce design costs and increase the market competitiveness of art works [43-44]. These findings support the importance of technological innovation within the field of art and design and provide designers with new tools and methods to achieve a more efficient and diverse creative process.

VII. RESEARCH RESULTS

The research results of this paper show that the application of deep neural network and human-computer interaction technology in the field of art design has greatly improved the quality and efficiency of design work. In particular, the automatic text-to-image conversion using Cycle GAN technology reduces the content loss rate by 74.5% compared with the traditional manual method, and increases the peak signal-to-noise ratio of image generation by 57.9%. In addition, through interactive technologies such as virtual reality and augmented reality, the display effect of art works and the immersive experience of the audience have been significantly enhanced. These technologies not only optimize the creative process, reduce costs, but also enhance the market competitiveness of art works and audience participation, and bring innovative development opportunities for the art design industry.

VIII. CONCLUSION

Through empirical analysis, this study verifies the significant effectiveness of deep neural networks and human-computer interaction technology in improving the quality and efficiency of art design. In particular, Cycle GAN technology reduces content loss in the automatic conversion of text to images and enhances the display of artworks and the immersive experience of the audience through interactive technologies such as virtual reality. Based on these findings, it is recommended that art and design practitioners actively adopt these advanced technologies in order to optimize the creative process, broaden the boundaries of creative expression, and improve the market competitiveness of their works. At the same time, further exploration of the application potential of these technologies in different art and design scenarios is encouraged to promote the continuous innovation and development of the art and design industry.

REFERENCES

[1] Changsheng WANG. "AI-driven digital image art creation: methods and case analysis". Chinese Journal of Intelligent Science and Technology 5.3 (2023): 406-414.

[2] Anantrasirichai, Nantheera, and David Bull. "Artificial intelligence in the creative industries: a review." Artificial intelligence review 55.1 (2022): 589-656.

[3] Vartiainen, Henriikka, and Matti Tedre. "Using artificial intelligence in craft education: crafting with text-to-image generative models." Digital Creativity 34.1 (2023): 1-21.

[4] Wang, Xiaolong, and Ling Cai. "Application of artificial intelligence in 6G internet of things communication in interactive installation art." International Journal of Grid and Utility Computing 13.2-3 (2022): 195-203.

[5] Cetinic, Eva, and James She. "Understanding and creating art with AI: Review and outlook." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18.2 (2022): 1-22.

[6] Mayo, Sherry. "Co-creating with AI in Art Education: On the Precipice of the Next Terrain." Education Journal 12.3 (2024): 124-132.

[7] Li, Sixian. "Research on The Innovative Development of Digital Media Art in The Era of Artificial Intelligence." Frontiers in Computing and Intelligent Systems 8.3 (2024): 33-36.

[8] Nie, Zexian, Ying Yu, and Yong Bao. "Application of human–computer interaction system based on machine learning algorithm in artistic visual communication." Soft Computing 27.14 (2023): 10199-10211.

[9] Huang, Lihua, and Peng Zheng. "Human-computer collaborative visual design creation assisted by artificial intelligence." ACM Transactions on Asian and Low-Resource Language Information Processing 22.9 (2023): 1-21.

[10] Liu, Lina. "The artistic design of user interaction experience for mobile systems based on context-awareness and machine learning." Neural Computing and Applications 34.9 (2022): 6721-6731.

[11] Guo, Qiongqiong, and Guofeng Ma. "Exploration of human-computer interaction system for product design in virtual reality environment based on computer-aided technology." Computer-Aided Design & Applications 19.S5 (2022): 87-98.

[12] Li Xiong, Su Jian-Ning, and Zhang Zhi-Peng." Product Conceptual Sketch Generation Design Using Deep Learning." JOURNAL OF MECHANICAL ENGINEERING 59.11 (2023): 16-30.

[13] Amin, Mostafa M., Erik Cambria, and Björn W. Schuller. "Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT." IEEE Intelligent Systems 38.2 (2023): 15-23.

[14] Xu, Wei. "Toward human-centered AI: a perspective from human-computer interaction." interactions 26.4 (2019): 42-46. https://doi.org/10.1145/3328485.

[15] Ren, Fuji, and Yanwei Bao. "A review on human-computer interaction and intelligent robots." International Journal of Information Technology & Decision Making 19.01 (2020): 5-47. https://doi.org/10.1142/S0219622019300052.

[16] McNicol, Sarah. "Using participant-created comics as a research method." Qualitative Research Journal 19.3 (2019): 236-247. https://doi.org/10.1108/QRJ-D-18-00054.

[17] Kuttner, Paul J., Marcus B. Weaver-Hightower, and Nick Sousanis. "Comics-based research: The affordances of comics for research across disciplines." Qualitative Research 21.2 (2021): 195-214. https://doi.org/10.1177/1468794120918845.

[18] Chessa, Manuela,Guido Maiello,Alessia Borsari,Peter J. Bex. "The perceptual quality of the oculus rift for immersive virtual reality." Human–computer interaction 34.1 (2019): 51-82. https://doi.org/10.1080/07370024.2016.1243478.

[19] Cetinic, Eva, and James She. "Understanding and creating art with AI: Review and outlook." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18.2 (2022): 1-22. https://doi.org/10.1145/3475799.

[20] Abdumutalibovich, Ashurov Marufjon. "Working on the Artistic Characteristics of Performance in the Teaching of Instruments and Ensemble for Students of Higher Education Music." International Journal on Integrated Education 4.11 (2021): 38-41. https://creativecommons.org/licenses /by/4.0.

[21] Wu, Shiqing,Zhonghou Wang,Bin Shen, Jia-Hai Wang,Li Dongdong . "Human-computer interaction based on machine vision of a smart assembly workbench." Assembly Automation 40.3 (2020): 475-482. https://doi.org/10.1108/AA-10-2018-0170.

[22] Wu, Wei-Long,Yen Hsu, Qi-Fan Yang,Jiang-Jie Chen,Morris Siu-Yung Jong. "Effects of the self-regulated strategy within the context of spherical video-based virtual reality on students' learning performances in an art history class." Interactive Learning Environments 31.4 (2023): 2244-2267. https://doi.org/10.1080/10494820.2021.1878231.

[23] Hermus, Margot, Arwin van Buuren, and Victor Bekkers. "Applying design in public administration: a literature review to explore the state of the art." Policy & Politics 48.1 (2020): 21-48. https://doi.org/10.1332/030557319X15579230420126.

[24] Azis, Nur, Azizah Silfa Azzahra, and Arianto Muditomo. "Analysis Of Human Computer Interaction Approach In Pospay Application." Jurnal Mantik 6.2 (2022): 1956-1963. https://doi.org/10.35335/mantik.v6i2.2712.

[25] Liang, Calvin A., Sean A. Munson, and Julie A. Kientz. "Embracing four tensions in human-computer interaction research with marginalized people." ACM Transactions on Computer-Human Interaction (TOCHI) 28.2 (2021): 1-47. https://doi.org/10.1145/3443686.

[26] Tian, Xi. "Research on the Application of Interaction Design in Museum Exhibition." Journal of Global Humanities and Social Sciences 3.2 (2022): 28-29.

[27] Howard, Matt C. "Virtual reality interventions for personal development: A meta-analysis of hardware and software." Human–Computer Interaction 34.3 (2019): 205-239. https://doi.org/10.1080/07370024.2018.1469408.

[28] Peng, Mengqi, Jun Xing, and Li-Yi Wei. "Autocomplete 3D sculpting." ACM Transactions on Graphics (ToG) 37.4 (2018): 1-15. https://doi.org/10.1145/3197517.3201297.

[29] Qiao, Xiuquan, Pei Ren,Schahram Dustdar,Ling Liu,Huadong Ma,Junliang Chen. "Web AR: A promising future for mobile augmented reality—State of the art, challenges, and insights." Proceedings of the IEEE 107.4 (2019): 651-666. DOI: 10.1109/JPROC.2019.2895105.

[30] Parmaxi, Antigoni, and Alan A. Demetriou. "Augmented reality in language learning: A state-of-the-art review of 2014–2019." Journal of Computer Assisted Learning 36.6 (2020): 861-875. https://doi.org/10.1111/jcal.12486.

[31] Tom Dieck, M. Claudia, Timothy Hyungsoo Jung, and Dario tom Dieck. "Enhancing art gallery visitors' learning experience using wearable augmented reality: generic learning outcomes perspective." Current Issues in Tourism 21.17 (2018): 2014-2034. https://doi.org/10.1080/13683500.2016.1224818.

[32] Aryanta, I. Ketut, Ni Wayan Sitiari, and Putu Ngurah Suyatna Yasa. "Influence of Motivation on Job Stress, Job Satisfaction and Job Performance at Alam Puri Villa Art Museum and Resort Denpasar." Jurnal Ekonomi & Bisnis JAGADITHA 6.2 (2019): 113-120. https://doi.org/10.22225/jj.6.2.1353.113-120.

[33] Kotis, Konstantinos, Sotiris Angelis,Maria Chondrogianni, Efstathia Marini. "Children's art museum collections as Linked Open Data." International Journal of Metadata, Semantics and Ontologies 15.1 (2021): 60-70. https://doi.org/10.1504/IJMSO.2021.117107.

[34] Soler Gallego, Silvia. "Audio descriptive guides in art museums: A corpus-based semantic analysis." Translation and Interpreting Studies 13.2 (2018): 230-249. https://doi.org/10.1075/tis.00013.sol.

[35] Daly, Nathan S., Michelle Sullivan, Lynn Lee, Karen Trentelman. "Multivariate analysis of Raman spectra of carbonaceous black drawing media for the in situ identification of historic artist materials." Journal of Raman Spectroscopy 49.9 (2018): 1497-1506. https://doi.org/10.1002/jrs.5417.

[36] Serafin, Stefania, Michele Geronazzo,Cumhur Erkut,Niels C. Nilsson,Rolf Nordahl. "Sonic interactions in virtual reality: State of the art, current challenges, and future directions." IEEE computer graphics and applications 38.2 (2018): 31-43. DOI: 10.1109/MCG.2018.193142628.

[37] Alhayani, Bilal SA, and Haci Llhan. "RETRACTED ARTICLE: Visual sensor intelligent module based image transmission in industrial manufacturing for monitoring and manipulation problems." Journal of Intelligent Manufacturing 32.2 (2021): 597-610. https://doi.org/10.1007/s10845-022-02063-3.

[38] Katona, Jozsef. "A review of human–computer interaction and virtual reality research fields in cognitive InfoCommunications." Applied Sciences 11.6 (2021): 2646. https://doi.org/10.3390/app11062646.

[39] Sutcliffe, Alistair G.,C. Poullis,A. Gregoriades,I. Katsouri,I. Katsouri,I. Katsouri. "Reflecting on the design process for virtual reality applications." International Journal of Human–Computer Interaction 35.2 (2019): 168-179. https://doi.org/10.1080/10447318.2018.1443898.

[40] Chakraborty, Biplab Ketan,Debajit Sarma, M.K. Bhuyan, Karl F MacDorman. "Review of constraints on vision-based gesture recognition for human–computer interaction." IET Computer Vision 12.1 (2018): 3-15. https://doi.org/10.1049/iet-cvi.2017.0052.

[41] Lv, Zhihan. "Virtual reality in the context of Internet of Things." Neural Computing and applications 32.13 (2020): 9593-9602. https://doi.org/10.1007/s00521-019-04472-7.

[42] Kim, Yong Min, Ilsun Rhiu, and Myung Hwan Yun. "A systematic review of a virtual reality system from the perspective of user experience." International Journal of Human–Computer Interaction 36.10 (2020): 893-910. https://doi.org/10.1080/10447318.2019.1699746.

[43] Alahira, Joshua, Nwakamma Ninduwezuor-Ehiobu, Kehinde Andrew Olu-lawal, Emmanuel Chigozie Ani, Irunna Ejibe. "ECO-Innovative Graphic Design Practices: Leveraging Fine Arts to Enhance Sustainability in Industrial Design." Engineering Science & Technology Journal 5.3 (2024): 783-793. https://doi.org/10.51594/estj.v5i3.902.

[44] Blazhev, Boyan. "Artificial Intelligence and Graphic Design." Cultural and Historical Heritage: Preservation, Presentation, Digitalization (KIN Journal) 9.1 (2023): 112-130.https://doi.org/10.55630/KINJ.2023.090109.

Development of Real Time Meteorological Grade Monitoring Stations with AI Analytics

Adrian Kok Eng Hock, Chan Yee Kit, Koo Voon Chet
Faculty of Engineering and Technology, Multimedia University, Melaka Campus, Malaysia

***Abstract*—Air pollution comes in many forms and the basis of measure is the concentration of particles in the air. The quality of air depends on the quantity of pollution measured by a particle sensor that is accurate down to micron-meter consistencies. The size of the pollutants will be ingested by humans and cause respiratory problems and its effects on health conditions. The research will study the measurement of particles using multiple types of light scattering sensors and reference them to the accuracy of meteorological standards for precision in measurement. The sensors will be subjected to extreme conditions to gauge the repeatability and behavior and also long-term deployment usage. This study is required as when deployed on the field, dust particles will degrade the sensors over time. Early detection of sensor sensitivity and maintenance is therefore considered part of the research. Air particle data is volatile and dynamic over time and with that said, mass deployment of these sensors will give a better measurement of pollution data. However, with more and more data, standard statistics used show a basic level indicator and hence the idea of using machine learning algorithms as part of artificial intelligence (AI) processing is adapted for analyzing and also predicting particle data. There is a foreseeable challenge on this as there is no one machine learning for use only for this and multiple models are considered and gauged with the best accuracy using R2 value as low as 0.75 during the entire research. Lastly, with the seamless Internet of Things sensing architecture, the improved spatial data resolution will be improved and can be used to complement the current pollution measurement data for Malaysia in particular.**

Keywords—Air pollution; air particles; PM2.5; PM10; real time; light scattering sensor; neural networks; AI; machine learning; R2; IoT; WSM

I. Introduction

Pollutants in the air are not clearly visible and they come from many different sources. Air pollution is a mix of particles and gases that can reach harmful concentrations both outdoors and indoors [1]. Its effects range from disease risks such as respiratory to cardiovascular symptoms [2]. Smoke, mold, pollen, methane, and carbon dioxide are just a few examples of common pollutants that can affect the air surroundings [3].

The measurement of air pollution is measured in PM2.5 and PM10. PM refers to particulate matter which means particles in the air. The number 2.5 and 10 is the measurement in microns [4]. Particles the size of 2.5 microns and below are considered harmful as they could enter the blood stream and do damages years before any side effects occurs [5].

A limited number of meteorological stations in Peninsular Malaysia nationwide as shown in Fig. 1 gives a basic level of the air quality readings in the country (https://apims.doe.gov.my/). Systems in place are also located mostly in open areas far from densely populated places such as schools and factories. Data gathered are recorded, processed and shown but there are no algorithms for any preventive methods or predictive approach that could be performed. Only when pollution occurs, the data will be of interest to be monitored [6].

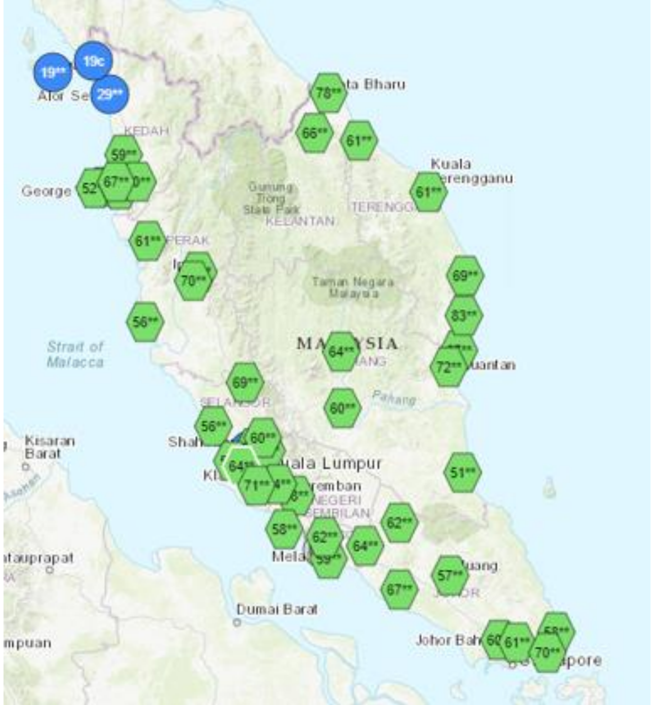


Fig. 1. Installation of meteorological stations in Malaysia.

The presented research focusses on market available particle sensors from various suppliers and the study of correlation of each sensor selection is benchmarked for its performance with the same environment and setup scenarios. The selection of particle sensor used will then be subjected to the meteorological reference to study its difference and an option to apply filtering mechanism for sensor data accuracy and traceability. Interface of the sensors and publishing the data to the cloud will be heart of the hardware being develop [7]. The hardware development will be in a form of an Internet of Things (IoT) node which is portable, energy efficient and ease of deployment and will be connected to the mobile network [8]. Altogether the hardware system will enable ease of deployment for mass monitoring and thus more perimeter of areas will be covered with real time data on concentration of the air pollutions [9]. As more systems deployed, the sensor data generated will be used for ingestion into the machine learning models for advanced processing on the prediction of the particulate matter. Machine learning models will be programed as a custom mode for particle data usage and models will be used based evaluated on based on the outcome of prediction accuracies and errors. In concluding the research presented, the IoT node will be able to measure particle sensor values accurately as of with meteorological standards with real time access and the overall system will be able to alert and provide predictive results based on artificial intelligence algorithms adapted that work behind the scenes [10].

Current and previous case studies and papers on air quality are based on meteorological sensors and as for Malaysia, there are not enough stations for monitoring dynamic conditions on the pollution that is causing the haze pollution. Therefore, additional stations whether it is meteorological or other particle sensor types are required. Cost-effective particle sensors that are used in some of the research are also not correlated to the meteorological side, hence it gives a sense of ambiguity on the results presented [11]. This research therefore would contribute and bridge the gap on these limitations using particle sensors that are based on the accuracy of the known standard with multiple deployment of sensors to complement the measurement data. Moreover, the systems designed and presented here are locally assembled.

The state of system design will be presented in the following chapters which include the study of sensors, the design of IoT and the correlation of the data with the Department of Environment (Malaysia) [12]. The backend of the systems utilizes the node and server architecture with web dashboard for first-hand information of IoT status. Machine learning models are then presented with comparative analysis against with R2 values as a scoring index on how well the models perform.

II. Methodology

A. Particle Sensor Selection and References

Particle measurement is related to the number of particles in an area at a point in time. These particles are extremely small; thus, it needs an accurate and precise sensor to be able to measure the particle sizes. The principle of particle measurement using a laser beam with particles blown over a detector and the reflection of the dust over time is counted and correlated with particle sizes and densities of the internal formulation of the sensor [13].

Three selections of particle sensors are considered from different manufacturers such as Honeywell, Sensirion and from China and shown in Fig. 2 [14]. The sensors are commonly used in industries and the reason of early selection. All three sensors are placed side by side and data is collected over a period of 45 days. This gives an equal level baseline on the particle data captured amongst the sensors being studied. The airflow, height, and simulated pollution levels are then conducted and data collected and compared. The co-relation of measured data is observed with an interval or 10 minutes and shown in Fig. 3 [15]. The hardware controlling the sensors were built to accommodate the three different sensors protocol and data is stored on the local storage. The dynamic range of all sensors are identical with co-relation of more than 90%. All three sensors are seen having the same identical behaviors. The resulting three sensors are equally suitable for deployment and the one being chosen to be used for this research is brand CN which comes with a protective outdoor casing. This sensor is chosen also due to the availability of the calibration certification ensuing the sensor is conformed to the specifications specified.

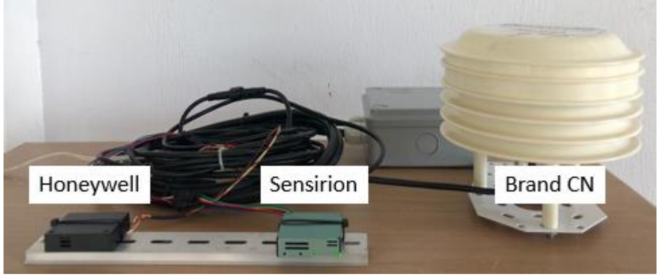


Fig. 2. Particle sensor selection and validation.

The brand CN sensor was then installed on the Department of Environment (DOE) in the Klang Valley meteorological station. There are two sets of the same sensors being placed for the study and comparison of the data repeatability alongside the reference standard. Data from the DOE particle sensor is then compared with the result of capture. Upon analysis, there are noticeable spikes in some of the range of the readings compared with the meteorological and the sensor used for the research. These spikes were due to the sensitivity and dynamic range of the sensors being used compared to the reference standard. Nevertheless, the trend of all the sensors is showing to be of identical trace.

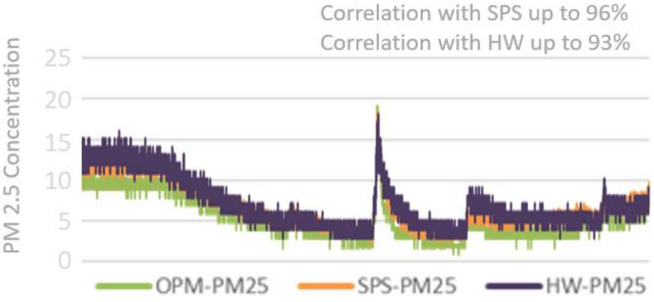


Fig. 3. Particle sensor correlation values and trend.

A study to improve the correlation of the sensor versus the DOE reference was embarked and Kalman filtering technique was adapted due to its usage for in noisy and dynamic sensors such as gyroscope sensors to normalize to the targeted reference [16]. The data that were captured were then passed to the Kalman algorithm to validate the outcome. The Kalman filtering consists of two main steps which are the prediction step and the update step with the following formula in Eq. (1).

$$x\hat{}k \mid k-1 = Fkx\hat{}k-1+Bkuk \qquad (1)$$

where, x^k ∣ k−1 is the predicted state estimate in time k and Fk is the state transition matrix which describes the dynamics of the system, x^k is the previous state estimate, Bk is the control input matrix and uk is the control input at time k.

The Kalman filter estimates the system’s states by predicting its next state using previous estimates and systems dynamics and minimizes the estimation error iteratively in turn makes the particle sensor referenced and improving its states to the known meteorological grade standard. The implementation of Kalman filter in the data processing of particles measurement data is shown in Fig. 4.

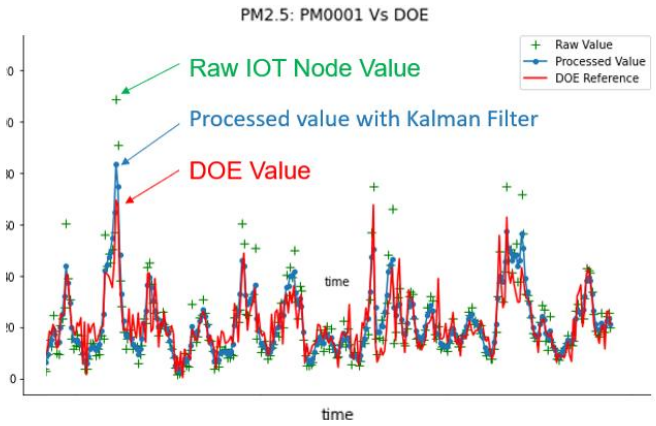


Fig. 4. Data processing using Kalman filtering algorithm.

B. Design of Wireless System Network (WSN) Hardware

Standalone API monitoring sensors form the basis of Wireless System Network (WSN) [17]. The development and requirement of the IoT hardware were carefully architected and specifications were designed to be adaptable and expandable to other types of sensors and connectivity. Other features such as low power usage <1Watts (idle) which is ideal for battery and solar power for off-the-grid deployment [18]. Splash-proof IP rating casings and connectors are also put in place for outdoor usage. The wireless communication used for the IoT nodes is Wi-Fi and option for 4G LTE mobile GSM networks for remote monitoring in rural areas. The communication protocols used were HTTP [19].

There are three iterations of prototypes of the IoT systems being designed and system validation and verification were carried out to finalize the first proper design of the deployment hardware [20]. Some problem that requires improvement such as RS485 protection circuitry were added as the IoT is likely to be damaged due to AC spikes and also lighting surges due to the system being placed open air in the field. Other improvements such as using a DC-DC converter for efficient power delivery versus a linear regulator were improved. This was crucial and required to be done as there are up to thirty-six systems that we build and to be deployed. The systems need to be robust and free of any maintenance or minimal problems on site. Fig. 5 showcases the hardware prototype being tested with the sensor connected to the 4G network while communicating with the server side. The hardware is then constructed and designed onto a PCB for final deployment with the proper housing and mounting in a pole with sensors altogether in one.

The IoT system generates several data from the sensors such as PM2.5, PM10, temperature, humidity, wind speed, direction and GPS location [21]. These sets of data are then pushed to the cloud server network for statistical data generation to build a web-based dashboard [22]. The interval of data transferred is set by default to be 30 seconds and can be triggered on demand as well for comprehensive monitoring if required. All the communication through and from the server is encrypted and error error-checking mechanism is used to ensure data transfer is successful and not corrupted. In the IoT system, there are various internal functions that are running in parallel, therefore the firmware that is built requires the system to be efficient and tasked deterministic and therefore a real time operating system is adapted with the structure is shown in Fig. 6. The structure itself is a simplified state machine architecture and this is the basic building block of the IoT system being built upon.

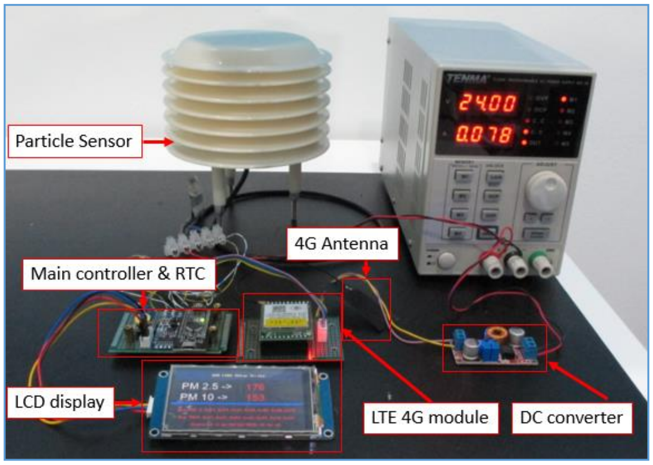


Fig. 5. Prototype IoT module.

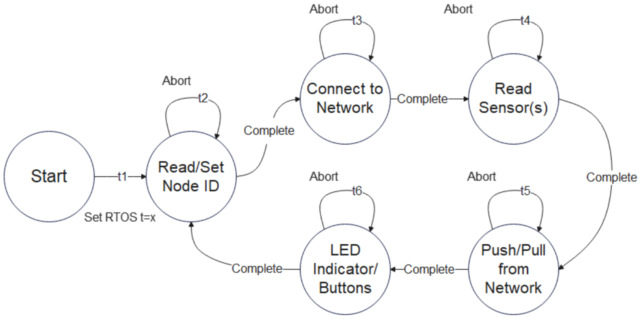


Fig. 6. Real-time operating system structure.

C. Data Processing

When the data is pushed to the cloud, a backend server-client dashboard is built in place to showcase the data based on time and day, and also a classification of API readings according to the standards on the IoT Nodes. This dashboard also serves as an indicator of the status of the connected sensor hardware that was deployed and activated on the field. This enables ease of measurement and also data storage from the sensor networks [23]. Basic analytical information based on mathematical statistic formulas is computed such as high, low, mean values of sensor data [24]. The client-side dashboard is shown in Fig. 7 which shows the real-time status of the connected IoT nodes along with the sensor measurement results.

Upon collecting and analyzing the measurements, the historical data will be used to serve as a base for AI modelling and processing. The AI algorithm will be used for predicting pollution ahead of time [25]. There are up to four types of machine learning models used and compared. They are the linear regression [26], ARIMA, Neural Prophet and the LSTM [27]. There are pros and cons in each algorithm being used and the basis of the modelling is time series type which is suited to the particle data being measured. Multiple machine learning models were studied and compared also as there the outcome of the prediction is different based on how the input dimensions and data types are being ingested and computed internally [28].

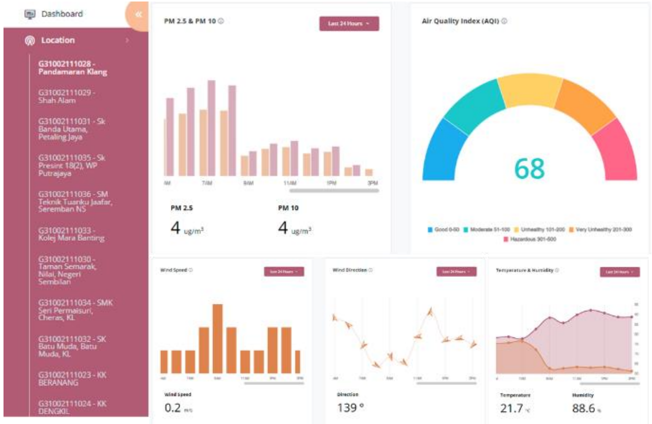

Dashboard
Location
G31002111028 - Pandamaran Klang
G31002111029 - Shah Alam
G31002111031 - Sk Banda Utama, Petaling Jaya
G31002111035 - Sk Presint 18(2), WP Putrajaya
G31002111036 - SM Teknik Tuanku Jaafar, Seremban NS
G31002111033 - Kolej Mara Banting
G31002111030 - Taman Semarak, Nilai, Negeri Sembilan
G31002111034 - SMK Seri Permaisuri, Cheras, KL
G31002111032 - SK Batu Muda, Batu Muda, KL
G31002111023 - KK BERANANG
G31002111024 - KK DENGKIL
PM 2.5 & PM 10
Last 24 Hours
7AM
9AM
11AM
1PM
3PM
PM 2.5
4 ug/m³
PM 10
4 ug/m³
Air Quality Index (AQI)
68
Good 0-50
Moderate 51-100
Unhealthy 101-200
Very Unhealthy 201-300
Hazardous 301-500
Wind Speed
Last 24 Hours
Wind Speed
0.2 m/s
Wind Direction
Last 24 Hours
Direction
139 °
Temperature & Humidity
Last 24 Hours
Temperature
21.7 °C
Humidity
88.6 %

Fig. 7. Web dashboard view on data from sensors.

III. Results

The overall research results focus on the machine learning computation outcome. The AI modelling uses the 80% training and 20% test case data and all the four models are fed with the same data period for comparison of outcome [29]. The data range used is a year collection of historical data which is from 15th March 2022 to 15th March 2023 on a particular IoT station (KK Nilai). The data that were ingested is from the particle measurement (PM2.5) sensor along with the time series points. There summary of outcome for the algorithm used is shown in Table I.

TABLE I. Comparison of AI Performance Index on Algorithm Used

	Linear Regression	ARIMA	Neural Prophet	LSTM
MAE	18.84	18.51	7.34	7.35
RMSE	22.43	22.00	10.26	9.72
R2	0.26	0.22	0.49	0.75

From the observation on the machine learning models used, it could be concluded that the statistical algorithm types are less accurate in prediction with the R2 value of below ⩽0.26. The Neural network types are suitable for prediction on this particle measurement (PM2.5) data and in the research presented as the R2 score is up to 0.75 with the LSTM method [30]. For pictorial visualization, the Fig. 8, Fig. 9, Fig. 10 and Fig. 11 shows a clear distinction on the statistical algorithm compared to neural types [31]. The prediction of the test data does not really contain any patterns and it could be said as of flat computation. On the other hand, for the neural network types, the prediction pattern can be seen and matching the flow of the test data [32].

Based on overall data and results, neural algorithm method can be seen as performing better as they are inspired by the structure and functioning of the human brain. It is a key component of machine learning, specifically a subset known as deep learning. Furthermore, neural networks are used for tasks such as pattern recognition, classification, regression, and more, by learning from data.

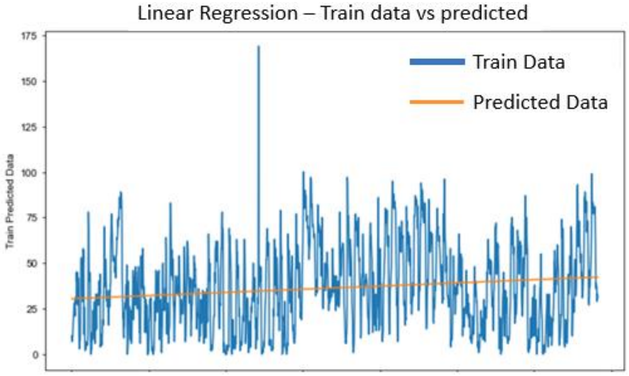


Fig. 8. Linear regression modelling.

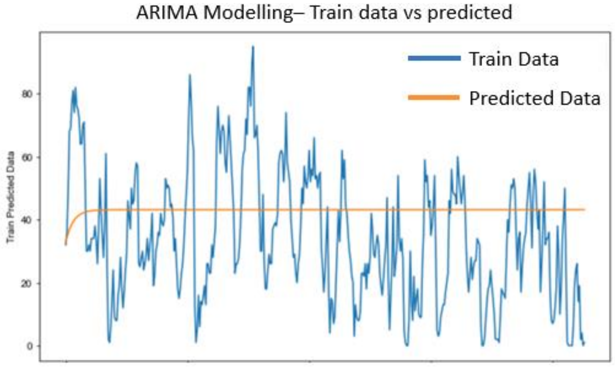


Fig. 9. ARIMA modelling.

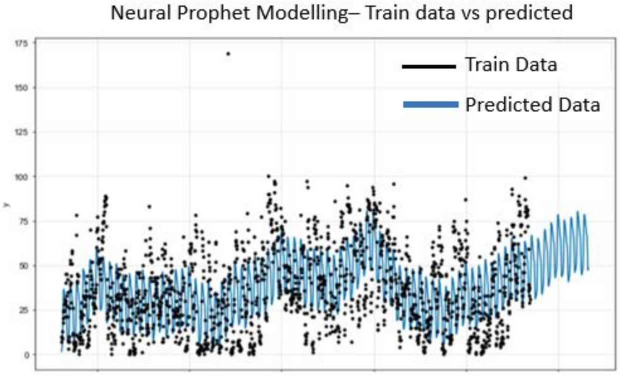


Fig. 10. Neural prophet modelling.

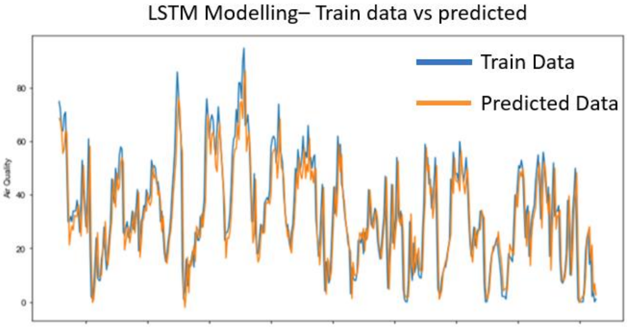


Fig. 11. LSTM modelling.

The LSTM model perform better is also due to the multi epoch used together with seasonality computation for ingestion while the rest uses the basis of time series modelling or single dimensional data for ingestion [33].

IV. Discussion

The successful integration of sensor selection, hardware design, and server-side processing has resulted in a comprehensive air quality monitoring system designed into an IoT based system. The deployment of this system on-site has demonstrated its capability to provide accurate, real-time air quality data, which is essential for informing public health strategies and environmental policies. Measurement of particle data is compared with DOE first hand to gauge the correlation and reproducibility of the measured air quality data [34]. Server based web API displays the IoT stations status with real-time on demand access using 4G GSM network. Future enhancements will focus on expanding the network of monitoring units and incorporating additional data sources to further improve prediction accuracy and system robustness.

There are two known limitations for the research which is on the long-term accuracy and maintenance of the sensors being used for the particle measurement. The Department of Environment (DOE) sensors are maintained regularly and calibrated to a traceable standard [35]. However, this sensitivity or performance of the sensor can be monitored using a machine learning algorithm using abnormal detection.

The second limitation is on the data processing with artificial intelligence algorithms. As the known research uses the data from meteorological sensors, there are years of historical data readily to be used and ingested into the algorithm [36]. The historical data from meteorological sensors contains additional sensor data such as pollution gases which is an advantage when using multivariate [37] AI modelling which serves as a multiple-dimensional processing and greatly influences the prediction outputs [38] [39].

V. Conclusion

From the sensor IoT node deployment to data generation and retrieval, this research presents the sensors being used is able to perform with high standards of measurement by referencing to a known standard. The IoT system hardware itself plays the main part in where sensors are being interconnected and data is fed back to the cloud for processing and storage. A step further in processing the data using AI algorithms is experimented and classification of the trend of data with anomalies seen of the data being produced and rectified with normalization [40]. The demonstrated predicted results using machine algorithm with an RMSE of 9.72 for LSTM is suitable for forecasting and detection of pollution levels ahead in time [41]. Hence, this information could be fed to other broadcast networks for notification. Lastly the presented system is not constrained to air pollution only and can be easily adapted to other types of pollution that compromise quality of life in urban areas, e.g. noise pollution, hazardous gases and more [42].

Acknowledgment

The author wishes to extend his profound gratitude to the Department of Environment Malaysia (DOE) on permission for installing the IoT sensors and data shared during the research. This work was partly financed by iRadar Sdn. Bhd. on hardware, engineer access and countless arrangements for deployment and setup on sites.

References

[1] MacNee, W. and Donaldson, K. (2003). Mechanism of lung injury caused by PM10 and ultrafine particles with special reference to COPD. European Respiratory Journal 21(40): 47s-51s.

[2] Khan S, Sahu V, Kumar N, Gurjar B. Particulate matters deposition in the human respiratory system: A health risk assessment at a technical university. JAPH. 2024;9(1):1-14.

[3] Ranathunga N, Perera P, Nandasena S, Sathiakumar N, Kasturiratne A, Wickremasinghe R. (2019). Effect of household air pollution due to solid fuel combustion on childhood respiratory diseases in a semi urban population in Sri Lanka. BMC Pediatr. 2019;19(1):306.

[4] Brunekreef, B. and Holgate, S. T. (2002). Air pollution and health. The lancet 360 (9341): 1233-1242.

[5] Zheng, Yijing, Ooi, Maggie C. G., Juneng, Liew, Wee, Hin B., Latif, M. Talib., M. Nadzir, M. Shahrul, Hanif, Norfazrin, Chan, Andy, Li, Li, Ahmad, Norfazilah, Tangang, Fredolin. (2023). Assessing the impacts of climate variables on long-term air quality trends in Peninsular Malaysia. The Science of the total environment. (901. 166430. 10.1016/j.scitotenv.2023.166430).

[6] Tian Y, Liu H, Zhao Z, Xiang X, et al. (2018). Association between ambient air pollution and daily hospital admissions for ischemic stroke: a nationwide time-series analysis. PLoS Med. 2018;15(10):e1002668.

[7] K. K. Johnson, M. H. Bergin, A. G. Russell, and G. S. W. Hagler (2018). Field test of several low-cost particulate matter sensors in high and low concentration urban environments. Aerosol and Air Quality Research.

[8] Mladen K., Biljana R. S., Kire T. (2019). Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization.IEEE EUROCON 2019 -18th International Conference on Smart Technologies.

[9] Tim K., Rea D., Samantha D., Royal G., Martha K., Melissa L., John P., Alison S., Kate T. (2023). Low-cost PM2.5 sensors can help identify driving factors of poor air quality and benefit communities. ISSN 2405-8440 (Volume 9, Issue 9, 2023).

[10] Tammy Noergaard (2013). Embedded Systems Architecture: A Comprehensive Guide for Engineers and Programmers. ISBN-10 0123821967.

[11] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38(18), 2895-2907.

[12] Hamza A. I. and Azman A (2015). Air quality pattern assessment in Malaysia using multivariate techniques (Malaysian Journal of Analytical Sciences ISSN 1394 – 2506).

[13] M. Budde, M. Köpke, and M. Beigl, (2016). Design of a light-scattering particle sensor for citizen science air quality monitoring with smartphones: tradeoffs and experiences. Pro Science 3: Conference Proceedings: 2nd International Conference on Atmospheric Dust - DUST 2016.Castellaneta Marina, Italy.

[14] Marek B., Piotr B., Anetta D, and Piotr M (2018). Evaluation of Low-Cost Sensors for Ambient PM2.5 Monitoring (Journal of Sensors, Volume 2018, Article ID 5096540).

[15] Ai S, Wang C, Qian ZM, Cui Y, Liu Y, Acharya BK, Sun X, Hinyard L, Jansson DR, Qin L, Lin H (2019). Hourly associations between ambient air pollution and emergency ambulance calls in one central Chinese city: Implications for hourly air quality standards (Journal Science of The Total Environment, Volume 696, 15 December 2019).

[16] Lai, Xiaozheng & Yang, Ting & Wang, Zetao & Chen, Peng. (2019). IoT Implementation of Kalman Filter to Improve Accuracy of Air Quality Monitoring and Prediction. Applied Sciences. 9. 1831. 10.3390/app9091831.

[17] Hermann Kopetz (2011). Real-Time Systems: Design Principles for Distributed Embedded Applications ISBN 9781441982377.

[18] Bilal, K.; Khalid, O.; Erbad, A.; Khan, S.U.(2018). Potentials, trends, and prospects in edge technologies: Fog, cloudlet, mobile edge, and micro data centers. Comput. Net.

[19] Sari, R.F., Rosyidi, L., Susilo, B., Asvial, M. (2021). A Comprehensive Review on Network Protocol Design for Autonomic Internet of Things. Information 2021, 12, 292. https://doi.org/10.3390/info12080292.

[20] Cuno Pfister (2011). Getting Started with the Internet of Things: Connecting Sensors and Microcontrollers to the Cloud (Make: Projects) 1st Edition. ISBN 978-1449393571.

[21] Pallavi S. and Smruti R. S. (2017). Internet of Things: Architectures, Protocols, and Applications (Journal of Electrical and Computer Engineering, Volume 2017 Article ID 9324035).

[22] Simwanda, Matamyo. (2019). Air Quality Monitoring Using Remote Sensing and GIS Applications: Exploring the Potential for Developing African Countries. Conference: WORLD ENVIRONMENT DAY COMMEMORATION 2019.

[23] Kevin Dooley (2015). The No-Sweat Guide to Network Topology.Auvik (www.auvik.com).

[24] Stawowy, M., Olchowik, W., Rosiński, A., Dąbrowski, T. (2021). The Analysis and Modelling of the Quality of Information Acquired from Weather Station Sensors. Remote Sens. 2021, 13, 693. https://doi.org/10.3390/rs13040693.

[25] J. K. Sethi and M. Mittal (2020). Analysis of Air Quality using Univariate and Multivariate Time Series Models. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 823-827, doi: 10.1109/Confluence47617.2020.9058303.

[26] Joseph M. Hilbe (2015). Practical Guide to Logistic Regression. ISBN 978-1-4987-0958-3.

[27] Zhao, Z., Wu, J., Cai, F. et al.(2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. Sci Rep 13, 1015 (2023). https://doi.org/10.1038/s41598-023-28287-8.

[28] Azid, A., Juahir, H., Toriman, M.E. (2014). Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: a Case Study in Malaysia (Water Air and Soil Pollution · July 2014).

[29] Prachi, Kumar N.and Matta, Gagan (2011) Artificial neural network applications in air quality monitoring and Management. International Journal for Environmental Rehabilitation and Conservation Volume II No. 1 2011 [30 – 64].

[30] Doreena D., Hafizan J., Mohd T. L., Sharifuddin M. Z., and Ahmad Z. A. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis (Atmospheric Environment 60 2012 172e181).

[31] Hamza A. I. and Azman A (2015). Air quality pattern assessment in Malaysia using multivariate techniques (Malaysian Journal of Analytical Sciences ISSN 1394 – 2506).

[32] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstorm (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK.

[33] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal (2015). Long short-term memory networks for anomaly detection in time series, European Symposium on Artificial Neural Networks, vol. 89.

[34] Abdullah, A. M., Abu Samah, M. A., and Jun, T. Y. (2012). An overview of the air pollution trend in Klang Valley, Malaysia. Open Environ Science 6:13-19.

[35] Department of Environment Malaysia (DOE) (2021). Malaysia Environmental Quality Report, Ministry of Science, Technology and Environment, Kuala Lumpur.

[36] Voukantsis, Dimitrios & Karatzas, Kostas & Kukkonen, Jaakko & Räsänen, Teemu & Karppinen, Ari & Kolehmainen, Mikko. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinkiinterface,".

[37] H. W. Xu, W. X. Chen, N. W. Zhao et al. (2018). Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications, in Proceedings of the World Wide Web Conference. Lyon, FranceA. Ito, S. Wakamatsu, T. Morikawa, S. Kobayashi (2021). 30 Years of Air Quality Trends in Japan. Atmosphere 2021. MDPI.

[38] Shaddick G, Thomas ML, Amini H, Broday D, Cohen A, Frostad J, Green A, Gumy S, Liu Y, Martin RV, et al. (2018). Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. Environ Sci Technol. 2018;52(16):9069–78.

[39] Y. J. Zheng, C.L. Zhang, and H. Y. Wang (2020). MTAD-TF: Multivariate Time Series Anomaly Detection Using the Combination of Temporal Pattern and Feature Pattern. Hindawi. Volume 2020 | Article ID 8846608.

[40] Mutalib, S. N. S. A., Juahir, H., Azid, A., Sharif, S. M., Latif, M. T., Aris, A. Z., Zain, S. M. and Dominick,D., (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. Environmental Science, Processes & Impacts 15(9): 1717-1728.

[41] Shang, L., Huang, L., Yang, W. et al.(2019). Maternal exposure to PM2.5 may increase the risk of congenital hypothyroidism in the offspring: a national database-based study in China. BMC Public Health 19, 1412.

[42] Dauchet L, Hulo S, Cherot-Kornobis N, Matran R, Amouyel P, Edme JL, Giovannelli J.(2018). Short-term exposure to air pollution: associations with lung function and inflammatory markers in non-smoking, healthy adults. Environ Int. 2018;121(Pt 1):610–9.

CNN-Based Salient Target Detection Method of UAV Video Reconnaissance Image

Li Na
Hainan Vocational College of Political Science and Law, Haikou City, Hainan Province, 570100, China

***Abstract*—In order to address the challenges of image complexity, capturing subtle information, fluctuating lighting, and dynamic background interference in drone video reconnaissance, this paper proposes a salient object detection method based on convolutional neural network (CNN). This method first preprocesses the drone video reconnaissance images to remove haze and improve image quality. Subsequently, the Faster R-CNN framework was utilized for detection, where in the Region Proposal Network (RPN) stage, the K-means clustering algorithm was used to generate optimized preset anchor boxes for specific datasets to enhance the accuracy of target candidate regions. The Fast R-CNN classification loss function is used to distinguish salient target regions in reconnaissance images, while the regression loss function precisely adjusts the target bounding boxes to ensure accurate detection of salient targets. In response to the potential failure of Faster R-CNN in extreme situations, this paper innovatively introduces a saliency screening strategy based on similarity analysis to finely screen superpixels, preliminarily locate target positions, and further optimize saliency object detection results. In addition, the use of saturation component enhancement and brightness component dual frequency coefficient enhancement techniques in the HSI color space significantly improves the visual effect of salient target images, enhancing image clarity while preserving the natural and soft colors, effectively improving the visual quality of detection results. The experimental results show that this method exhibits significant advantages of high accuracy and low false detection rate in salient object detection of unmanned aerial vehicle (UAV) video reconnaissance images. Especially in complex scenes, it can still stably and accurately identify targets, significantly improving detection performance.**

Keywords*—Regional convolutional neural network; K-means clustering; UAV reconnaissance image; salient target detection; task loss function

I. INTRODUCTION

The salient target detection of UAV video reconnaissance image is a method that uses computer vision and depth learning technology to quickly identify and locate key targets from the image taken by UAV. It automatically detects salient targets in the image, such as people, vehicles, buildings, etc., by analyzing the texture, color, shape, and other characteristics of the image [1]. The detection of the salient target of a UAV video reconnaissance image has extensive application value in many fields, such as safety monitoring, public safety, environmental monitoring, disaster rescue, etc. [2]. Through real-time monitoring and early warning, this technology can help people find abnormal situations in time, improve monitoring accuracy and response speed, and provide strong support for preventing and responding to various emergencies [3]. Therefore, the detection of salient targets of UAV video reconnaissance images is of great significance in ensuring public safety and improving emergency response capability.

Jirayupat C, et al. proposed a method of detecting salient targets in UAV video reconnaissance images based on chromatography-mass spectrometry [4]. This method pre-processes the image using the chromatography-mass spectrometry method, extracts the salient target features in the image, and then detects the target using the trained machine learning model. By combining image processing and machine learning, the salient target in the UAV video can be quickly and accurately recognized. This method involves the integration of many technologies, including image processing, chromatography, mass spectrometry, and machine learning. This requires researchers to have extensive professional knowledge and skills, and it is difficult to achieve and optimize. Cherri A K et al. proposed a method of detecting salient targets of UAV video reconnaissance images based on a joint transform correlator [5]. This method reduces the size of the UAV video reconnaissance image to speed up the processing time and increase the storage capacity of the recognition system. By using fringe adjustment joint transform correlation technology, multiple targets based on compression are successfully detected. The use of shift phase encoding and reference phase encoding technology can eliminate the false detection and missed detection caused by multiple expected and unwanted targets, thus realizing the detection of salient targets of UAV video reconnaissance images more accurately. This method reduces the size of the UAV video reconnaissance image, which will lead to the reduction of image resolution and will affect the accuracy of salient target detection and detail recognition. Iván García-Aguilar and others proposed the method of detecting salient targets of UAV video reconnaissance images using CNN and super-resolution [6]. This method improves the resolution of UAV video images through super-resolution technology and then uses a convolutional neural network (CNN) for feature extraction and target detection. Train CNN models to identify and locate salient targets. Combining the advantages of deep learning and super-resolution technology, it can accurately detect salient targets in UAV video reconnaissance images of low resolution. In this method, the super-resolution technology is sensitive to the noise in the image, which will affect the accuracy and reliability of target detection. Ezequiel López-Rubio a b and others proposed a method of detecting salient targets of UAV video reconnaissance images based on deep learning and PTZ camera controller [7]. This method uses deep learning to detect the salient target of a UAV video reconnaissance image. This method includes three modules: target detection, salient target detection, and PTZ camera controller. The deep learning network is used to detect the target in the scene. The salient target detection module automatically detects the salient target using the Dirichlet distributed hybrid model, while the PTZ camera controller allows it to follow and focus on the salient target to achieve the salient target detection of UAV video reconnaissance image. This method uses deep learning to detect the salient target, which faces a challenge in the real-time environment. On low-performance hardware, the processing speed cannot meet the real-time requirements, resulting in the limitation of this method in practical applications.

Faster R-CNN can effectively deal with complex factors in drone video reconnaissance images, such as subtle information, lighting changes, and dynamic backgrounds, ensuring high-precision salient object detection under various conditions. Therefore, this paper proposes a CNN-based method of detecting salient targets of UAV video reconnaissance images. By performing haze removal operations in the preprocessing stage, image quality is improved, and the Faster R-CNN framework combined with K-means clustering is used to optimize anchor boxes, achieving accurate detection of targets in complex scenes. Specifically, in response to the limitations of Faster R-CNN in extreme situations, this paper introduces a superpixel saliency filtering strategy and combines it with image enhancement techniques in the HSI color space. This not only significantly improves the accuracy of object detection and image clarity, but also ensures the authenticity and softness of colors, providing more reliable and efficient technical support for drone video reconnaissance. The specific research approach is as follows:

1) After removing haze from drone video surveillance images, a salient object detection framework for drone video surveillance images is constructed.

2) Preset anchor boxes in RPN and determine the final target bounding box through a multi task loss function.

3) Significant object detection calculation is used for undetectable targets, including calculating target similarity and target connectivity maps to obtain target probabilities.

4) The enhancement of saturation and brightness components based on HSI color space improves the accuracy of salient object detection.

II. Salient Target Detection of UAV Video Reconnaissance Image

A. Dehazing of UAV Video Reconnaissance Image

UAV imaging will be interfered with a variety of factors. Due to the presence of haze in the air or chemical particles and other particles, in the propagation process, the light will be scattered or refracted to varying degrees, resulting in some detailed information not being received by the sensor [8]. The haze image degradation model formula is expressed by Eq. (1):

$$A(x) = Z(x)t(x) + \rho[1 - t(x)] \quad (1)$$

In the equation, $A(x)$is the haze image degradation model, $Z(x)$ is a realistic image, $t(x)$ is medium transmission, ρ is atmospheric scattered light.

Light sources are scattered or reflected by the haze present in the air as they travel through the atmosphere. Then the light reflected through the object is also scattered or reflected by the haze [9]. After that, only the energy of$Z(x)t(x)$can reach the sensor. At the same time, the atmospheric scattered lightρgenerated due to the scattering of various particles in the air will also be absorbed by the sensor.

The medium transmission is expressed by Eq. (2):

$$t(x) = h^{-\varepsilon d(x)} \quad (2)$$

The medium transmission is related to the scattering coefficient $\varepsilon$of the atmosphere and also related to the distance $d(x)$from the sensor to the object. When$d(x)$goes to infinity $t(x)$ is close to zero. Then the atmospheric scattered light can be expressed by Eq. (3):

$$\rho = A(x), d(x) \to inf \quad (3)$$

In practical imaging, instead of relying on equations, obtaining the atmospheric scattered light$\rho$is rather a more stable calculation based on Eq. (4). $d(x)$cannot be infinite, but gives a very low transmittance t_0.

$$\rho = \max_{y \in \{x | t(x) \leq t_0\}} A(y) \quad (4)$$

If the atmospheric light in the corresponding area is given, the medium transmission can be calculated based on physical principles, and a clear image can be obtained after the haze is removed. The following procedures are all performed on the dehazed UAV video reconnaissance image.

B. Target Detection of UAV Video Reconnaissance Image Based on Faster R-CNN

After getting a clear image after removing the haze, the corresponding targets can be detected from it. For UAV video reconnaissance, multiple targets need to be monitored at the same time. To detect multiple targets at the same time and ensure detection accuracy, Faster R-CNN is used to build a detection framework, and training is used to adapt to different scenes and target types, to achieve the target detection of UAV video reconnaissance images.

1) Target detection framework: Faster R-CNN (Faster Region-based Convolutional Neural Network) is a new generation of target detection model that optimizes CNN [10]. Faster R-CNN can be seen as a combination of an RPN (Region Proposal Network) and a Fast R-CNN (Region-based Convolutional Network), which share the basic convolutional network [11]. RPN is responsible for generating high-quality proposed target areas in UAV video reconnaissance images, and Fast R-CNN will complete the salient target classification and position regression of the proposed target areas [12]. The improved target detection framework in this paper is shown in Fig. 1. First, the anchor frame is reset by the clustering method to make the reference anchor frame more suitable for the target characteristics of the dataset. Second, a new full connection layer (behind the ROI layer) is added to the Fast R-CNN model to effectively improve the detection performance of the algorithm.

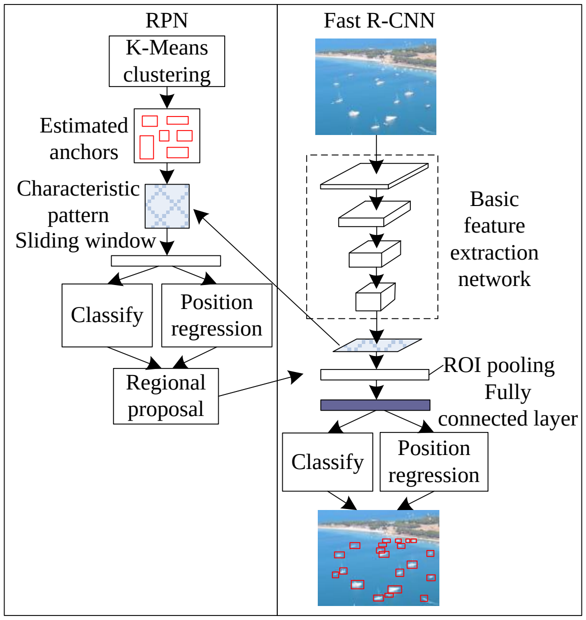


Fig. 1. Improved faster R-CNN target detection framework.

As displayed in Fig. 1, RPN uses a sliding window and anchor mechanism to generate a proposed target area of multi-scale. After the image passes through the basic feature extraction network, a convolution kernel of 3×3 slides on the feature map to get a feature vector of 512 dimensions (corresponding to the VGG model) each time, and then the vector is sent to the full connection layer: (1) Object/non-object binary layer, to predict whether there is a target in the window; (2) Position regression layer, to calculate the position correction of the anchor frame relative to the target boundary frame, and to obtain the position coordinates of the candidate frame. Fast R-CNN and RPN share the feature extraction network. First, ROI pooling is used to obtain the feature representation of each candidate frame, and then ROI features are sent to the full connection layer, and the features are sent to the two parallel task layers: (1) Softmax classification layer, to calculate the probability of candidate frames on the C+1(Class C target + background) class; (2) Position regression layer, to calculate the relative offset between the candidate frame and the target boundary frame, and to further correct the position of the predicted target.

2) Anchor frame setting: In the RPN model, anchor frames with three different aspect ratios are set according to experience. However, for different datasets, the scale of the anchor frame is inconsistent. Choosing an appropriate anchor frame can effectively improve the learning speed of network model training, and can also improve the target detection accuracy of UAV video reconnaissance images [13]. In this paper, K-Means clustering is used to select anchor frames. The algorithm flow is shown in Fig. 2.

- Collect the truth frames of all target samples in the UAV video reconnaissance image set $a = (xmaxmaxmin_{min})$, and set the number of clusters, i.e., the number of anchor frames k, and then randomly selected$k$samples as initial clustering centers;
- Calculate the distances between the remaining target samples and k centers, based on the Euclidean distance. The center with the smallest distance is used as the target sample class;
- Calculate the new clustering center according to the target classification results, and set the clustering center be o. The criterion function is calculated as shown in Eq. (5):

$$I(a,o) = \rho \sum_{i=1}^{k} \left\| a^i - o_c^i \right\|^2 \qquad (5)$$

- Repeat steps (2) and (3) until the error of the criterion function is within the allowable range, and end the loop to obtain the final target classification results.

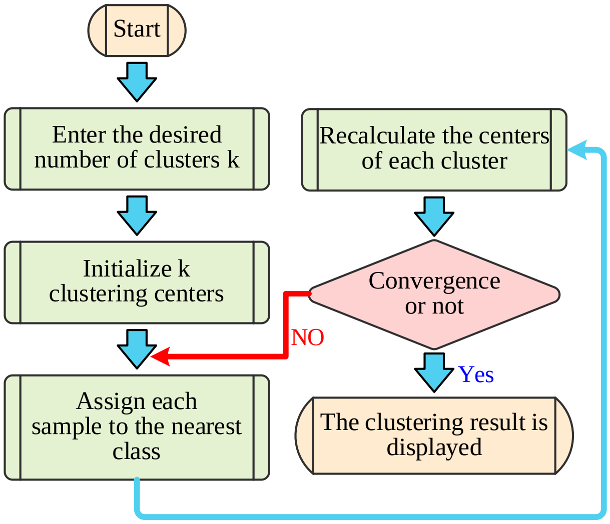

Start
Enter the desired number of clusters k
Initialize k clustering centers
Assign each sample to the nearest class
Recalculate the centers of each cluster
Convergence or not
NO
Yes
The clustering result is displayed

Fig. 2. Clustering flowchart of anchor frame.

3) Multitask loss function: The multitask loss function in the Fast R-CNN model is mainly composed of the classification loss function and regression loss function [14]. The classification loss function is used for the classification of salient targets in the proposed target areas of UAV video reconnaissance images, and the regression loss function is used for the regression of salient target boundary frames in UAV video reconnaissance images [15]. The loss function definition formula for a UAV video reconnaissance image is expressed by Eq. (6):

$$H[(r_i),(s_i)] = \frac{I(a,o)}{N_{cls}} \sum_i H_{cls}(r_i, r_i^*) \qquad (6)$$

$$+\alpha \frac{1}{N_{reg}} \sum_i r_i^* H_{reg}(s_i, s_i^*)$$

In the equation, r_i is the prediction probability of the i -th anchor, r_i^* is the predicted probability of the actual boundary frame $(GT, Grond \;\; Truth)$corresponding to the i -th anchor.

If the recall rate between the i -th anchor boundary frame andGT,$IoU > 0.7$ (at this point $r_i^* r_i^* = 1$), then the anchor is considered as a salient target. If $0.3 < IoU < 0.7$, then the anchor will not participate in the training. If $IoU < 0.3$ (at this point $r_i^* = 0$), then the anchor is considered as the background.

s_i is a vector, indicated as$s_i\{s_x, s_y, s_w, s_v\}$, corresponding to the four parametric coordinates of the boundary frame of the predicted salient target. s_x, $s_y$are corresponding to the central coordinates of the boundary frame of the salient target. s_w, s_v are corresponding to the width and height of the boundary frame of the salient target.$s_i^*$is the coordinate vector of the salient target GTcorresponding to the anchor.

The classified loss is the logarithmic loss of salient target class and non-salient target class, and the calculation formula is expressed by Eq. (7):

$$H_{cls}(r_i, r_i^*) = -lg[r_i^* r_i + (1 - r_i^*)(1 - r_i)] \quad (7)$$

The regression loss is calculated as shown in Eq. (8):

$$H_{reg}(s_i, s_i^*) = B(s_i * -s_i^*) \quad (8)$$

In the equations, B is the defined robust loss function $(smooth\ H_1)$, and the calculation formula is expressed by Eq. (9):

$$smooth\ H_1(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (9)$$

Using a four-dimensional vector$(x, y, w, v)$for the target display window, respectively represents the center point coordinates, width, and height of the target window. Through the regression learning of the salient target boundary frame in the UAV video reconnaissance image, a relational mapping is found, to obtain the regression target window G'which is close to the real target window G by input of the original anchor C through the mapping method, i.e., given $C = (C_x, C_y, C_w, C_v)$, and look for a relational mapping$f$, and make $f(C_x, C_y, C_w, C_v) = (G'_x, G'_y, G'_w, G'_v)$, of which$(G'_x, G'_y, G'_w, G'_v) \approx (G_x, G_y, G_w, G_v)$.

The following two methods of translation transformation and scaling transformation are used to implement the transition from anchor to approximate GT.

The translation transformation is calculated as shown in Eq. (10):

$$\begin{cases} G'_x = C_w d_x(C) + C_x \\ G'_y = C_v d_y(C) + C_y \end{cases} \quad (10)$$

The scaling transformation is calculated as shown in Eq. (11):

$$\begin{cases} G'_w = C_w\, exp[d_w(C)] \\ G'_v = C_v\, exp[d_v(C)] \end{cases} \quad (11)$$

The amount of translation(s_x, s_y), $(s_x^*, s_y^*)$and the scale factor(s_w, s_v), $(s_w^*, s_v^*)$are expressed by Eq. (12), (13), (14), and (15):

$$s_x = (x - x_c)/w_c; s_y = (y - y_c)/v_c \quad (12)$$

$$s_x^* = (x^* - x_c)/w_c; s_y^* = (y^* - y_c)/v_c \quad (13)$$

$$s_w = lg(w/w_c)\,; s_v = lg(v/v_c) \quad (14)$$

$$s_w^* = lg(w^*/w_c)\,; s_v^* = lg(v^*/v_c) \quad (15)$$

In the equations,x , $y$indicate the coordinates $x$and the coordinates $y$of the center of the predicted target boundary frame. w, v indicate the width and height of the target boundary frame. The coordinate parameters of the boundary frame of the anchor are respectively expressed asx_c , y_c , w_c , v_c . The coordinate parameters of the boundary frame of GT are respectively expressed asx^*, $*\, y^*$, w^*, v^*.

C. Detection of Salient Target of UAV Video Reconnaissance Image

The above calculation formula can be understood as, by the regression learning of the target boundary frame, that is, the regression from the anchor boundary frame to the nearbyGT boundary frame, a boundary frame of the regression target windowG' which is much closer to the actual target window$G$ is obtained. The target is detected according to the boundary frame, but due to some certain occlusion or mistaken recognition of the target as an unrecognized object in the UAV video detection, it is further optimized to achieve salient target detection.

1) Calculation of target similarity: By using Faster R-CNN, through the training of salient target detection in UAV video reconnaissance images, the identification of salient targets within the image can be efficiently identified [16], followed by the extraction of potential targets based on their distinctive features [17]. Following that, begin to create the similarity graph. Faster R-CNN exhibits a notable detection rate; however, it may struggle to detect salient targets in exceptional circumstances. If Faster R-CNN cannot identify the salient target, it will process the entire image as a target.

The likelihood of this window containing a target is denoted by the target similarity score. The pixel-level similarity score [18] is derived from the potential targets to determine the likelihood of a pixel being a component of the target. The pixel-level similarity score is determined as shown in Eq. (16) as follows:

$$PixObj(q) = \sum_{i=1}^{N} e_i\, L_i(x, y) \quad (16)$$

The equation, $e_i$indicates that if the pixel $q$is contained in the target window i detected by Faster R-CNN, $L_i$is a Gaussian filter window of equal dimension to the window, x, $y$is the relative coordinate of pixels$q$ in one of the detection windows, $N$is the number of possible target windows detected by Faster R-CNN.

The sum of the similarity scores of all pixels in the superpixel region is the similarity score of the current superpixel region, defined by Eq. (17).

$$Objectness(Tq_i) = \sum_{i \in M} PixObj(q_j) \quad (17)$$

In the equation, q_j is one pixel in the$i$ -th superpixel regionTq_i. The similarity graph is then further optimized using a threshold-based method, where the threshold is set as 1.5 times the quantity of pixels in the similarity graph divided by the overall number of pixels in the graph.

2) *Calculation of target connectivity graph*: The threshold similarity graph is only a part of the target superpixels that are roughly acquired. This paper adopts the "target connectivity" method, which assigns a value to the predicted target according to the salience value of the superpixel connectivity [19]. Build a graph using superpixels as nodes. The adjacent superpixel nodes have edges, and the weight of these edges is specified as the Euclidean space of the mean Lab color of the two nodes. The target connectivity of the i -th superpixel Tq_i is defined by Eq. (18):

$$F(Tq_i) = \frac{\sum_{m=1}^{N_1} d(Tq_i,Tq_m)\cdot\beta(Tq_m)}{\sum_{m=1}^{N_1} d(Tq_i,Tq_m)\cdot[1-\beta(Tq_m)]} \quad (18)$$

In the equation, $d(Tq_i,Tq_m)$indicates the shortest distance between superpixels Tq_i and Tq_m. If the superpixel Tq_m is predicted as a target in the similarity graph, then assign the value of $\beta(\cdot)$ as 1, and N_1 is the total number of superpixels.

The more superpixel similarities that are predicted as targets, the lower the numerator value and the higher the denominator values, which makes the lower value of F. Set the reciprocal of Fas the target probabilityf_i, which means that the algorithm has identified the possible salient target in the UAV video reconnaissance image. To improve the detection accuracy of the salient target, following image enhancement algorithm is used to improve the detection accuracy of the salient target.

D. Enhancement of Salient Target of UAV Video Reconnaissance Image Based on HSI Color Space

First, convert the low illuminance image from RGB space to HSI space [20], and then different enhancement algorithms are used respectively according to component (S) and component(I). The process of this algorithm is: (1) Color space conversion, from RGB space to HSI space; (2) Enhance components$(S)$and(I)respectively by "Piecewise exponential transformation" and "V-transformation + Retinex enhancement + improved fuzzy enhancement"; (3) Return to RGB color space to get the enhanced image. The flow chart is shown in Fig. 3.

The enhancement algorithms for each stage are described in detail below.

1) *Enhancement of saturation component*: Generally, the enhancement of the saturation component is a simple linear transformation [21]. This paper proposes a piecewise exponential enhancement algorithm, characterized by its nonlinear nature. The saturation levels of distinct areas can be processed individually to enhance the overall visual impact. The salient target image of the UAV video reconnaissance image is divided into three regions based on its saturation level: high, medium, and low. When $x > 0$, $u^x - 1 > x$, and $\lim_{x\to 0}\frac{u^x-1}{x} = 1$, that is, when$x$is extremely small, $u^x - 1$ and$x$are equivalent. Therefore, the low saturation region is stretched by the exponential transformation to enlarge the saturation; for the medium saturation region, only exponential transformation is done to make appropriate adjustments; for the high saturation region, the saturation is reduced appropriately through the reduction of the exponential transformation. The piecewise exponential enhancement algorithm proposed in this paper is expressed by Eq. (19):

$$P'(m,n) = \begin{cases} \eta\left[u^{P(m,n)} - 1\right], & P(m,n) \le 0.2 \\ u^{P(m,n)} - 1, & 0.2 < P(m,n) \le 0.7 \\ \mu\left[u^{P(m,n)} - 1\right], & else \end{cases} \quad (19)$$

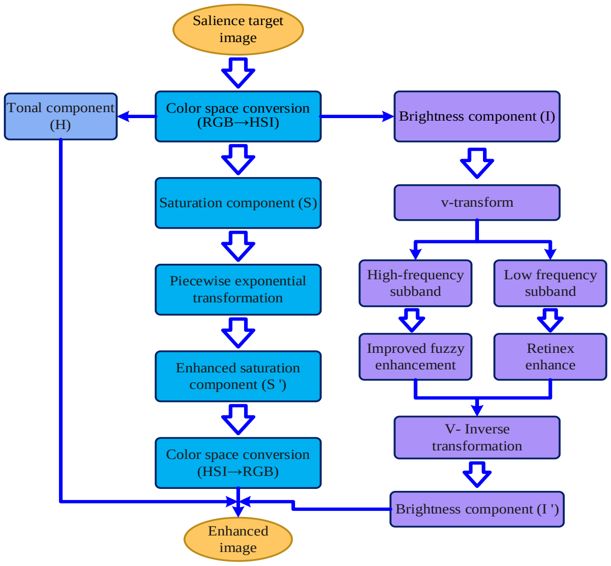

Salience target image
Tonal component (H)
Color space conversion (RGB→HSI)
Brightness component (I)
Saturation component (S)
v-transform
High-frequency subband
Low frequency subband
Piecewise exponential transformation
Improved fuzzy enhancement
Retinex enhance
Enhanced saturation component (S ')
V- Inverse transformation
Color space conversion (HSI→RGB)
Brightness component (I ')
Enhanced image

Fig. 3. Flow chart of image enhancement algorithm.

In the equation, $P$and$P'$are the saturation before and after enhancement, respectively; the parameters η, $\mu$are used to adjust the scaling of the transformation. When$\eta$is 1.2~1.5, and $\mu$is 0.7~0.9, the enhancement effect is the best. In this paper, the value of η is 1.4, and the value of μ is 0.8.

2) *Enhancement of brightness component*: V-transform the brightness component to obtain the V-spectrum matrix. The data of the 1/4 of the upper left corner of the V-spectrum matrix is the low-frequency sub-band $I_L$of the brightness component, with the rest of the data being the high-frequency sub-bandI_H. Different methods of enhancement are used for low and high frequencies.

a) *Low-frequency coefficient enhancement (Retinex):* Because the low-frequency sub-band $\mathrm{I_L}$concentrates the overall information of the brightness of the salient target image of the UAV video reconnaissance image, it depicts the general outline of the image, and the Retinex enhancement algorithm can be used.

Assuming that the brightness low-frequency coefficient is$P(x,y)$, the enhanced low-frequency coefficient is calculated by Eq. (20):

$$\begin{aligned} z(x,y) &= ln\,J\,(x,y) \\ &= ln\,P\,(x,y) - ln[D(x,y) * P(x,y)] \end{aligned} \quad (20)$$

$z(x,y)$ stands for the output image after enhancement and $*$ stands for the convolution symbol; $D(x,y)$ is the center-surround function, a Gaussian filter function is usually chosen as shown in Eq. (21):

$$D(x,y) = \sigma \cdot u^{\frac{-(x^2+y^2)}{c^2}} \tag{21}$$

In the equation, c is the Gaussian surround scale; σ is a constant which makes the integral of $D(x,y)$ as 1. The reflected image$J(x,y)$ represents the intrinsic property of an image and carries detailed information about the image.

The key to Retinex theory is to reasonably assume the composition of the salient target image of the UAV video reconnaissance image [22]. If an image is regarded as an image with noise, then the component of the incident light can be regarded as a multiplicative, relatively uniform, and slowly transformed noise [23]. The Retinex algorithm can fairly estimate the noise present at every position within the image and remove it to acquire a noticeable Impact. The obtained image reduces the impact of incident light [24], and retains the reflection attribute of the object essence, that is, the essence of the image.

b) High-frequency coefficient enhancement (Improved fuzzy optimization): In the sub-bandI_H characterized by high frequency, the wavelet coefficient of noise exhibits a diminutive magnitude, which is further reduced compared to the coefficient of the signal. To enhance the details of the salient target image and suppress noise [25], this paper proposes an improved fuzzy enhancement algorithm, which aims to enhance the clarity of details while simultaneously suppressing noise [26]. The specific algorithm process is outlined below:

Step 1: Construct the affiliation function as shown in Eq. (22):

$$E_{mn} = \frac{x_{mn} - x_{min}}{xmin_{max}} \tag{22}$$

Transforming the high-frequency coefficient into fuzzy sets is aiming to normalize the coefficient in the high-frequency subband to the interval$[0,1]$. In the equation, x_{max}andx_{min}stand respectively for the highest and lowest values of the coefficient in the high-frequency subband; x_{mn} is the coefficient.

Step 1: Design the fuzzy affiliation transformation as expressed by Eq. (23):

$$E'_{mn} = \frac{1}{2} + \left(E_{mn} - \frac{1}{2}\right)^{1/3} \tag{23}$$

This function is a nonlinear and monotonically increasing function and $\left(\frac{1}{2}, \frac{1}{2}\right)$ is its inflection point. It makes the numbers less than $\frac{1}{2}$ shrinking, and makes the numbers greater than$\frac{1}{2}$amplified. When acting on the high-frequency coefficient, it can achieve the purpose of enhancing the details of the salient target image of the UAV video reconnaissance image while suppressing noise.

Step 1: Convert the fuzzy set to the high-frequency subband as shown in Eq. (24):

$$E''_{mn} = E'_{mn} \cdot (xmin_{max} + x_{min}) \tag{24}$$

The enhanced high-frequency coefficients are obtained.

So far, the enhanced low-frequency and high-frequency coefficients are obtained, and then the enhanced brightness component I' is obtained through V-inverse transformation. Finally, the obtained enhanced saturation$S'$and the enhanced brightness I', and the hue component of the salient target image$(H)$are synthesized and converted to RGB space to output the enhanced salient target image.

III. Experimental Analysis

To verify the effect of the method in this paper on the detection of salient targets of UAV video reconnaissance images, the GPU configuration of experimental hardware equipment is selected as displayed in Table I.

The UAV used for reconnaissance shooting is shown in Fig. 4. The configuration parameters are shown in Table II.

TABLE I. GPU Configuration

Name	Parameter
Brand	Shadow Chi
Model	GTXTITAN-6GD5
Craftsmanship	28mm+
Stream processor	2688pcs
Core frequency	954 MHze
Video memory capacity	6G GDDR5+
Video memory bit width	384Bite
Video memory frequency	6008MHz
Graphics card power consumption	300W+
Heat-dissipating method	Cooling fan
Size	280mm×127mm×42mm

TABLE II. UAV Configuration Table

Name	Parameter
Model	RQ-8 drone
Maximum safe takeoff weight	7kg
No-load	6.5kg
Load	2kg
Size	The length is 970mm
	Width 920mm, 500mm after folding
	Height 240mm, folded back 180mm
Wheelbase size	1100mm
Propeller size	26inch
Duration of flight	>50min
Maximum speed	36km/h
Wind loading rating	Strong breeze
Operating radius	>10km
Dynamic type	Whoring polymer cell
Working altitude	4500m
Operating temperature	-20 to 60 degrees Celsius

Fig. 4. UAV used for reconnaissance shooting.

The experimental test dataset consists of images captured by a drone as shown in Fig. 4. This dataset consists of 500 frames of video reconnaissance images captured by drones in different scenes, lighting conditions, and dynamic backgrounds. These images contain complex and subtle information, such as hidden targets, varied backgrounds, and challenging factors such as lighting changes. The dataset is divided into two parts: (1) Training set: containing 300 frames of images, used to train a CNN based saliency object detection model. These images cover various possible scenarios and conditions in the dataset to ensure that the model can learn enough features to cope with complex detection tasks. (2) Test set: Contains the remaining 200 frames of images to evaluate the performance of the trained model on unknown data. These images maintain similar diversity to the training set, but are completely independent to ensure the objectivity and accuracy of the test results.

After completing the above preparations, proceed with the experiment according to the following steps:

Step 1: Dataset preparation and annotation

Collect and organize drone video surveillance images, accurately label salient targets in the images through manual means, and provide accurate data foundation for model training.

Step 2: Remove haze from the image

To remove haze from the original image, improve image quality, reduce the impact of haze on subsequent object detection, and enhance detection accuracy.

Step 3: Faster R-CNN model training

Using annotated training set data, train the Faster R-CNN model to learn the features of salient targets and possess the ability to classify and perform bounding box regression.

Step 4: Test Set Evaluation

Evaluate the trained Faster R-CNN model using an independent test set to validate its detection performance on unknown data, including subjective and objective testing metrics such as accuracy, recall, and F1 score.

Step 5: Extreme situation handling and image enhancement

In response to extreme situations where the Faster R-CNN model cannot accurately detect, methods such as superpixel saliency screening are used for supplementary detection, and HSI color space enhancement is applied to the detected saliency target images to improve image clarity and visual effects.

Step 6: Result analysis and optimization

Conduct a comprehensive analysis of the experimental results, evaluate the advantages and disadvantages of the model, and further optimize the model based on the analysis results to improve detection performance and robustness.

In the test set, 1 frame image is selected to perform the target detection experiment using the trained network, and the detection results are plotted as shown in Fig. 5. And compare these detection results with the detection results of the chromatography-mass spectrometry method in study [4] and the joint transform correlator method in study [5], which are shown in Fig. 6 and Fig. 7. The red boxes are the correct targets, the yellow circles are the false detection, and the blue boxes are the missed detection.

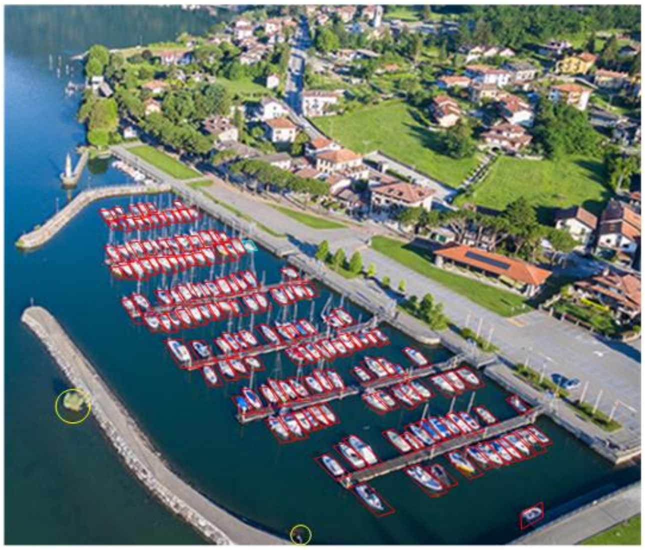


Fig. 5. Detection results of the proposed method.

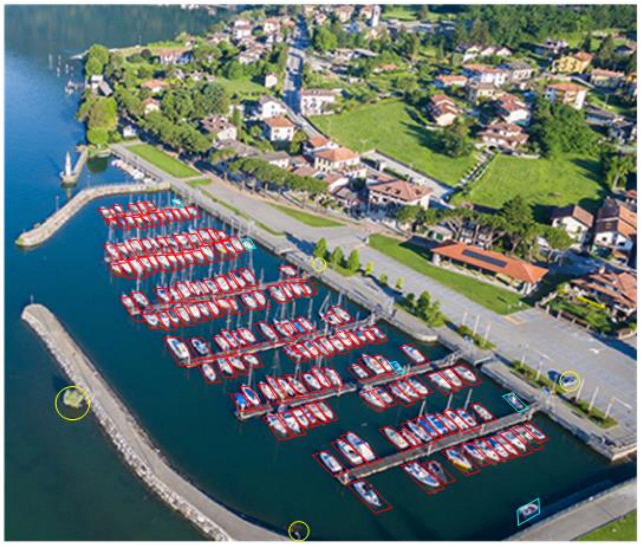


Fig. 6. Test results by chromatography-mass spectrometry.

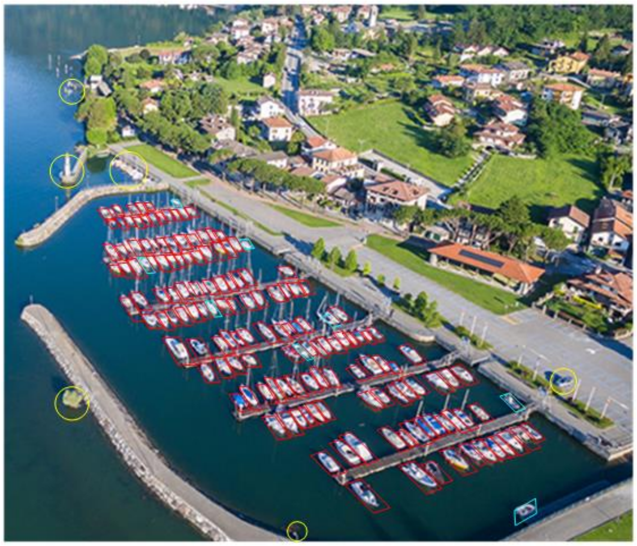


Fig. 7. Detection results of the joint transform correlator method.

From Fig. 5, it can be seen clearly that the method in this study successfully detects all targets, except one missing detection and two false detections. This result shows that the method in this study has high accuracy and reliability in the detection of salient targets of UAV video reconnaissance images. Through the analysis of the results, it can be seen that the method in this paper provides an effective way for the accurate detection of salient targets, with high feasibility and significant effectiveness. In contrast, there are five missed and four false detections in the detection results of the chromatography-mass spectrometry method shown in Fig. 6. This data is significantly higher than the data of the method in this paper, indicating that the effect of the method of chromatography-mass spectrometry in target detection is not ideal. The detection results of the joint transform correlator method in Fig. 7 are even more disappointing, with 8 missed and 6 false detections. This result shows the shortcomings of the method of joint transform correlator in multi-target detection. In conclusion, by comparing the target detection effect of the three methods, it can be seen that the method in this paper has high accuracy and low false detection rate in multi-target detection, and can play an important role in UAV video reconnaissance.

To verify the performance of the salient target detection method in this paper, 1 frame of image is selected in the test set for the experiment. Also compare the result with the experimental result of the chromatography-mass spectrometry method and the joint transform correlator method, as shown in Fig. 8 to Fig. 11.

Fig. 8 shows a scene with multiple targets and a complex background, increasing the difficulty of target detection. By comparing the detection results in Fig. 9, Fig. 10, Fig. 11, and Fig. 8, the advantages of this method in dealing with such complex scenes are evident. In the detection result of this method, salient targets are detected, and there is no false detection or missed detection. On the contrary, both the chromatography-mass spectrometry method and the joint transform correlator method have problems in processing Fig. 8. There are three missed detections in the detection of salient targets by the chromatography-mass spectrometry method, and the detected targets are not clear, which has the problem of shadow occlusion. The detection effect of the joint transform correlator method is worse. There are not only five missed detections but also the green belt in the middle of the road has been detected. In addition, the detected target is accompanied by obvious shadows, which makes the target unclear. By conducting a comparative analysis, it is concluded that the method in this study can detect salient targets more accurately and clearly, which proves the effectiveness of this method in complex scenes.

The frame images in the test set are selected for the image enhancement experiment. Compared with the joint transform correlator method and the chromatography-mass spectrometry method, the original plots and the experimental results of the three methods are shown in Fig. 12 to Fig. 15.

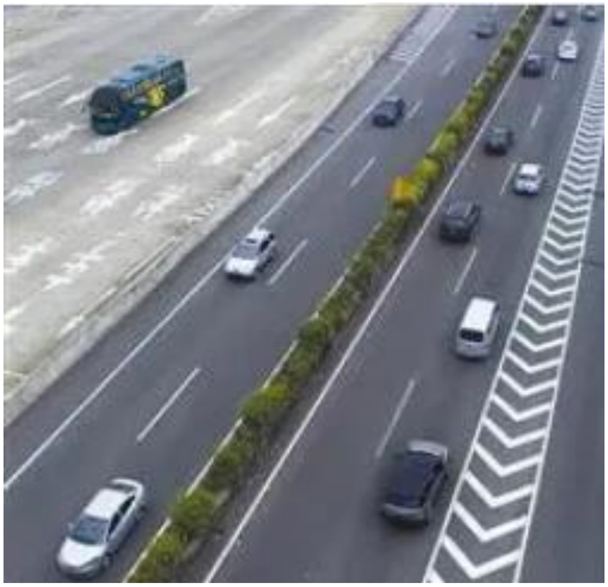


Fig. 8. Original drawing.

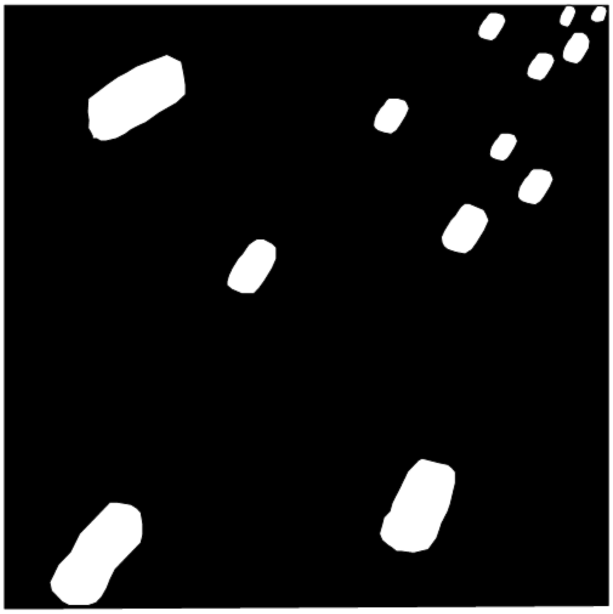


Fig. 9. Result of salient target detection of the proposed method.

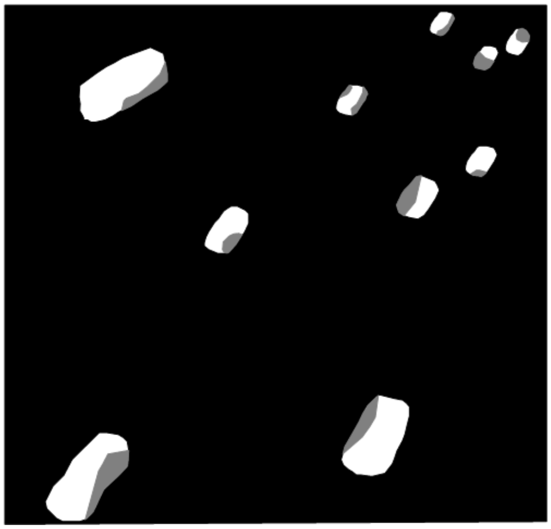


Fig. 10. Result of salient target detection by chromatography-mass spectrometry.

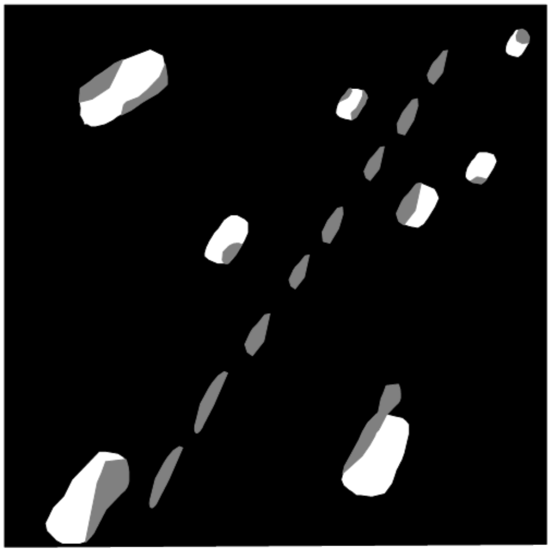


Fig. 11. Result of the salient target detection by joint transform correlator.

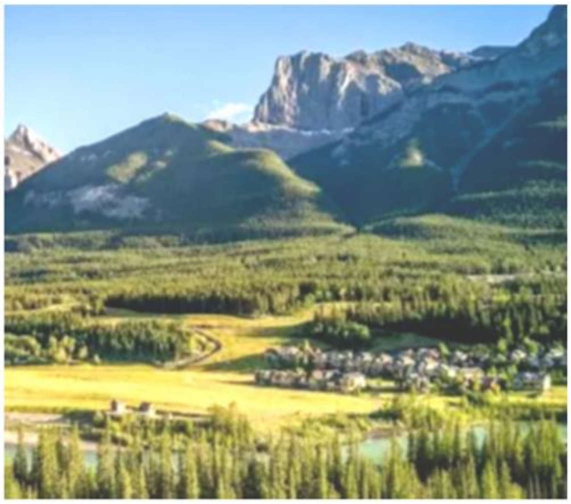


Fig. 12. Original drawing.

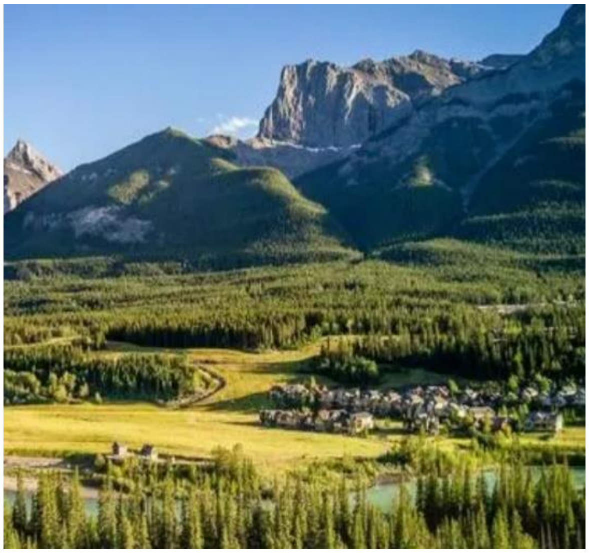


Fig. 13. Enhancement effect of the proposed method.

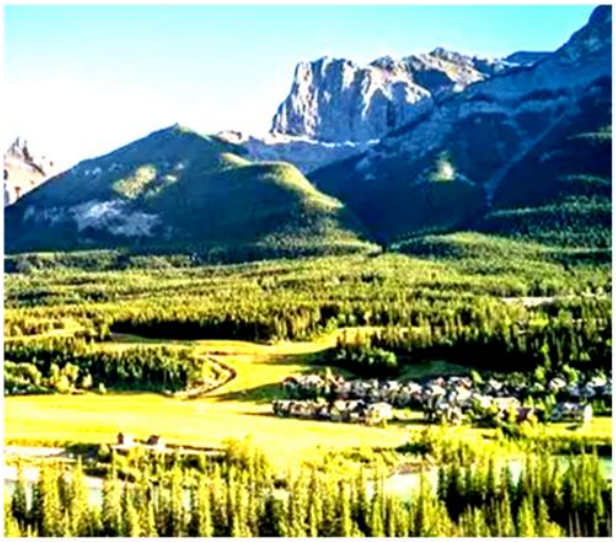


Fig. 14. Enhancement effect of joint transform correlator.

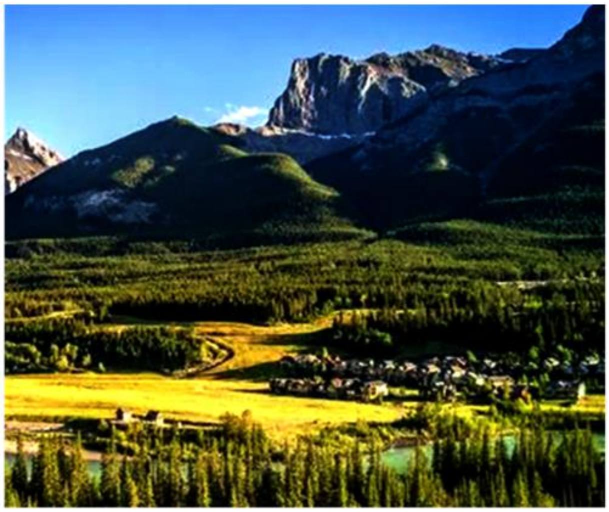


Fig. 15. Enhancement effect of chromatography-mass spectrometry method.

By comparing Fig. 13 and Fig. 12, the advantages of this method in image enhancement are apparent. The original image in Fig. 12 is slightly fuzzy and the details are not clear enough, but after the enhancement processing of this method, the image in Fig. 13 becomes very clear and the details are presented vividly. More importantly, the color of the enhanced image is soft, which is closer to the visual effect of the real scene, providing a good basis for subsequent tasks such as target detection. In contrast, although the joint transform correlator method in Fig. 14 has some effect of enhancement, the color is too bright, even some dazzling, giving a sense of unnatural. This kind of too-bright color will cover up some important details, causing trouble for subsequent tasks. The chromatography-mass spectrometry method in Fig. 15 has insufficient effect of enhancement. The overall image is dark and some details are not clear enough. Such an enhancement effect will make subsequent tasks such as target detection more difficult. In conclusion, the method in this paper performs well in the enhancement of UAV video reconnaissance images. It can not only significantly improve the image clarity, but also maintain the authenticity and softness of colors. The enhancement effect has a positive role in promoting the detection of salient targets.

To further validate the object detection performance of the proposed method, the detection performance of the three methods was compared using accuracy, recall, F1 score, and average detection time as objective indicators. The results are shown in Table III.

TABLE III. UAV CONFIGURATION TABLE

Method	Precision	Recall	F1 Score	Average detection time (seconds)
Chromatography-mass Spectrometry Method	0.75	0.68	0.71	5.2
Joint Transform Correlator	0.80	0.72	0.76	4.8
Proposed Method	0.90	0.85	0.87	0.3

According to Table III, the chromatography-mass spectrometry method shows relatively low accuracy and recall, with values of 0.75 and 0.68, respectively. This indicates that the method has certain errors in distinguishing significant and non-significant targets, and may miss some targets. The joint transformation correlator method has improved in accuracy and recall, reaching 0.80 and 0.72 respectively, demonstrating better object detection capability. The method proposed in this paper performs well in both accuracy and recall, reaching 0.90 and 0.85 respectively, significantly higher than the other two methods, indicating that this method can more accurately identify and detect significant targets. The F1 score of proposed method reached 0.87, which is much higher than that of the chromatography-mass spectrometry method (0.71) and the combined transform correlation method (0.76), further verifying the superiority of the proposed method. In terms of detection speed, the method proposed in this paper demonstrates significant advantages, with an average detection time of only 0.3 seconds, far lower than the chromatography-mass spectrometry method (5.2 seconds) and the combined transform correlation method (4.8 seconds). This indicates that the method proposed in this paper has higher real-time performance in processing drone video reconnaissance images. In summary, the CNN based method for detecting salient objects in unmanned aerial vehicle (UAV) video reconnaissance images proposed in this paper outperforms the compared methods in terms of accuracy, recall, F1 score, and detection speed. The experimental results show that the method proposed in this paper can accurately and efficiently detect salient targets in drone video reconnaissance images, which is of great significance for improving the efficiency and accuracy of reconnaissance work.

IV. CONCLUSION

With the popularization of UAV technology, the application of UAVs in the field of video reconnaissance is more and more extensive. However, the images taken by UAVs are often affected by factors such as illumination variation, camera angles, etc., which brings certain difficulties to target detection. To solve the problem, this study proposes a CNN-based salient target detection method for UAV video reconnaissance images. According to Faster R-CNN, the salient target detection of UAV video reconnaissance images is realized. For the salient target that cannot be detected in extreme cases, the detection is completed using the salient target calculation method of this paper. To make the detection effect better, the salient target enhancement algorithm is set to complete the salient target detection of UAV video reconnaissance image. Through experimental verification, the method of this paper has high accuracy and low false detection rate, can detect salient targets more accurately and clearly, and performs well in the enhancement of UAV video reconnaissance images. However, the effectiveness of this method largely depends on the quality and diversity of the training dataset. If the training dataset cannot fully cover various complex scenarios that may be encountered in practical applications, the generalization ability of the model may be limited. In addition, image differences under different drone platforms and shooting conditions may also affect the detection performance of the model. Therefore, in the future, addressing this issue will be the focus.

COMPETING OF INTERESTS

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Li Na: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

DATA AVAILABILITY

On Request

DECLARATIONS

Not applicable

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

AUTHORS' STATEMENT

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

FUNDING

Not applicable

ETHICAL APPROVAL

All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

COMPETING OF INTERESTS

The authors declare no competing of interests.

REFERENCES

[1] Y. Hu, X. Wang, W. Li, X. Hei, and G. Xie, “A New Ship Target Detecting Method Based on Saliency in SAR Image,” in 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC), IEEE, 2021, pp. 241–245.

[2] V. R. S. Mani, A. Saravanaselvan, and N. Arumugam, “Performance comparison of CNN, QNN and BNN deep neural networks for real-time object detection using ZYNQ FPGA node,” Microelectronics J, vol. 119, p. 105319, 2022.

[3] K. Lee et al., “STEM image analysis based on deep learning: identification of vacancy defects and polymorphs of MoS2,” Nano Lett, vol. 22, no. 12, pp. 4677–4685, 2022.

[4] C. Jirayupat et al., “Image Processing and Machine Learning for Automated Identification of Chemo-/Biomarkers in Chromatography–Mass Spectrometry,” Anal Chem, vol. 93, no. 44, pp. 14708–14715, 2021.

[5] A. K. Cherri and A. S. Nazar, “Class-associative multiple target recognition for highly compressed color images in a joint transform correlator,” Optical Engineering, vol. 61, no. 12, p. 123102, 2022.

[6] I. García-Aguilar, R. M. Luque-Baena, and E. López-Rubio, “Improved detection of small objects in road network sequences using CNN and super resolution,” Expert Syst, vol. 39, no. 2, p. e12930, 2022.

[7] E. López-Rubio, M. A. Molina-Cabello, F. M. Castro, R. M. Luque-Baena, M. J. Marín-Jiménez, and N. Guil, “Anomalous object detection by active search with PTZ cameras,” Expert Syst Appl, vol. 181, p. 115150, 2021.

[8] M. Gazzea, M. Pacevicius, D. O. Dammann, A. Sapronova, T. M. Lunde, and R. Arghandeh, “Automated power lines vegetation monitoring using high-resolution satellite imagery,” IEEE Transactions on Power Delivery, vol. 37, no. 1, pp. 308–316, 2021.

[9] A. W. S. Putra, H. Kato, and T. Maruyama, “Infrared LED marker for target recognition in indoor and outdoor applications of optical wireless power transmission system,” Jpn J Appl Phys, vol. 59, no. SO, p. SOOD06, 2020.

[10] R. Akter, V.-S. Doan, T. Huynh-The, and D.-S. Kim, “RFDOA-Net: An efficient ConvNet for RF-based DOA estimation in UAV surveillance systems,” IEEE Trans Veh Technol, vol. 70, no. 11, pp. 12209–12214, 2021.

[11] D. Mishra, S. K. Singh, R. K. Singh, and D. Kedia, “Multi-scale network (MsSG-CNN) for joint image and saliency map learning-based compression,” Neurocomputing, vol. 460, pp. 95–105, 2021.

[12] K. Ogohara and R. Gichu, “Automated segmentation of textured dust storms on mars remote sensing images using an encoder-decoder type convolutional neural network,” Comput Geosci, vol. 160, p. 105043, 2022.

[13] Z.-H. Lin, A. Y. Chen, and S.-H. Hsieh, “Temporal image analytics for abnormal construction activity identification,” Autom Constr, vol. 124, p. 103572, 2021.

[14] S. Molavi Vardanjani, A. Fathi, and K. Moradkhani, “Grsnet: gated residual supervision network for pixel-wise building segmentation in remote sensing imagery,” Int J Remote Sens, vol. 43, no. 13, pp. 4872–4887, 2022.

[15] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, “Scalable recurrent neural network for hyperspectral image classification,” J Supercomput, vol. 76, no. 11, pp. 8866–8882, 2020.

[16] A. Ferdowsi, M. A. Abd-Elmagid, W. Saad, and H. S. Dhillon, “Neural combinatorial deep reinforcement learning for age-optimal joint trajectory and scheduling design in UAV-assisted networks,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 5, pp. 1250–1265, 2021.

[17] P. D. Ledger, B. A. Wilson, A. A. S. Amad, and W. R. B. Lionheart, “Identification of Metallic Objects using Spectral MPT Signatures: Object Characterisation and Invariants,” arXiv preprint arXiv:2012.10376, 2020.

[18] S. El Mohtar, B. Ait-El-Fquih, O. Knio, I. Lakkis, and I. Hoteit, “Bayesian identification of oil spill source parameters from image contours,” Mar Pollut Bull, vol. 169, p. 112514, 2021.

[19] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, and X. Chen, “Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+,” Comput Geosci, vol. 158, p. 104969, 2022.

[20] A. Patra, A. Saha, and K. Bhattacharya, “High-resolution image multiplexing using amplitude grating for remote sensing applications,” Optical Engineering, vol. 60, no. 7, p. 73104, 2021.

[21] S. V. S. Diddi and L.-W. Ko, “Course-grained multi-scale EMD based fuzzy entropy for multi-target classification during simultaneous SSVEP-RSVP hybrid BCI paradigm,” International Journal of Fuzzy Systems, vol. 24, no. 5, pp. 2157–2173, 2022.

[22] R. Theagarajan et al., “Integrating deep learning-based data driven and model-based approaches for inverse synthetic aperture radar target recognition,” Optical Engineering, vol. 59, no. 5, p. 51407, 2020.

[23] X. Chen, R. Proietti, C.-Y. Liu, and S. J. Ben Yoo, “A multi-task-learning-based transfer deep reinforcement learning design for autonomic optical networks,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 9, pp. 2878–2889, 2021.

[24] M. Goudarzi, M. Palaniswami, and R. Buyya, “A distributed deep reinforcement learning technique for application placement in edge and fog computing environments,” IEEE Trans Mob Comput, vol. 22, no. 5, pp. 2491–2505, 2021.

[25] S. Gupta, P. K. Rai, A. Kumar, P. K. Yalavarthy, and L. R. Cenkeramaddi, “Target classification by mmWave FMCW radars using machine learning on range-angle images,” IEEE Sens J, vol. 21, no. 18, pp. 19993–20001, 2021.

[26] A. Rizik, E. Tavanti, H. Chible, D. D. Caviglia, and A. Randazzo, “Cost-efficient FMCW radar for multi-target classification in security gate monitoring,” IEEE Sens J, vol. 21, no. 18, pp. 20447–20461, 2021.

Construction of Image Retrieval Module for Ethnic Art Design Products Based on DF-CNN

Yaru He
Chengdu Vocational and Technical College, Chengdu, 610041, China

***Abstract*—With the increasing interest of consumers in ethnic art, more design products with ethnic art characteristics are being displayed. In order to help users easily retrieve related art products, an image retrieval model that can effectively extract data is proposed. The research method strengthens the depth of data mining through weighted methods, main characteristics and local features in images based on the multi-window combination, and uses the deep forest algorithm to expand the decision path and select information gain nodes. By adjusting the weights of convolutional neural networks, the retrieval ability of the model is enhanced. The gradient problem in the propagation process is optimized using residual modules, and the prominent features of the features are strengthened using a bar attention mechanism to optimize the retrieval ability. The results indicated that the loss function of the research model converged within 20 iterations, and the matching degree of the retrieved images in the testing set reached 91.28% after iterative training. The AUC of the research model was 0.876, indicating that the model had a good performance in image retrieval and classification. The retrieval accuracy of the research model was higher than other methods for image data of different specifications. This indicates that the research model has universality for multi-scale image retrieval, which can provide theoretical support for the development of ethnic art design products.**

***Keywords*—Image retrieval; main characteristics; local features; deep forest; convolutional neural network; bar attention mechanism; residual module**

I. Introduction

Ethnic art contains the cultural heritage and artistic ideas of a nation, and products created based on ethnic art have distinct cultural characteristics [1]. More art museums are using ethnic art design as a featured theme to share related types of exhibits. However, the patterns of ethnic characteristic exhibits are usually formed by simple line and pattern combinations, which do not have obvious characteristics of things, making retrieval difficult [2-3]. In order to retrieve images more effectively, various image retrieval method shave been developed. Early image retrieval relied on textual keywords to express image features, using the main parameters of the image as a reference standard, and obtaining retrieval results through data comparison and visual combination. However, method comparison requires manual labeling, resulting in low retrieval efficiency [4]. With the advancement of retrieval technology, content-based image retrieval methods have been proposed. The image search method can avoid manual errors by comparing image attributes with image features to obtain retrieval results. However, the recognition effect on local features is unstable and easily affected by noise [5]. Therefore, D. Lowe proposed the Scale-Invariant Feature Transform (SIFT) feature method and the Local Binary Pattern (LBP) feature method. The local features of the image are enhanced to deepen the model's memory of feature points. Moreover, it has good recognizability for behaviors such as image rotation and scaling, but the method places too much emphasis on local features, resulting in a lack of global perspective [6]. With the development of intelligent technology, Convolutional Neural Networks (CNN) have been applied in image retrieval. The self-learning ability based on CNN can accurately extract image features and enhance the adaptability of network retrieval through dataset training. However, excessive feature vectors may lead to long training time and slow retrieval efficiency of the network. The research aims to design a retrieval module that ensures the accuracy of image retrieval while improving the speed of image retrieval. To optimize the user's search experience and match the search needs of art galleries or museums. Therefore, the DF-CNN algorithm is innovatively proposed for image retrieval in the field of ethnic art products. It strengthens local features through a bar attention mechanism, combines local features with subject features using an overlap pooling method, and optimizes the retrieval speed of the model through parallel multi-channel convolution. A new ethnic art design product retrieval model is constructed to provide certain technical support for the dissemination of ethnic art.

II. Related Work

With the development of Internet technology, the number of image data in the network is gradually rich, but it is difficult to accurately retrieve the required image in a rich database. Therefore, researchers have conducted research on image retrieval methods. Suganyadevi S et al. proposed a research method based on deep learning for medical image retrieval. The method established a learning system by exploring data information in the medical field, and reduced the dimensionality of functions using deep conventional and extreme learning methods to complete the retrieval and analysis of medical images. The results showed that the proposed method was accurate for retrieving medical image data [7]. Kelishadrokhi M K et al. proposed a research method based on novel local texture descriptors and color features for image retrieval problems in computers. The texture and color information database was used to train the image recognition ability. Effective features were extracted using the local neighborhood difference method. The results indicated that the proposed method had a good corresponding effect on computer content image retrieval [8]. Keisham Net al. proposed a research method based on depth search and rescue algorithm for image retrieval on the Internet. The method used advanced visual feature extraction techniques to enhance the feature points of the image, and optimized the feature clustering of the image through feature fusion and filtering. The results indicated that the proposed method significantly improved the efficiency of image retrieval [9]. Wang W et al. proposed a method based on sparse representation and feature fusion for image retrieval in databases. The method used a generalized search tree to retrieve similar scenes in the image and enhanced local features of the image through sparse coding. The results indicated that the proposed method improved the accuracy of database image retrieval [10]. Ning C et al. proposed a method based on deep metric learning for image retrieval of clothing. The method called similar features from the database through pre-set regional scenarios, and optimized the recognition of clothing patterns based on feature similarity comparison and feature point analysis. The results indicated that the proposed method had a fast retrieval speed for clothing patterns [11].

Zhuang H et al. adopted the Deep Forest (DF) algorithm to address the urban land change. A dimensional space was constructed based on terrain modeling. Advanced features in structural data were mined on the basis of deep learning methods, and the changing state of land was simulated through community comparison. The designed method had a relatively accurate prediction level for land change issues [12]. Hamedianfar et al. proposed a method based on the DF algorithm to address the application issues of remote sensing methods. Multi-scale features were established through shallow machine learning, the network architecture was expanded through a time series strategy and remote sensing method was optimized through multiple training methods. The developed method had a good simulation effect on remote sensing data modeling [13]. Shaaban M A et al. adopted a deep convolutional forest to detect text spam emails. Machine learning was used to perform the basic classification of email types. The dynamic deep integration method automatically adjusted the classification complexity and extracted effective features with the help of classifiers. The results indicated that the proposed method accurately isolated phishing emails [14]. Huang et al. proposed a method based on CNN for fault diagnosis in complex systems. The method used sample feature extraction from multivariate time series for multi-layer transmission, optimized data retrieval through sliding processing of data windows, and combined model training to enhance the diagnostic performance. The results indicated that the proposed method had high prediction accuracy for fault diagnosis [15]. Tayal A proposed a research method based on CNN for the diagnosis of retinal diseases. The method automatically identified disease types through the constructed intelligent learning framework, determined disease categories based on the feature set of medical images, and filtered out interference items through image denoising. The results indicated that the proposed method had an assisting role in the diagnosis of retinal diseases [16].

In summary, Kelishadrokhi et al. enhanced the depth of image feature extraction through local texture and color feature recognition, but it affected the retrieval speed of the image. Ning C et al. used deep metric learning to enhance the speed of retrieval, but due to the single feature extraction method, the matching degree of image retrieval decreased. Moreover, a single CNN algorithm lacks feature type differentiation, which can affect the efficiency of ethnic art image retrieval. However, there is currently limited data on the combined application of CNN and DF algorithms in the field. Therefore, research attempts to achieve synchronous improvement in retrieval performance and retrieval speed through the fusion of the two algorithms.

III. Methods and Materials

A. *Design of Image Retrieval Module for Ethnic Art Products Based on DF*

With the exchange and dissemination of culture, ethnic art patterns have gradually emerged as an artwork in the public eye. The application of ethnic patterns has developed from the initial daily life to ethnic art design [17-19]. Art design products with distinctive features during tourism are often given as souvenirs to friends and family. However, faced with a wide variety of ethnic art patterns, the search process is often dazzling [20-22]. In order to efficiently and accurately retrieve ethnic artworks with diverse patterns, the study introduces the DF for the image retrieval process. The DF algorithm, as an efficient classifier, can accurately extract image features from high-dimensional image data and optimize the model's fitting through its self-learning ability [23-25]. The cascading structure processing process of the DF algorithm is shown in Fig. 1.

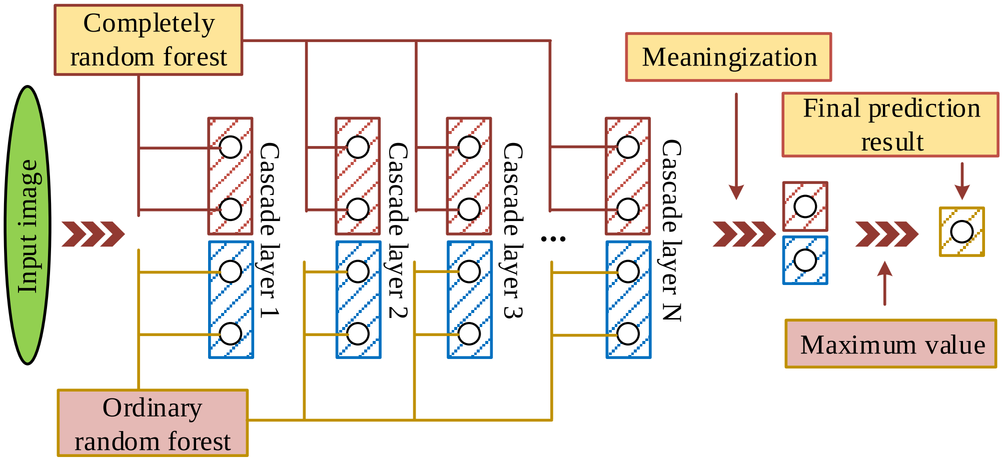


Fig. 1. The cascade structure of deep forests.

In Fig. 1, the DF algorithm processes input data through multiple layers. The process strengthens the data mining depth using the weighted method, and assigns hierarchical weights to the forest based on weight factors. Based on the predicted probability value of the forest, the weight factor's proportion is optimized. Through hierarchical multi-source data synchronization analysis, discrete image features can be effectively detected. To optimize the feature partitioning performance in the DF, the decision tree in the DF algorithm is used for information entropy optimization. The optimization process defines the information entropy of the sample set, as shown in Eq. (1).

$$Ent(N) = -\sum_{i}^{n} \frac{C^i}{N} \log_2 \frac{C^i}{N} \quad (1)$$

In Eq. (1), $Ent(N)$ represents the sample set information entropy. N represents the total sum of the sample set. C^i represents the sample size. The sample ratio is set to $p_i = \frac{C^i}{N}$ during the optimization process. Therefore, the information entropy of the sample set is converted, as shown in Eq. (2).

$$Ent(N) = -\sum_{i}^{n} p_i \log_2 p_i \quad (2)$$

In Eq. (2), p_i represents the sample ratio. Information entropy, as a metric in decision trees, can represent the set purity of a sample set. Based on the difference in information entropy before and after the feature decision dataset, the change in information gain is calculated. The value of information gain is measured by the influence proportion of branch nodes. The gain calculation process is shown in Eq. (3).

$$Gain = (N, a) = Ent(D) - \sum_{v=1}^{V} \frac{|N^v|}{|N|} Ent(N^v) \quad (3)$$

In Eq. (3), $Gain$ represents the information gain of dataset N after being partitioned by feature attribute a. a represents the selected feature attribute. V represents the number of branch nodes that may be formed during the process. N^v represents the sample data when the branch node is V. v represents the range of values for feature attribute a. $|N^v|/|N|$ represents the weight assigned to branch nodes. The feature partitioning of the input image by the decision tree is evaluated by the Gini coefficient, and the evaluation calculation process is expressed as Eq. (4).

$$Gain(D) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k}^{K} p_k^2 \quad (4)$$

In Eq. (4), $Gain(D)$ represents the Gini index. D represents the selected dataset. D represents the probability of decision-making. K signifies the number of branch nodes in the current dataset D. k signifies the value of the current branch node. To extract features from the entire image layer, dynamic sliding windows are used to scan image information in different regions. The complete feature extraction of the image is completed by concatenating the information from overlapping windows. The research process adopts multi-granularity scanning. The principle of the scanning model is shown in Fig. 2.

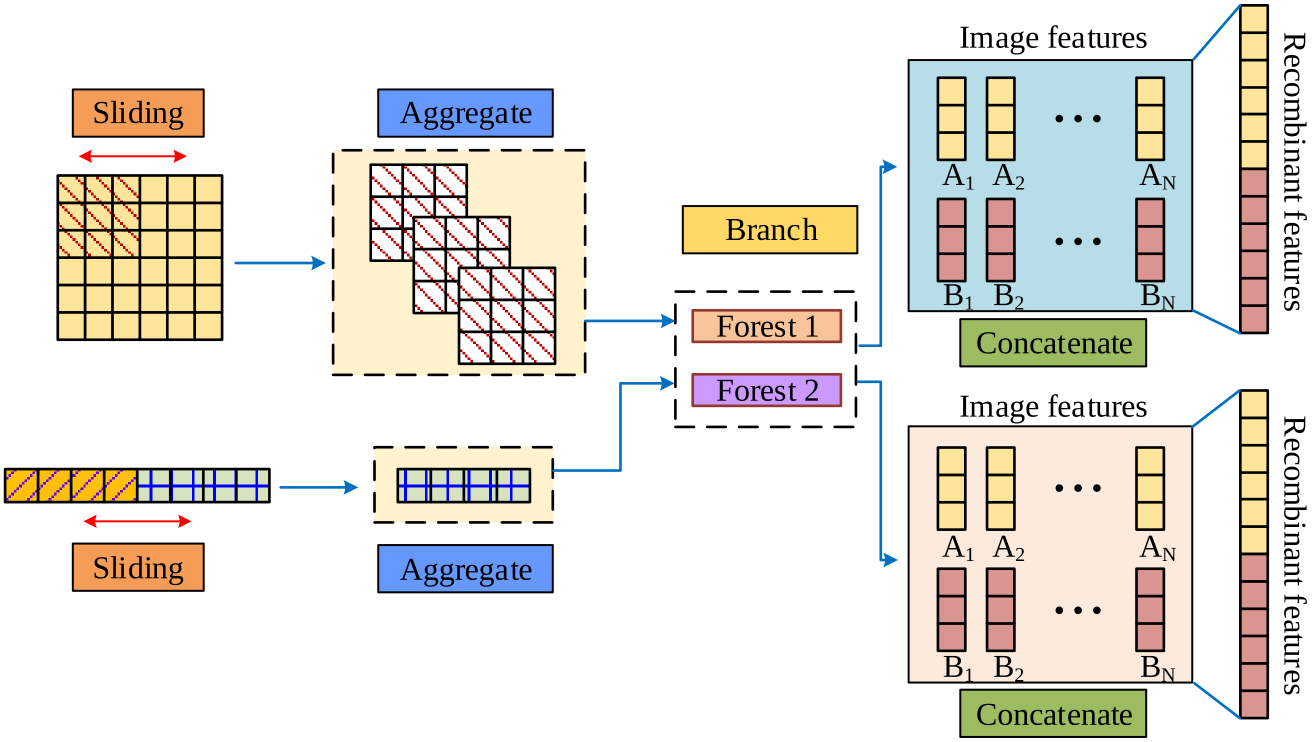


Fig. 2. Multi-granularity scanning process.

In Fig. 2, the multi-granularity scanning process of the DF algorithm uses window frames of different sizes as feature extraction windows. The main and local features in the image are identified through a combination of multiple windows. The feature output extracted by sliding window is used as a probability vector, and the main output and local output are hierarchically placed into a cascaded forest. The image dimension is enhanced through multi-vector scanning and transmission. In order to enhance local data feature processing, both completely random forest and ordinary random forest are used as decision paths, and split nodes with information gain are selected. To address the feature loss caused by dimensional differences, a linear discriminant analysis method is introduced to perform correlation recognition of image features. The mean vector is expressed as Eq. (5) during the calculation process.

$$u_j = \frac{1}{N_j} \sum_{x \in X_j} x\left(j = 0,1 \right) \tag{5}$$

In Eq. (5), u_j represents the mean vector of the classes j in the sample. N_j signifies the number of classes j in the sample. X_i signifies the dimension vector value of the sample. X_j signifies the class sample set of group j. The covariance matrix of sample features is calculated, as shown in Eq. (6).

$$\sum j = \sum_{x \in X_j} (x-u_j)(x-u_j)^T (j=0,1) \tag{6}$$

In Eq. (6), $\sum j$ represents the covariance matrix of the sample. T represents the transpose of a matrix. In order to calculate the projection point positions of two sets of samples, the divergence matrix is calculated in such a way that similar samples are close and dissimilar samples are far away. The optimization objective is rewrite through high-dimensional to low dimensional vector mapping. The maximum eigen value of the matrix is calculated by the generalized Rayleigh quotient. The DF algorithm generates its own feature vectors at each level when transmitting image features step by step. However, the algorithm's extraction step for the image is based on direct learning of the overall features. Too many levels in the cascaded forest may cause the enhanced features to be covered by ordinary feature vectors, resulting in weak fitting of the model to image samples even after learning. Due to the urgent need for retrieving images of ethnic art products in the scene, the study introduces the prior box method in the scanning stage to make the image extraction process faster. The specific working process of the prior box is shown in Fig. 3.

In Fig. 3, the prior box sets the window size that matches the extraction target through rough measurement of the input image. The window is directly generated based on the original image size using clustering algorithm to avoid the occupation of the main features by invalid information windows. The scanning process of the DF algorithm simplifies the traversal and sampling process of image frames, using effective feature data links to generate intersecting anchor boxes, and quickly locking the effective information window for feature transmission through algorithm training. Finally, the clustering pruning module and joint crossover method are used to achieve rapid output of effective information.

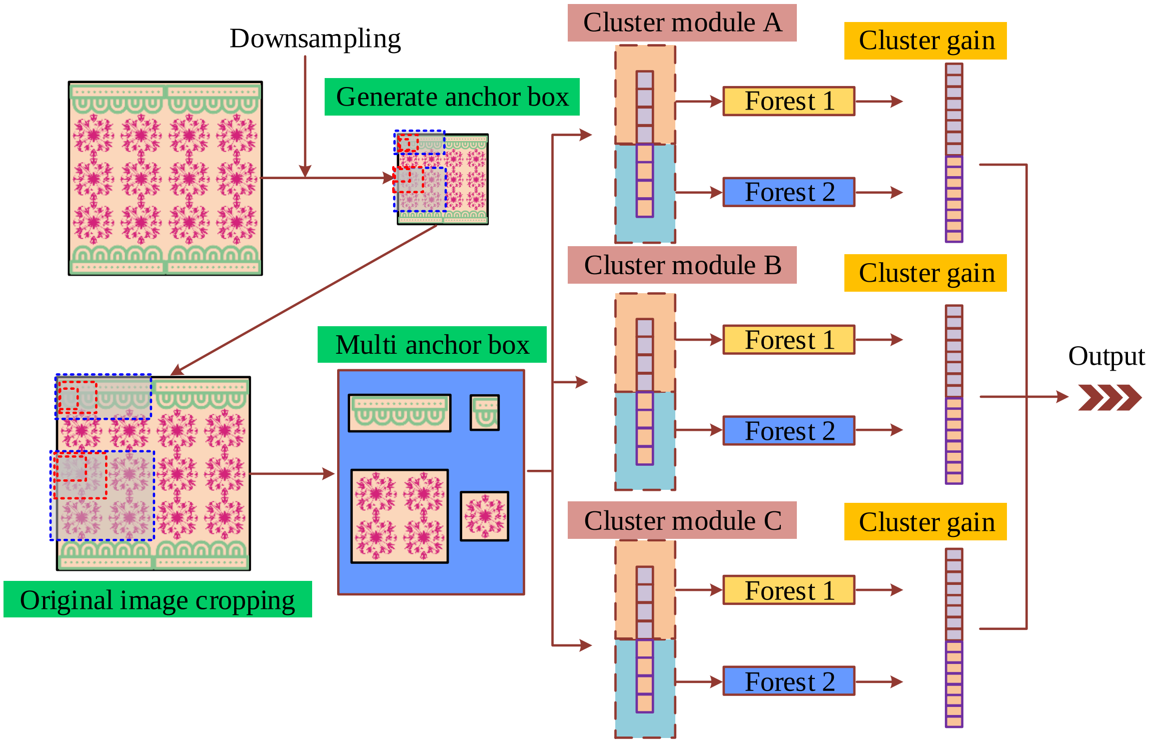


Fig. 3. Prior box workflow.

B. Construction of Ethnic Art Product Image Retrieval Module Combined with CNN

The image retrieval involves users as the main users, so the process inevitably involves personal subjective expression issues [26-27]. There is a semantic gap between the image information subjectively expressed by humans and the actual images recognized by computers. Therefore, the retrieval process needs to convert user expression information into image information and use the computer vision field to process the retrieval problem between images [28-29]. Image retrieval technology based on deep learning has precise and fast processing capabilities for complex and diverse pattern information. The core method of image retrieval technology is the adaptive ability of CNN [30]. Therefore, CNN is combined with DFto jointly construct an image retrieval module for ethnic art design products. The DF-CNN algorithm transmits feature images through sparse connections. The specification of the output feature map is represented by Eq. (7).

$$h_o = \frac{h_{im} + 2\times padding\_size - k\_h}{stride\_h} \tag{7}$$

In Eq. (7), h_o represents the size of the feature window. h_{im} represents the input image size. $padding\_size$ represents the pixel value of edge extension. k_h signifies the length of the convolution kernel. $stride\_h$ represents the step size. The CNN algorithm achieves image recognition through weight feedback and convolution calculation. The CNN is shown in Fig. 4.

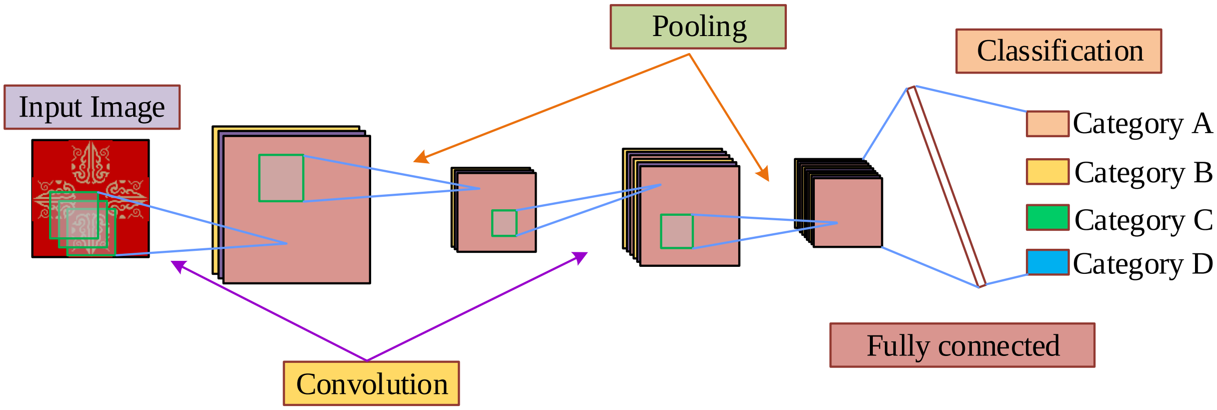


Fig. 4. Convolutional neural network process.

Fig. 4 shows that after the input of the CNN algorithm, the convolutional layers and sub-sampling layers are arranged in an overlapping manner. The collected image features are output through multi-layer convolution and pooling operations. The sub-sampling layer can process and transmit data within the selected range, and the processed features are reduced in model complexity through fuzzy operations. To further enhance the learning ability of the network, ResNet residual network technology is used to complete multi-layer data learning. The calculation process of the residual network is represented by Eq. (8).

$$H(x) = F(x) + x \tag{8}$$

In Eq. (8), $H(x)$ represents the functional expression of the residual process. $F(x)$ represents the module expression function in the residual process. x serves as the feature image for the input stage. The residual formula obtained through transformation processing is Eq. (9).

$$y = R(H(x)) \tag{9}$$

In Eq. (9), y represents the output value. R represents the ReLU activation function used in the network model. The residual module can effectively alleviate the gradient vanishing during signal propagation. The study strengthens the application of residual modules by setting a Bottle-neck structure, further increasing the depth of the network and ensuring that the enhanced deep network maintains efficient learning ability. In order to optimize the process of transforming image features from high dimensions to low dimensions, a 1×1 convolutional layer is added to the original CNN model, and local normalization is adopted after each convolutional layer to enhance the data transmission process. The extracted image features are detected by calculating the similarity between the network input and the matched ethnic art patterns using Euclidean distance. The retrieval results are determined based on the similarity measure. The calculation process of Euclidean distance is represented by Eq. (10).

$$d(x,y)\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{10}$$

In Eq. (10), d represents the feature distance. x_i represents the i -th feature of one of the two images used to calculate distance. x_i represents the i -th feature of another image for calculating distance. Ethnic art products prioritize symmetrical or rotated patterns with simple patterns. The processed patterns are arranged and stacked to form complex patterns with distinctive features. Therefore, in image feature extraction, it is necessary to more effectively identify the basic image patterns. To optimize the visual feature extraction efficiency in image retrieval, the bar attention mechanism is used to process different features. The specific working mode is shown in Fig. 5.

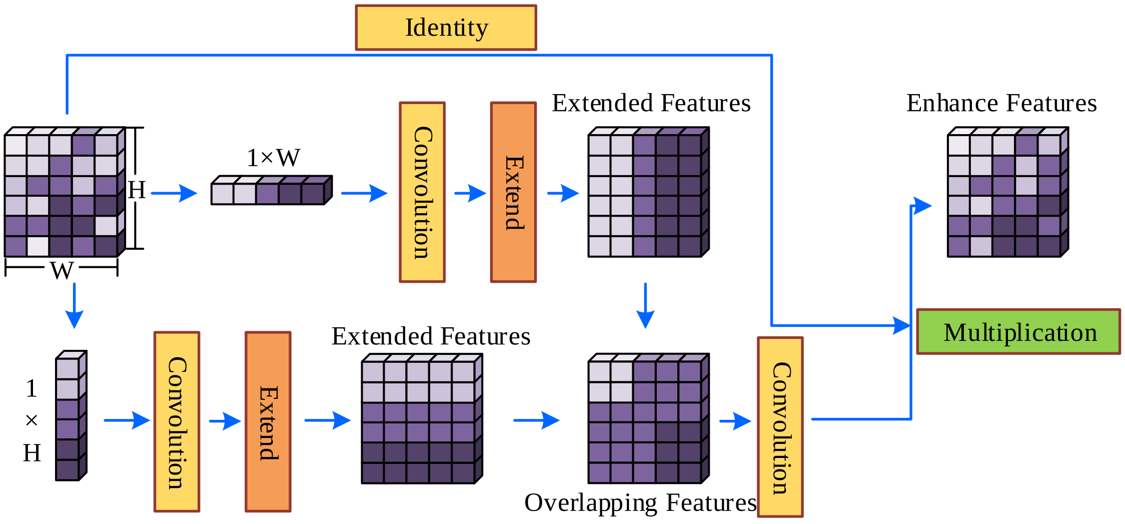


Fig. 5. Bar attention mechanism.

In Fig. 5, the feature extraction mode of the model is composed of a single vertical bar box and a single horizontal bar box. The two sets of standard bar boxes respectively enhance the feature attention points in the current mode, and form a set of enhanced feature data through expansion and combination, filtering out feature maps with less feature information. By optimizing the attention mechanism, the network model magnifies the local feature points of the pattern, resulting in more accurate recognition of effective patterns. The perceptual field of the horizontal bar box in the feature enhancement process is represented by Eq. (11).

$$y_i^h = \frac{1}{W}\sum_{j=1}^{W} x_{i,j} \tag{11}$$

In Eq. (11), y_i^h is the perceived visual field output of the horizontal box. W represents the width value of the space. i signifies the current row value. j signifies the current column value. The perceived visual field of the corresponding vertical bar frame is represented by Eq. (12).

$$y_j^w = \frac{1}{H}\sum_{i=1}^{H} x_{i,j} \tag{12}$$

In Eq. (12), y_j^w D represents the perceived visual field output of the vertical box. H represents the height value of space. Due to the shape setting of the window visual field, a long line connection is established between the discrete region and the main features. The value box of the bar range also makes the feature extraction of the image more targeted, which can enhance attention from two directions and integrate feature maps. To optimize the efficiency of image feature processing, a multi-channel convolution operation is performed, as shown in Fig. 6.

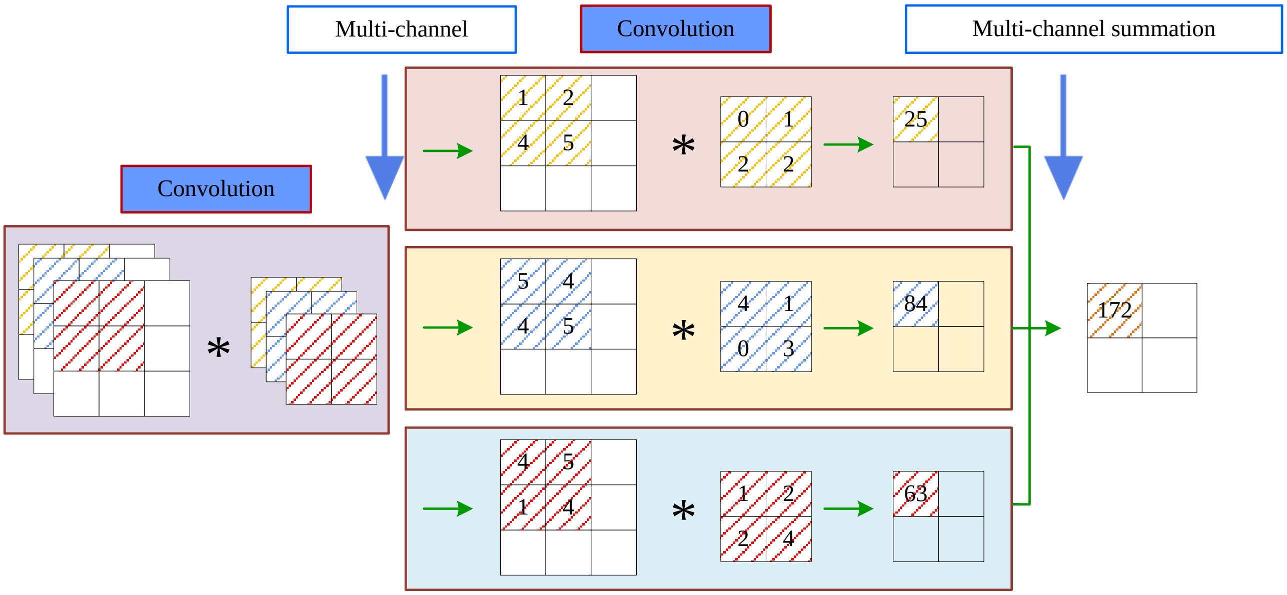


Fig. 6. Multi-channel convolution operation.

In Fig. 6, in the input stage, the convolution kernel dimensionality determines the convolution calculation method. Multiple channels are given priority in completing their respective convolution operations. In the output stage, the sum of convolution values for multiple channels is calculated. The convolution kernel moves with a set stride, and the output size of the feature values is consistent with the dimension size of the convolution kernel. The calculation process of convolution operation is shown in Eq. (13).

$$x^{l+1} = w^l * x^l + b^l \tag{13}$$

In Eq. (13), x^{l+1} represents the output value after convolution operation. x^l represents the operation output of the upper. w^l represents the computed convolution kernel. $*$ represents convolution operation. b^l represents the bias term of the current layer. l represents the number of layers in the convolution operation. To make the self-learning mechanism of the network more efficient, dependency relationships can be established between network channels. The channel attention mechanism can be strengthened through training feedback, which selectively integrates the extracted local features to obtain feature maps with more information. The research model divides adjacent pooling layers using the overlap pooling algorithm, which focuses on global features through the overlapping parts of the region. The specific algorithm process is shown in Fig. 7.

In Fig. 7, it is shown that the feature map after overlapping pooling can cover global features. The input image is convolved and overlapped with pooled feature regions. The overlapped pooled feature map is output using the global average pooling method. The feature maps that have undergone overlapping pooling highlight important features and weaken a small amount of information regions, reducing the image blur problem of global average pooling output. The specification calculation of the new feature map generated by overlap pooling is shown in Eq. (14).

$$n = \frac{m - f}{s} + 1 \tag{14}$$

In Eq. (14), n represents the output value specification. m represents the input value specification. f signifies the size of the convolution kernel. s represents the selected step size. The single channel calculation process of global average pooling is displayed in Eq. (15).

$$\zeta'_n(I) = \frac{\sum_{y_i=1}^{H'}\sum_{x_i=1}^{W'} f(x_i, y_i)}{H' \times W'} \tag{15}$$

In Eq. (15), $\zeta'_n(I)$ is used as the output value of feature map I on channel n. H' represents the height of the newly output feature map. W' represents the width of the newly output feature map. $f(x_i, y_i)$ represents the feature points of the new output feature map. Therefore, when constructing the image retrieval module for ethnic art design products, the DF algorithm is prioritized to enhance feature classification extraction. The DF-CNN is designed in combination with a CNN, and the residual network is used to optimize feature map transmission. The strip attention mechanism and multi-channel operation are used to optimize the image feature extraction process. The global features are integrated through overlap pooling to achieve image retrieval output.

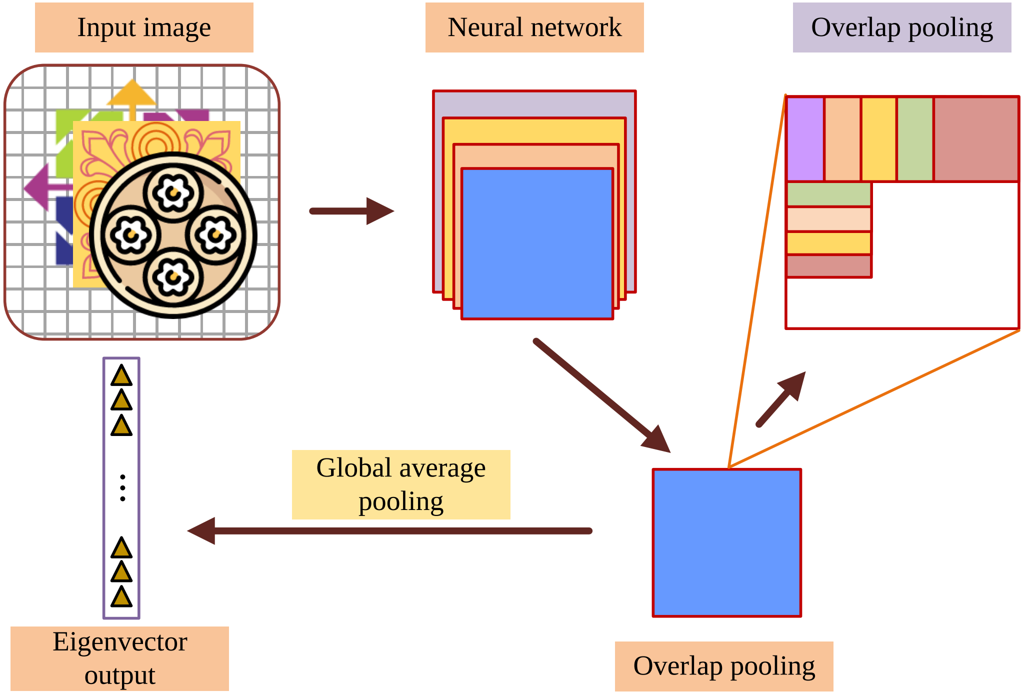
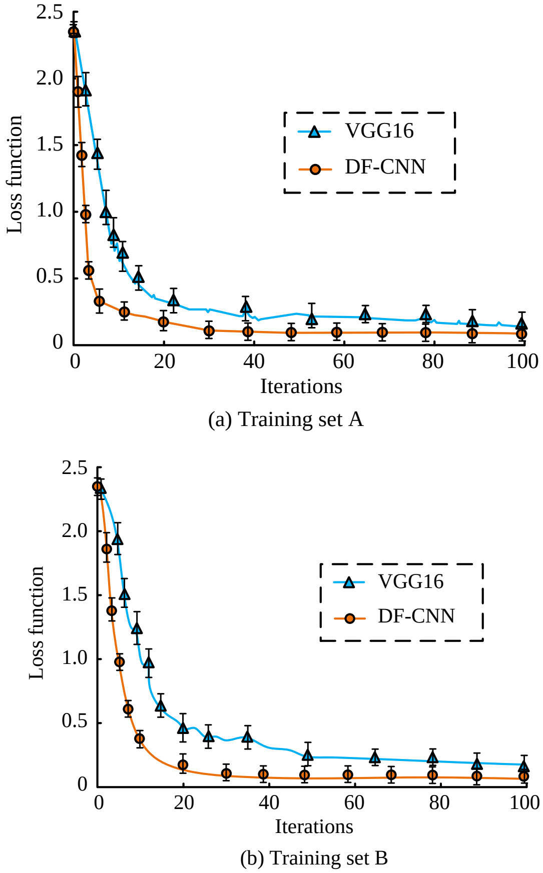


Fig. 7. Overlap pooling and global average pooling.

IV. Results

A. Performance Testing of Image Retrieval Model for Ethnic Art Products

To test the image detection performance of the model, a training set is selected to enhance the learning ability of the model. The pre-testing ratio is set to 10%, and the network threshold of the model is 0.5. The sample images are selected from the CIFAR-10 dataset, and 200 images are set for each group based on the shape of the image pattern as test samples. The convolutional layer during the testing process is set to 16 layers, and the fully connected layer is set to three layers. The learning rate is set to 0.001 and the training iterations are 100. The iterative changes of the loss function in the training set is shown in Fig. 8.

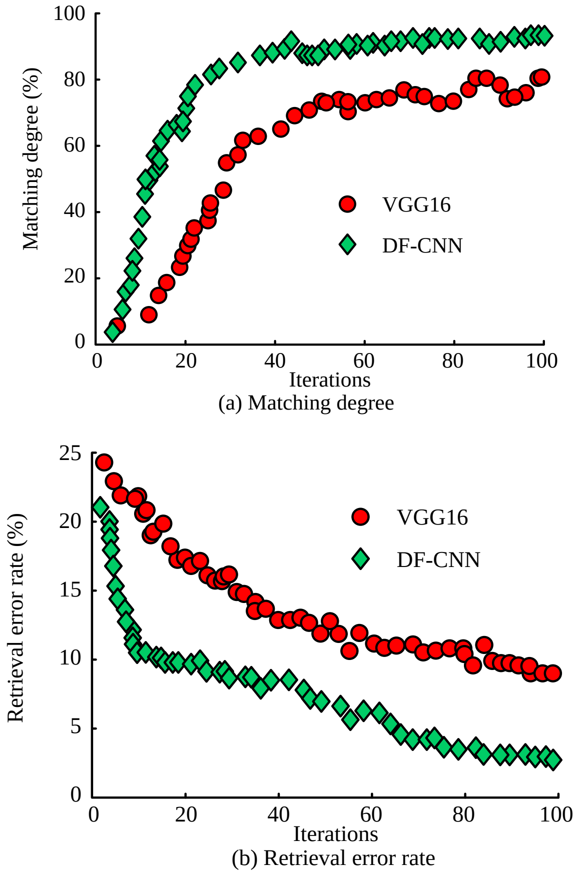


(a) Training set A

(b) Training set B

Fig. 8. Changes in the loss function in the training set.

In Fig. 8 (a), the loss function of the DF-CNN model in training set A decreased with increasing iterations. When the iterations reached 10, it realized stable convergence. The loss function variation of the Visual Geometry Group Network (VGG16) model [31] in training set A also decreased with increasing iterations. When the number of iterations reached 20, the loss function tended to converge, but showed slight fluctuations in subsequent iterations. In Fig. 8 (b), the loss function of the DF-CNN model in training set B was consistent with the results in training set A, but the convergence speed was slightly slower. After 20 iterations, the loss function converged stably. The loss function of VGG16 model in training set B tended to converge after 25 iterations, but the converged loss function value showed a slight rebound. The research model in the training set shows a stable trend and performs better in image detection compared with the VGG16 model. To further validate the pattern retrieval matching of the model in the testing set, the retrieval results of the model for the samples are shown in Fig. 9.



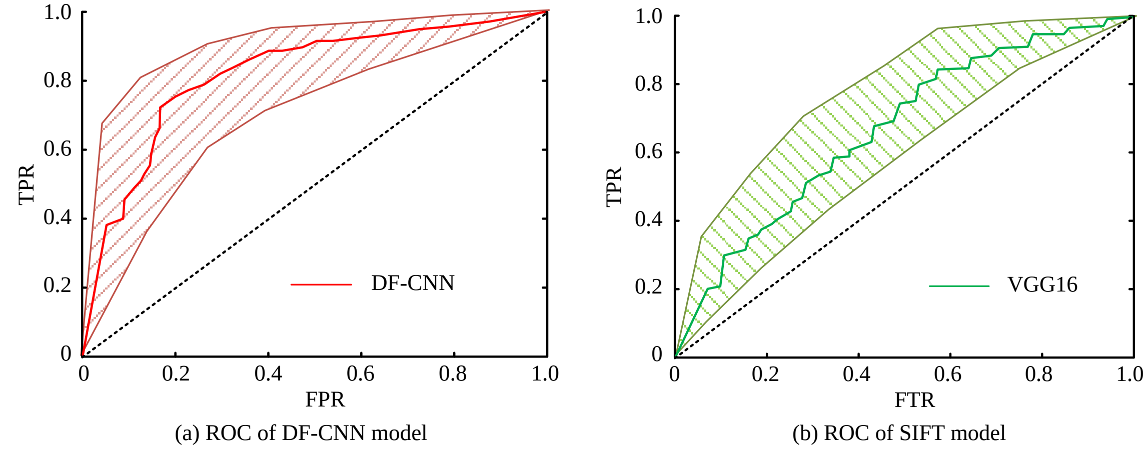
(a) Matching degree

(b) Retrieval error rate

Fig. 9. Pattern retrieval performance in the testing set.

In Fig. 9 (a), the DF-CNN model gradually increased its matching rate with image samples during the iterative learning process. After 23 iterations, the matching degree of the research model reached over 80%. When the model completed learning, the retrieval matching degree of the image reached 91.28%. The VGG16 model had a lower image retrieval matching degree compared with the research model, with an image matching degree of 80.69% after the model completed training and learning. In Fig. 9 (b), the error rate of image retrieval in the DF-CNN model gradually decreased during the iteration process. When the iteration reached 10 times, the image retrieval rate of the model decreased to within 10%. The error rate of the model after training was reduced to 2.91%. The error rate of the VGG16 model in the testing set steadily decreased with increasing iterations, but the error rate optimization effect in the research model during iterations was better than that of the VGG16 model. After 100 iterations of learning, the error rate of the VGG16 model was 9.04%. The trained research model demonstrates good performance in recognizing various types of image patterns, with lower retrieval error rates compared with the VGG16 method, and better image retrieval performance. To ensure the application effectiveness of the research model in image retrieval, the Receiver Operating Characteristic Curve (ROC) is used to evaluate the model.

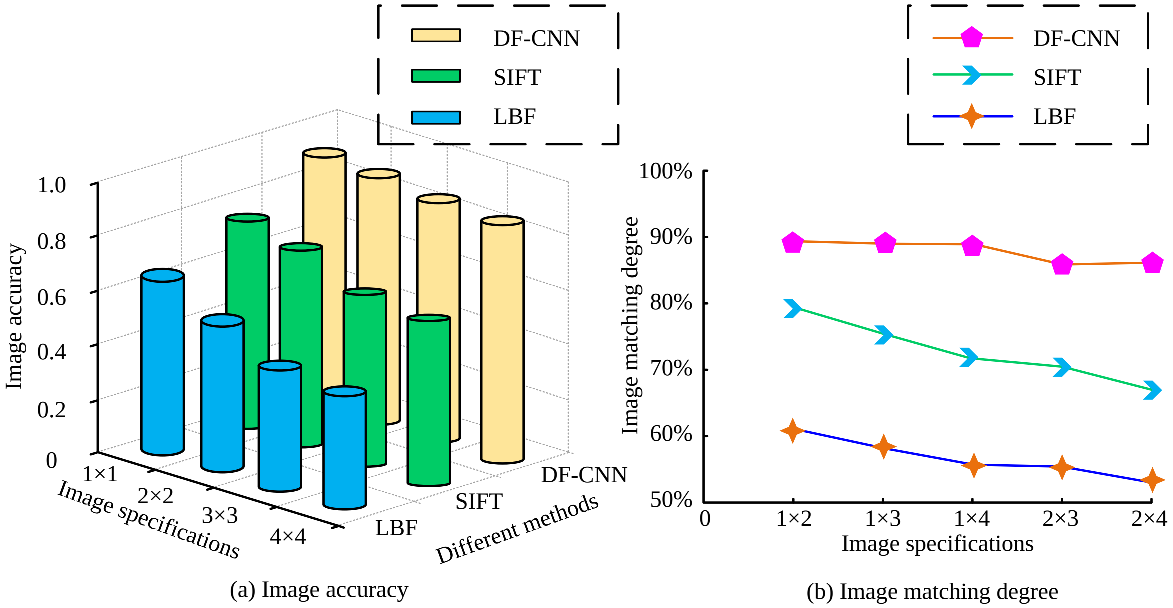
Fig. 10(a) shows the ROC of the DF-CNN model. The ROC curve of the DF-CNN was convex towards the upper left and away from the threshold line, with an Area Under Curve (AUC) of 0.876. Fig. 10(b) shows the ROC curve of the SIFT algorithm. The convex distance of the curve was slightly further to the upper left and the curve range was completely in the lower layer of the research model. The classification performance of the research model performs better throughout the entire process. The ROC curve of the SIFT algorithm was closer to the threshold line and the final AUC value was 0.716. Therefore, the classification performance of the SIFT algorithm [32] is average. This indicates that the image classification performance of the research model has significant advantages, and its ability to retrieve images of ethnic art products is stronger than that of SIFT method in terms of recognition. To further analyze the image comparison ability of the model, the results of non-uniform specification image retrieval under different methods are shown in Fig. 11.



(a) ROC of DF-CNN model

(b) ROC of SIFT model

Fig. 10. Comparison of ROC curves of different models.



(a) Image accuracy

(b) Image matching degree

Fig. 11. Comparison of retrieval effects for multi-specification images.

In Fig. 11 (a), the DF-CNN model exhibits different retrieval accuracy in the selected images of various specifications, with a decreasing trend in retrieval accuracy as the size of the image increases. The retrieval rates of SIFT and LBF algorithms for images of various specifications were consistent with the research method. However, the research model had a relatively small range of accuracy changes for image retrieval of different specifications, with a difference of 4.16% in accuracy under different specifications. The accuracy difference of LBF algorithm and SIFT algorithm for image retrieval of different specifications was 16.21% and 18.97%, respectively. In Fig. 11 (b), the image retrieval accuracy of the three models decreased with increasing image size under irregular image sizes. The DF-CNN model had an average accuracy of 88.49% for non-equal length image type retrieval, while the LBF algorithm and SIFT algorithm had average accuracy of 73.94% and 57.59% for non-equal length image type retrieval, respectively. The results indicate that LBF and SIFT have poor image retrieval performance under the current specifications, while the research model performs well in retrieving images of multiple specifications, which also shows good adaptability in retrieving images with a wide variety of specifications.

B. *Application Effect Test of Image Retrieval Model for Ethnic Art Products*

To visualize the application effect, a designed image retrieval webpage is used to complete the simulation image retrieval of ethnic art design products. The computer processor used for the simulation experiment is Intel(R) Xeon(R)Platinum, and the GPU specification is RTX 2080 8GB*4. An example of retrieving ethnic art images using research methods is shown in Fig. 12.

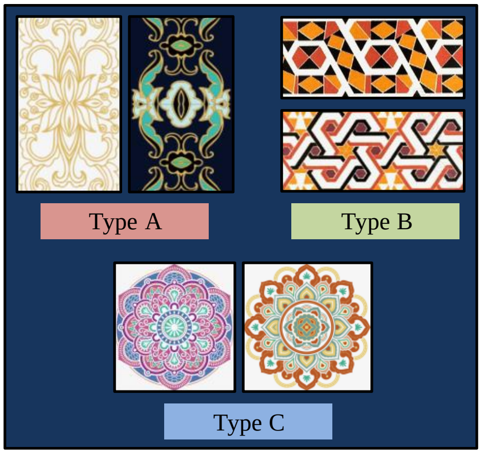
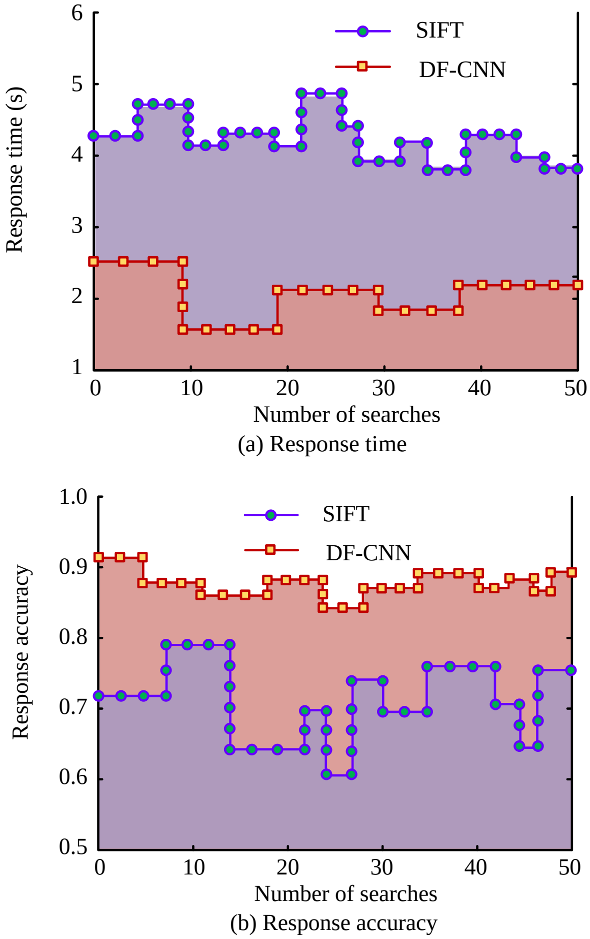


Fig. 12. Retrieve image examples.

Fig. 12 shows the partial image results obtained through keyword retrieval, indicating a clear distinction in the retrieval types of the images. To further test the detection efficiency of the image, the retrieval response time and response accuracy were obtained through 50 retrieval processes, as shown in Fig. 13.

The average response time of the DF-CNN model for 50 image retrieval processes shown in Fig. 13(a) was 1.87s, with the longest response time of the model for retrieval results being 2.47s and the shortest response time being 1.44s. The average response time of the SIFT model for 50 retrieval processes was 4.21s, with the longest response time being 4.86s and the shortest response time being 3.88s. In Fig. 13 (b), the average image retrieval matching degree of the DF-CNN model in 50 searches was 0.879. The optimal matching degree of the research model for images was 0.913, and the matching degree of the image with the worst retrieval effect was 0.843. The average image matching degree of the SIFT model for 50 retrieval processes was 0.701. This indicates that the research model shows a high search speed in the image retrieval, with a response speed improvement of 51.8% compared with the SIFT algorithm. The DF-CNN model also shows a high matching rate for image retrieval, which is 16.8% higher than the image matching rate of the SIFT algorithm. To compare the image retrieval performance under various conditions, the image retrieval results of the research model within one week are summarized in Table I.



(a) Response time

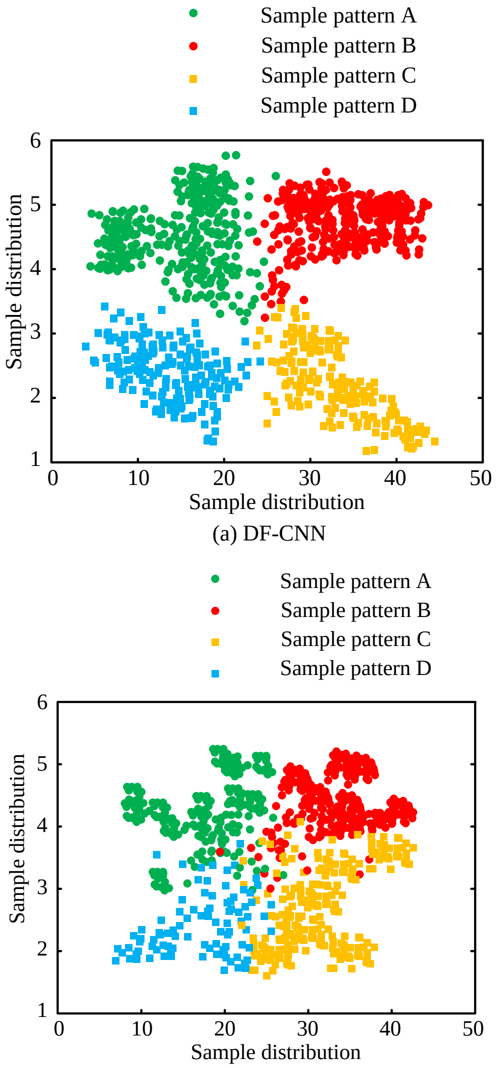
(b) Response accuracy

Fig. 13. Search response during the application process.

In Table I, the three types of models with a code length of 128 had shorter response times for image detection, among which the DF-CNN model had the fastest response speed in the detection process. The three types of models with a code length of 512 showed longer response times for image detection, and the DF-CNN model remained the fastest response speed. However, as the code length increased, the mean Average Precision (mAP) and recall of the three types of models showed an upward trend. The DF-CNN model demonstrated good retrieval precision and sensitivity. Compared with the VGG16 algorithm and DF-CNN algorithm, the mAP value optimization range of the research model was 6.1% -29.8%, and the sensitivity optimization range was 15.33% -36.9%. The research model shows that the retrieval speed of images is faster under shorter code lengths, but the retrieval precision is correspondingly reduced. The retrieval precision of images is improved under longer code lengths, but the corresponding retrieval time is longer. The DF-CNN model shows excellent performance in both detection precision and response time during the detection process. To distinguish and recognize different patterns, the clustering results of detection samples under different algorithms are compared, as shown in Fig. 14.

TABLE I. COMPARISON OF RETRIEVAL PERFORMANCE OF DIFFERENT MODELS

Retrieval Model	Code-Length	Recall	mAP	Time/s	References
DF-CNN	128	53.30%	76.4%	1.64	/
DF-CNN	512	57.40%	90.1%	4.62	/
VGG16	128	46.80%	69.3%	1.82	[31]
VGG16	512	48.60%	84.6%	5.07	[31]
SIFT	128	33.90%	45.3%	2.85	[32]
SIFT	512	36.20%	63.2%	8.74	[32]

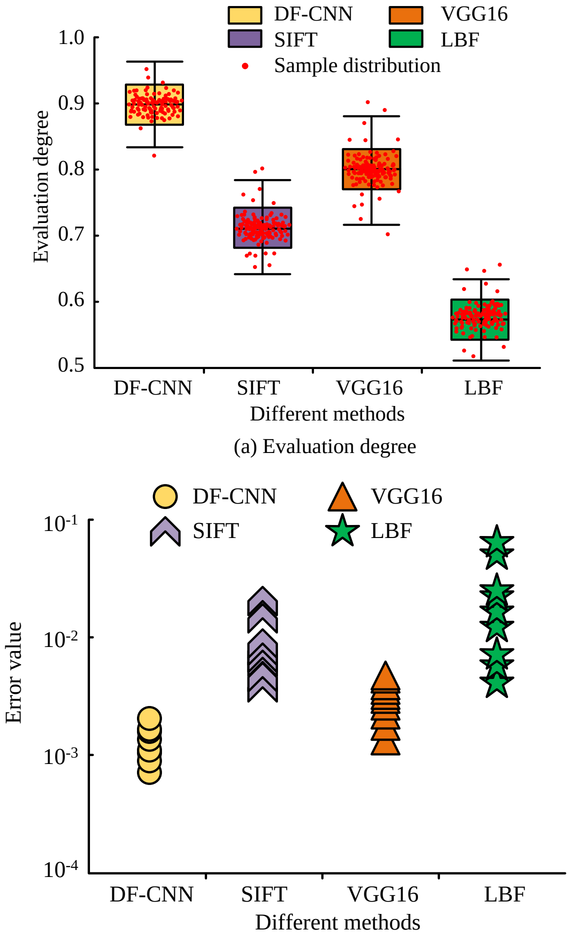


(a) DF-CNN

(b) VGG16

Fig. 14. Clustering of detection samples under different algorithms.

In Fig. 14 (a), the four types of pattern detection samples in the DF-CNN model were clustered separately, and the boundary parts of the four types of samples were adjacent but not intersecting. In Fig. 14 (b), the four pattern type detection samples in the VGG16 model also exhibited their own clustering, but there was partial intersection at the boundaries of the four types of samples. The DF-CNN model showed good recognition performance for four patterns and strong recognition ability for similar images. The VGG16 model can achieve basic recognition and detection for four types of patterns. When the patterns presented in the retrieved images are similar, the VGG16 model may exhibit confusion in image detection. The patterns of ethnic art products may appear similar but not identical, and the VGG16 model shows a significant lack of retrieval ability for such images. The DF-CNN model can accurately identify the differences in such images, indicating that the research model has a better retrieval application effect for ethnic art products. To verify the detection accuracy of the model, the multi-sample image detection accuracy and image detection error values under different models are shown in Fig. 15.



(a) Evaluation degree

(b) Error range

Fig. 15. Image detection accuracy and detection error.

In Fig. 15 (a), the average image accuracy evaluation value of the DF-CNN model in multi-sample image detection was 0.898, with only a few samples scattered outside the detection floating range. The image accuracy evaluation values of SIFT algorithm, VGG16 algorithm, and LBF algorithm were 0.731, 0.804, and 0.579, respectively, and all three algorithms had some samples scattered outside the detection floating range during the image detection process. In Fig. 15 (b), the detection error range of the DF-CNN model in multi-sample image detection was 8.7×10-4-4.3×10-3, while the image detection error ranges of the SIFT algorithm, VGG16 algorithm, and LBF algorithm were 4.7×10-3-4.1×10-2, 1.1×10-3-7.9×10-3, and 4.8×10-3-9.1×10-2, respectively. This demonstrates that the research method has good stability in image detection function. Compared with other algorithms, the image accuracy has always been maintained at a good level. The image detection error of the research model is smaller and the error value is lower compared with other algorithms, indicating that the research model has good image detection ability during operation.

C. Discussion

The above research results show that the research method has good matching performance for the retrieval of images of ethnic art products, and the accuracy of image retrieval is better than that of the SIFT algorithm. Due to the SIFT algorithm's emphasis on local features in image detection, and the lack of significant differences in local features of ethnic art products, the retrieval matching degree of images decreases. The research model utilizes the cascaded structure of the DF algorithm to enhance the depth of image feature mining, strengthen the correlation between retrieval keywords and corresponding images, and thus improve the retrieval accuracy of the design module. At the same time, the retrieval efficiency of the research method also shows significant advantages. In the process of retrieving ethnic art images, the retrieval response time of the research method is significantly better than that of the VGG16 algorithm. The VGG16 algorithm enhances its feature extraction ability through multi-layer convolution, but due to the complex calculation process, the retrieval speed of the algorithm decreases. The prior box design of the research model increases the pre-screening of image features, reduces the computational complexity of the image-matching process, and thus achieves an improvement in retrieval efficiency. It can be seen that the research model achieves synchronous improvement in retrieval matching and retrieval efficiency through the fusion of advantages between different algorithms.

V. Conclusion

In order to enhance the retrieval ability of art galleries for ethnic art patterns, an image retrieval model is constructed by combining the DF algorithm with CNN. The decision tree was used to explore the transformation paths of images, and the overlap window was used to enhance the global features of features. Multi-granularity scanning processes optimized feature loss during dimension transformation. The Euclidean distance algorithm was used to calculate the similarity of image features. Parallel multi-channel convolution was applied to optimize the transmission rate of features, focusing on global features through overlapping pooling algorithm, and completing image retrieval output by combining global and local features. The research results indicated that the error rate of the DF-CNN model was reduced to 2.91% after training, which was 6.13% lower than the retrieval error rate of the VGG16. The difference in retrieval accuracy of the DF-CNN model for multi-specification image retrieval was 4.16%, which was 12.05%-14.81% higher than other algorithms, indicating that the research model had good adaptability to retrieval of various image types and stable retrieval performance for images. In the application process, the average retrieval response time of the DF-CNN model was 1.87s, and the longest response time was 2.47s, which was 51.8% higher than the response rate of the SIFT algorithm. The research model had good image retrieval accuracy while maintaining a high response rate. The average image retrieval matching degree of the model was 0.879, which was 16.8% higher than the SIFT algorithm. The application effect of the research model in practical processes is good, and the efficiency of image retrieval is high. With the widespread dissemination of ethnic art products in more fields, they have also been applied in virtual scenes. However, in the practical application of retrieval methods, virtual scenes are often not included in the retrieval scope, resulting in a lack of retrieval data for virtual scenes. At the same time, the research method has increased the retrieval ability of ethnic art images by adding multiple complex graphic processing layers, which has resulted in a decrease in the retrieval speed of simple images. Therefore, in future research plans, image retrieval scenarios can be expanded to achieve further expansion of retrieval databases. It is possible to attempt to classify the retrieval paths of images, achieving hierarchical processing of simple and complex images, thereby further improving the efficiency of image retrieval and meeting the retrieval needs of different users.

References

[1] C. A. Burks, T. I. Russell, D. Goss, D. Goss, G. Ortega, and G. W. Randolph, "Strategies to increase racial and ethnic diversity in the surgical workforce: A state of the art review," Otolaryng. Head Neck, vol. 166, no. 6, pp. 1182-1191, April, 2022.

[2] A. Olivares and J. Piatak, "Exhibiting inclusion: An examination of race, ethnicity, and museum participation," Int. J. Voluntary Nonprofit Organ., vol. 33, no. 1, pp. 121-133, February, 2022.

[3] T. Zhang, B. Li, and N. Hua, "Chinese cultural theme parks: text mining and sentiment analysis," J. Tour. Cult. Change, vol. 20, no. 1-2, pp. 37-57, January, 2022.

[4] N. Arora and S. C. Sharma, "ETLBP and ERDLBP descriptors for efficient facial image retrieval in CBIR systems," Multimedia Tools Appl., vol. 83, no. 4, pp. 9817-9851, June, 2024.

[5] M. Majhi, A. K. Pal, J. Pradhan, S. H. Islam, and M. K. Khan, "Computational intelligence based secure three-party CBIR scheme for medical data for cloud-assisted healthcare applications," Multimedia Tools Appl., vol. 81, no. 29, pp. 41545-41577, February, 2022.

[6] S. Saurav, R. Saini, and S. Singh, "Fast facial expression recognition using boosted histogram of oriented gradient (BHOG) features," Pattern Anal. Appl., vol. 26, no. 1, pp. 381-402, September, 2023.

[7] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," Int. J. Multimed. Inf. R., vol. 11, no. 1, pp. 19-38, September, 2022.

[8] M. K. Kelishadrokhi, M.Ghattaei, and S. Fekri-Ershad, "Innovative local texture descriptor in joint of human-based color features for content-based image retrieval," Signal, Image Video P., vol. 17, no. 8, pp. 4009-4017, July, 2023.

[9] N. Keisham and A. Neelima, "Efficient content-based image retrieval using deep search and rescue algorithm," Soft Comput., vol. 26, no. 4, pp. 1597-1616, January, 2022.

[10] W. Wang, P. Jiao, H. Liu, X. Ma, and Z. Shang, "Two-stage content based image retrieval using sparse representation and feature fusion," Multimed. Tools Appl., vol. 81, no. 12, pp. 16621-16644, March, 2022.

[11] C. Ning, Y. Di, and L. Menglu, “Survey on clothing image retrieval with cross-domain," Complex Intell. Syst., vol. 8, no. 6, pp. 5531-5544, May, 2022.

[12] H. Zhuang, X. Liu, Y. Yan, D. Zhang, J. He, J. He, X. Zhang, H. Zhang, and M. Li, "Integrating a deep forest algorithm with vector-based cellular automata for urban land change simulation," Trans. GIS, vol. 26, no. 4, pp. 2056-2080, April, 2022.

[13] A. Hamedianfar, C. Mohamedou, A. Kangas, and J. Vauhkonen, "Deep learning for forest inventory and planning: A critical review on the remote

sensing approaches so far and prospects for further applications," Forestry, vol. 95, no. 4, pp. 451-465, October, 2022.

[14] M. A.Shaaban, Y. F. Hassan, and S. K. Guirguis, "Deep convolutional forest: a dynamic deep ensemble approach for spam detection in text," Complex Intell. Syst., vol. 8, no. 6, pp. 4897-4909, April, 2022.

[15] T. Huang, Q. Zhang, X. Tang, S. Zhao, and X. Lu, "A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems," Artifi. Intell. Rev. vol. 55, no. 2, pp. 1289-1315, April, 2022.

[16] A. Tayal, J. Gupta, A. Solanki, K. Bisht, A. Nayyar, and M. Masud, "DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases," Multimedia Syst., vol. 28, no. 4, pp. 1417-1438, March, 2022.

[17] M. Zhitomirsky-Geffet and S. Minster, "Cultural information bubbles: A new approach for automatic ethical evaluation of digital artwork collections based on Wikidata," Digit. Scholarsh. Hum., vol. 38, no. 2, pp. 891-911, June, 2023.

[18] W. Serhan, "Symbolic capital and the inclusion of ethnic minority artists in Dublin and Warsaw," Ethnicities, vol. 24, no. 3, pp. 475-496, June, 2024.

[19] M. Zhitomirsky-Geffet, I. izhner, and S. Minster, "What do they make us see: a comparative study of cultural bias in online databases of two large museums," J. Doc., vol. 79, no. 2, pp. 320-340, July, 2023.

[20] C. Sutherland, "Transcending ethnicity through photography: representing the Cham," Asian Ethn., vol. 23, no. 1, pp. 127-145, March, 2022.

[21] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," ACM Trans. Inform. Syst., vol. 40, no. 4, pp. 1-42, March, 2022.

[22] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal graph matching network for image-text retrieval," ACM Trans. Multim. Comput., vol. 18, no. 4, pp. 1-23, March, 2022.

[23] N. T. Dinh, N. T. U. Nhi, T. M. Le, and T. T. Van, "A model of image retrieval based on KD-tree random forest," Data Technol. Appl., vol. 57, no. 4, pp. 514-536, May, 2023.

[24] N. Bharatha Devi, "Satellite image retrieval of random forest (rf-PNN) based probablistic neural network," Earth Sci. Inform., vol. 15, no. 2, pp. 941-949, February, 2022.

[25] P. Ma, Y. Wu, Y. Li, L. Guo, H. Jiang, and X. Zhu, "HW-Forest: Deep forest with hashing screening and window screening," ACM Trans. Knowle. Discov. Data, vol. 16, no. 6, pp. 1-24, July, 2022.

[26] S. Kumar, A. K. Pal, S. K. H. Islam, and M. Hammoudeh, "Secure and efficient image retrieval through invariant features selection in insecure cloud environments," Neural Comput. Appl., vol. 35, no. 7, pp. 4855-4880, June, 2023.

[27] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Þ.Jónsson, and J. Lokoč, "Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown," Int. J. Multimed. Inf. Retr., vol. 11, no. 1, pp. 1-18, January, 2022.

[28] L. Shi, J. Du, G. Cheng, X. Liu, Z. Xiong, and J. Luo, "Cross-media search method based on complementary attention and generative adversarial network for social networks," Int. J. Intell. Syst., vol. 37, no. 8, pp. 4393-4416, November, 2022.

[29] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," ACM Trans. Inform. Syst., vol. 40, no. 4, pp. 1-42, March, 2022.

[30] P. Preethi and H. R. Mamatha, "Region-based convolutional neural network for segmenting text in epigraphical images," Artif. Intell. Appl., vol. 1, no. 2, pp. 119-127, September, 2023.

[31] S. Kumar, M. K. Singh, and M. Mishra, "Efficient deep feature based semantic image retrieval," Neural Process. Lett., vol. 55, no. 3, pp. 2225-2248, January, 2023.

[32] W. Wang, P. Jiao, H. Liu, X. Ma, and Z. Shang, "Two-stage content based image retrieval using sparse representation and feature fusion," Multimed. Tools Appl., vol. 81, no. 12, pp. 16621-16644, March, 2022.

Application of Speech Recognition Technology Based on Multimodal Information in Human-Computer Interaction

Yuan Zhang
Xuchang Vocational Technical College
Henan Province, Data Intelligence and Security Application Engineering Technology Research Center, Xuchang 461000, China

Abstract*—Multimodal human-computer interaction is an important trend in the development of human-computer interaction field. In order to accelerate the technological change of human-computer interaction system, the study firstly fuses Connectionist Temporal Classification algorithm and attention mechanism to design a speech recognition architecture, and then further optimizes the end-to-end architecture of speech recognition by using the improved artificial swarming algorithm, to obtain a speech recognition model suitable for multimodal human-computer interaction system. One of them, Connectionist Temporal Classification, is a machine learning algorithm that deals with sequence-to-sequence problems; and the Attention Mechanism allows the model to process the input data in such a way that it can focus its attention on the relevant parts. The experimental results show that, the hypervolume of the improved swarm algorithm converges to 0.861, which is 0.099 and 0.059 compared to the ant colony and differential evolution algorithms, while the traditional swarm algorithm takes the value of 0.676; the inverse generation distance of the improved swarm algorithm converges to 0.194, while that of the traditional swarm, ant colony, and differential evolution algorithms converge to 0.263, 0.342, and 0.246, respectively. Hypervolume and Inverse Generation Distance Measures the diversity and convergence of the solution set. The speech recognition model takes higher values than the other speech recognition models in the evaluation metrics of accuracy, precision, and recall, and the lowest values of the error rate at the character, word, and sentence levels are respectively 0.037, 0.036 and 0.035, ensuring higher recognition accuracy while weighing the real-time rate. In the multimodal interactive system, the experimental group's average opinion scores, objective ratings of speech quality, and short-term goal comprehensibility scores, and the overall user experience showed a significant advantage over the control group of the other methods, and the application scores were at a high level. The speech processing technology designed in this study is of great significance for improving the interaction efficiency and user experience, and provides certain references and lessons for the research in the field of human-computer interaction and speech recognition.

Keywords—Multimodal information; speech recognition; intelligent optimization algorithm; multimodal human-computer interaction; CTC; attention mechanisms; artificial bee colony algorithms

I. INTRODUCTION

Human-computer interaction (HCI) aims to realize information exchange and interaction between human and computer through computer technology. Single-modal human-computer interaction has been unable to meet the growing user needs, and multimodal human-computer interaction system is gradually becoming a research hotspot [1, 2]. Multimodal human-computer interaction (MMHCI) is a system that uses multiple input and output modes to carry out human-computer interaction, integrating various technologies such as speech recognition, gesture recognition, speech synthesis, expression parsing, etc., which is able to better adapt to the user's personalized needs, and has the advantages of improving interaction experience and interaction efficiency. It has the advantage of improving interaction experience and interaction efficiency [3].

At present, MMHCI is still a relatively new type of human-computer interaction and most of its key technologies are still in the exploration stage. Speech recognition is a key technology in the cognitive decision-making aspect of MMHCI, which affects the overall performance of the system. Currently, Hidden Markov, Recurrent Neural Network, Transformer structure and Connectionist Temporal Classification (CTC) are mostly used in the construction of speech recognition models, but the key technology enhancement of the models focuses on the accuracy rate, feature extraction, etc., and does not pay much attention to the specific speech recognition tasks and application environments. The recognition accuracy, real-time efficiency, semantic and emotional understanding, and single-signal processing of existing speech recognition technology still cannot meet the application requirements of MMHCI [4, 5].

In order to enhance the accuracy, robustness and improve the interactivity and flexibility of speech recognition in the interaction process of MMHCI speech recognition technology, the study utilizes inputs in the form of audio and video formed by multimodal information such as speech, semantics, expression, text and video, and designs a framework for speech recognition model with the help of improved CTC algorithm and a mixture of attention mechanisms. The adaptive multiple search strategy is also implemented on the basis of Artificial Bee Colony Algorithm (ABC) and this is used to complete the improved optimization of speech recognition model.

The study takes MMHCI as the research object, and carries out research on the key technologies of its speech recognition module, which helps to improve the accuracy and robustness of speech recognition, and obtains technological innovations, and provides theoretical support for the speech recognition algorithm of MMHCI. The introduction of ABC algorithm provides a new solution to the optimisation problem of speech recognition. The research is expected to improve the reliability and efficiency of speech recognition technology in practical applications, enriching the user interaction experience; at the same time, it expands the application scenarios of speech recognition, promotes the development of MMHCI, and helps to realize a more intelligent and humanized interaction experience. The optimization research results of the key technologies of speech recognition in MMHCI provide new ideas and technologies for the development of the field of human-computer interaction, and promote the integration of the industrial and computer fields.

The study consists of six main sections, Section II is to carry out a summary of existing related research work in the field of speech recognition and human-computer interaction; then Section III elaborates on the construction and improvement of the speech recognition framework of the CTC joint attention mechanism, and designs the optimization algorithm of the improved ABC speech recognition framework; in Section IV, completes the performance test of the speech recognition model and the application analysis; Discussion is given in Section V and finally, Section VI concludes the paper.

II. Related Work

With the development of computer technology and artificial intelligence, multimodal human-computer interaction and speech recognition technology related to human-computer interaction have become a hot topic of concern in the community. Various neural networks and deep learning are widely used in speech recognition. Among the speech recognition researches in different countries, there have been relatively more researches on the recognition of English, Japanese or Chinese. Mukhamadiyev et al. designed an end-to-end deep neural network-hidden Markov model speech recognition model and a hybrid CTC-attention network with the small language Uzbek as the research object. The method can effectively utilize the connection time classification objective function and achieves improved recognition efficiency and accuracy, with a speech recognition error rate of 14.3% on the Uzbek dataset [6]. Deep learning models have been used effectively in speech recognition tasks, Dua et al. extended the application of convolutional neural networks to the recognition of speech signals and developed a speech-to-text recognition system. The method achieves recognition accuracy of 89.15 per cent and word error rate of 10.56 per cent on continuous and extensive lexical sentences of speech signals of different pitches [7]. Świetlicka et al. have used principal component analysis in combination with multilayer perceptual networks for fluent and interfering speech signals to analyse their application in describing the dimensionality reduction of speech signal variables. The experimental results show that the method achieves 76% total classification accuracy compared to Kohonen network [8]. Based on the network training techniques, Reza et al. designed a stacked five-layer custom residual convolutional neural network and seven-layer bi-directional gated recurrent units, where the network units all contain learnable layer normalization techniques based on element affine parameters. The character error rate of the model is 4.7 and 3.61% based on the public datasets librisspeech, LJ Speech validation [9].

With the rapid development of deep learning, artificial intelligence and other computer technologies, gesture recognition, speech recognition and other technologies related to human-computer interaction have also made great progress. The application of speech recognition technology in the field of human-computer interaction is gradually increasing, and researchers have conducted performance improvement studies around specific application tasks of speech recognition technology. Lv et al. summarized the research on human-computer interaction and speech recognition based on Web of Science, an academic literature database. The study found that intelligent human-robot interaction and deep learning have made great progress in gesture recognition, speech recognition and emotion recognition, and deep learning can effectively improve the recognition accuracy [10]. In order to improve the service quality and control effect of robots, Pan designed a command understanding method based on command intent understanding and key information extraction, as well as a human-robot voice interaction system with good application effect based on microphone array, voice wake-up and speech recognition [11]. Mavropoulos et al. designed an MMHC system based on knowledge representation, speech recognition and synthesis, sensor data analysis and Computer vision designed a MMHCI. The system can collect and monitor patient related information for healthcare [12]. Liu et al. conducted a study on speech interaction based on emotional Internet of Things, designing a multi-stage deep transfer learning scheme for the problem of limited large-scale emotion labeled datasets, and the experimental results show that the model is effective and superior in terms of naturalness and emotional expressiveness [13]. The use of speech recognition technology in the teaching process to assist teachers in correcting the pronunciation of spoken English has significant application effects. Ran et al. improved the speech recognition algorithm based on artificial intelligence speech recognition technology and designed a speech cutting model based on phonemic level speech error correction, including speech front-end processing and feature parameter extraction. The experimental results verify the effectiveness of the model [14].

In summary, there have been many applications and researches on speech recognition technology and human-computer interaction, but the technical improvement of the model mainly focuses on improvement of recognition accuracy and the reduction of error rate, and does not pay much attention to the specific speech recognition task and application environment. Moreover, there are relatively few studies on speech recognition for human-computer interaction with multimodal information, and the application of the existing speech recognition technology in multimodal human-computer interaction system is still immature and poorly adapted. In this study, speech recognition technology is improved by utilizing multimodal information.

III. Optimized Speech Recognition Model Design Based on Multimodal Information

In order to meet the technical demands of multimodal human-computer interaction, the study firstly designs a speech recognition framework for multimodal human-computer interaction based on the CTC and Attention mechanism; then improves and optimizes the CTC-Attention decoding scheme;

and finally introduces the ABC algorithm to improve the global searching ability and adaptive ability of the hybrid speech recognition model.

A. Speech Recognition Model Design Based on CTC and Attention Mechanism

Traditional human-computer interaction is mostly limited to keyboard, mouse or touch screen inputs, with a single mode of interaction, mostly relying on independent speech or gesture recognition. MMHCI technology is an important innovation in the field of artificial intelligence and interaction, integrating multiple input modes, including speech, gesture, touch, facial expression and eye movement, etc., and fusing the input data of different modes. Combining data from various input modes enhances the accuracy and intelligence of responses in multimodal human-computer interaction. At the same time, MMHCI can adaptively learn according to the user's interactive behavior habits and provide more convenient and personalized interactive services. MMHCI integrates multiple technologies such as speech recognition, image recognition, motion sensing, etc., and utilizes multi-modal information to complete the interaction, which involves multiple modules such as information input, multi-modal interaction information fusion and processing, multi-modal interaction information feedback, etc. Speech recognition technology is one of the key components of MMHCI, and it is also the most important component of MMHCI, as it can be used for the interaction between different modalities. Speech recognition technology is an important part of MMHCI, which is the key link to enhance user experience and interaction richness [15]. The MMHCI framework designed in the study consists of three parts: data input, cognitive decision control and output. The cognitive decision control includes the recognition of multimodal information such as speech, image and video, text dialogue, and facial expression. The study is aimed at the improvement of speech recognition accuracy, cross-modal information fusion and recognition efficiency, and the basic techniques such as the CTC algorithm and attention mechanism are utilized for speech recognition.

The speech recognition model designed in the study is an end-to-end encoder-decoder network structure. CTC is an algorithm for processing sequence data, which mainly solves the label alignment problem due to the change in the length of the sequence data [16, 17]. CTC contains a CTC loss function, which can be applied to the sequence data with variable-length labels during the training of the neural network without the need to align the speech. The calculation process of CTC network update is shown in Eq. (1), in which X denotes the speech feature sequence, $X = x_1, x_2, ..., x_T$; Y denotes the text sequence, $Y = y_1, y_2, ..., y_U$; $A(Y)$ denotes the set of all the aligned sequences corresponding to Y , a_i ; a_i denotes the path mapped from X to Y ; P denotes the likelihood probability of the mapping; and $blank(\in)$ is the special character introduced by the CTC to solve the problem of the model's input/output alignment.

$$\begin{cases} CTC_{Loss} = \sum_{(X,Y)\in D} -\log P(Y|X) \\ P(Y|X) = \sum_{A\in A(Y)} P(A|X) = \sum_{A\in A(Y)} \prod_{i}^{T} P(a_i|X) \end{cases} \quad (1)$$

Attention Mechanism (AM) is a model that simulates the mechanism of human attention allocation for solving the problem of information filtering and weighting when processing sequence data. The network architecture of the speech recognition model designed for the study draws on the Transformer model. The Transformer model is a model that solves the sequence-to-sequence problem based on the Attention mechanism, and the structure of the Transformer feature extractor is shown in Fig. 1.

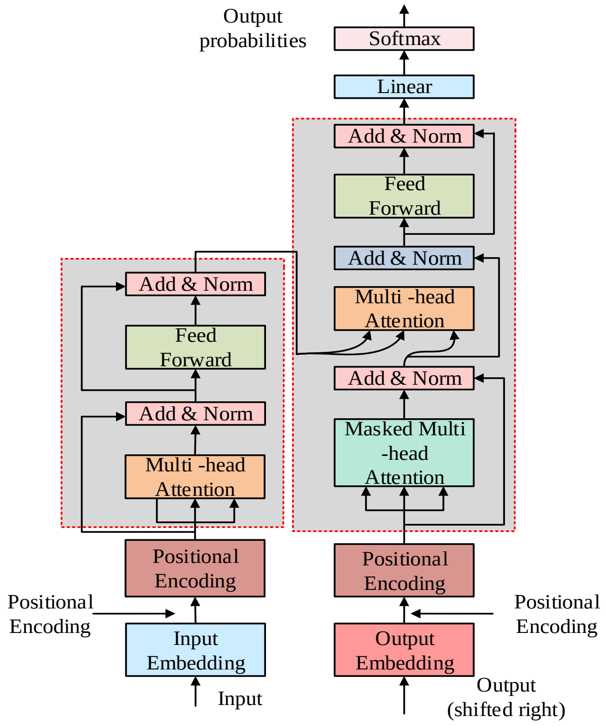


Fig. 1. Structure of transformer feature extractor.

The Transformer model treats the input and output sequences as a series of encoder and decoder stacked layers, with the different layers being composed of Multi-Headed Self-Attention Mechanism (MHSA) and Feedforward neural network (FFN). FFN), AM can weigh different positions in sequence processing. However, Transformer is weak in dealing with fine-grained local feature extraction problems, the study adopts the Conformer model, an improved structure of Transformer, to fully learn the local feature information, and the model structure is shown in Fig. 2. The Conformer model contains a Conformer module, which contains the Self-AM and CNN which is placed between the FFN layers. The Conformer introduces a position-sensitive sinusoidal function when calculating the AM scores, which helps the model to handle the weighting of acoustic features at different time steps. In addition, the Conformer model employs forward weighting and layer normalization to further improve the model performance [18, 19]. The study hybridizes CTC with Conformer model, firstly, Fbank features and 3-dimensional fundamental frequency features are selected for training the speech recognition model, and SpecAugment method is utilized for data enhancement of speech. Then after convolutional sampling, linear mapping and regularization the input encoder Conformer block and finally decoding is done by CTC and AM decoder.

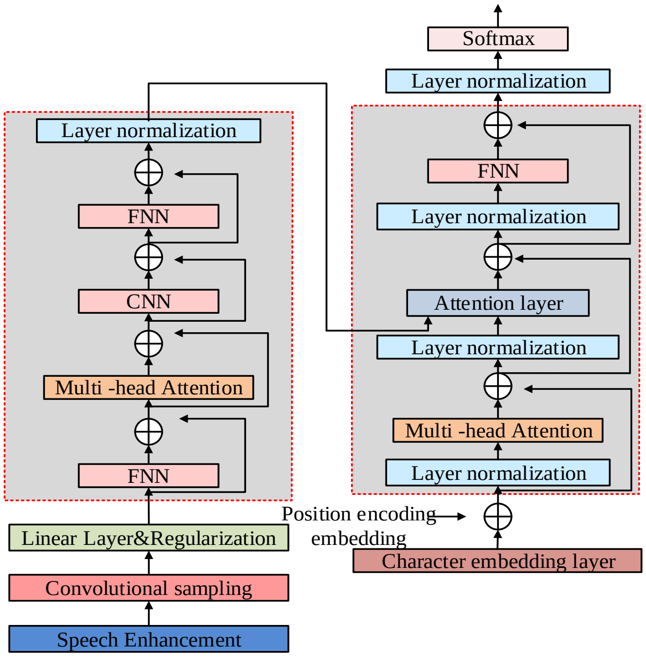


Fig. 2. Structural diagram of the transformer model.

The completion process of the Conformer model is shown in Eq. (2) and x denotes the input.

$$\begin{cases} x_{FFN1} = x + \frac{1}{2} FFN(x) \\ x_{MHSA} = x_{FFN1} + MHSA\left(x + \frac{1}{2} FFN(x)\right) \\ x_{Conv} = x_{MHSA} + Conv(x_{MHSA}) \\ x_{FFN2} = Layernorm\left(x_{Conv} + \frac{1}{2} FFN(x_{Conv})\right) \end{cases} \quad (2)$$

The loss function calculation of the CTC-Conformer hybrid model is shown in Eq. (3), where x, y denotes the acoustic features, the real text, respectively; and λ denotes the hyperparameters balancing CTC and AM.

$$Loss_{combined}(x,y) = \lambda * Loss_{ctc}(x,y) + Loss_{attention}(x,y) \quad (3)$$

The decoding score for the CTC-Conformer hybrid model is calculated in Eq. (4).

$$Score_{final} = \lambda * Score_{ctc} + Score_{attention} \quad (4)$$

B. Improvement of CTC-AM Speech Recognition Framework

CTC-Conformer hybrid model has achieved a large advantage over the traditional Transformer model in speech recognition, but it still has shortcomings such as model convergence difficulty and complex calculation. In this study, improvements are made to AM and CTC algorithms. The decoding process of CTC-Conformer is shown in Fig. 3, as seen in Fig. 3, it is difficult for the model to converge in the face of a longer feature sequence input; and the computational complexity of Self-AM is $O(L^2)$, and the computational complexity of decoding of the hybrid model is higher [20].

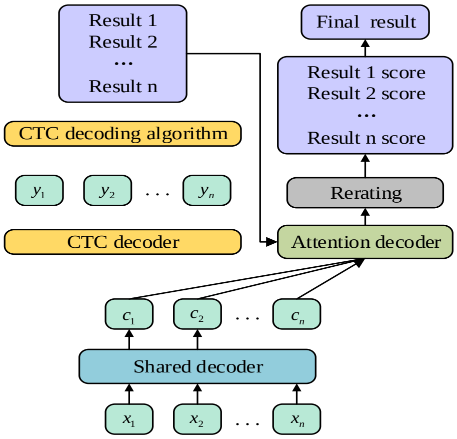


Fig. 3. Schematic diagram of CTC transformer decoding process.

For the long sequence speech recognition task, the study introduces the concept of Probability Sparse (Prob-Sparse). The AM used in the study is soft attention, as shown in Eq. (5). In Eq. (5), Q, K, V corresponds to query, key, and value, respectively; d denotes the sequence dimension;

$$A(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (5)$$

The query matrix Q has some sparsity, which will lead to more redundant computation if the attention of all queries of the query vector is computed. The study has calculated the attention score and distribution difference using Kullback-Leibler (KL) scatter, which measures the difference between the generated sample distribution and the target distribution. The calculation procedure is shown in Eq. (6). In Eq. (6), L denotes the length of the sequence; q_i, k_i, v_i denotes the i th line of Q, K, V; and p, U denotes the distribution of the attention scores and the uniform distribution, respectively.

$$KL(p \| U) = \ln \sum_{j=1}^{L} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L} \sum_{j=1}^{L} \frac{q_i k_j^T}{\sqrt{d}} - \ln L \quad (6)$$

The sparsity metric value $M_{Sparse}(q_i, K)$ of the query matrix Q is calculated from Eq. (6), see Eq. (7). The calculation process can be accelerated by keeping the larger $M_{Sparse}(q_i, K)$ queries. In Eq. (7), K denotes the random sampling of K; L denotes the number of samples.

$$M_{Sparse}\left(q_i,K\right)=\max_j\left\{\frac{q_i k_j^T}{\sqrt{d}}\right\}-\frac{1}{L}\sum_{j=1}^{L}\frac{q_i k_j^T}{\sqrt{d}} \tag{7}$$

L the calculation process is shown in Eq. (8), and r_{sample} denotes the sample sampling factor.

$$L=r_{sample}\ln L \tag{8}$$

In summary, the computation of Prob-Sparse Attention for the research design is shown in Eq. (9), where I_{Sparse} denotes the index of the L_{Sparse} query and $L_{Sparse}=r_{sample}L$.

$$\left(q_i,K,V\right)=\begin{cases}\sum_{j}^{L}p\left(k_j|q_i\right)v_j & if\ \ i\in I_{Sparse}\\ v_i & else\end{cases} \tag{9}$$

The CTC algorithm employs a dynamic programming algorithm to learn the mapping relationship between sequences, and uses maximum likelihood estimation to learn the probability of mapping paths, but the increase in the length of the input sequences is not conducive to the CTC algorithm to find feasible paths, and the blank labels introduced by the CTC will lead to the model easily falling into the local optimal situation. The conditional probability $p\left(l|X_{1:T}\right)$ for a given target sequence $X_{1:T}$ is computed in Eq. (10), where ϖ denotes the path of temporally concatenated observation labels; l denotes the true output; and B denotes the many-to-one mapping from π to l.

$$p\left(l|X_{1:T}\right)=\sum_{\pi\in B^{-1}(l)}p\left(\varpi|X_{1:T}\right) \tag{10}$$

The training process of CTC is constantly optimizing the CTC loss function $Loss_{ctc}$ to complete, but CTC exists a large number of ϖ, the optimization of the loss function will cause the model to produce poorly aligned output; and the error signal is positively correlated with π, the positive feedback of the error signal makes the probability prone to fall into a certain single path, which leads to the model overfitting phenomenon. In this regard, the study introduces the maximum conditional entropy to improve the CTC, and the network structure of the improved CTC-Attention decoding scheme is shown in Fig. 4.

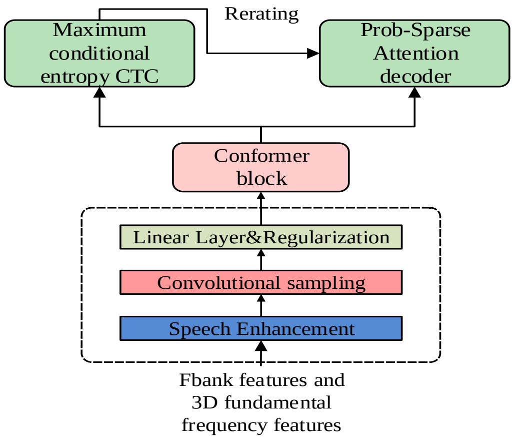

Rerating
Maximum conditional entropy CTC
Prob-Sparse Attention decoder
Conformer block
Linear Layer&Regularization
Convolutional sampling
Speech Enhancement
Fbank features and 3D fundamental frequency features

Fig. 4. Network structure diagram of improved CTC attention decoding scheme.

The CTC damage function based on maximum conditional entropy $Loss_{ctc}^{'}$ is calculated in Eq. (11), α denotes the coefficient of maximum conditional entropy regularization; $H\left(p\left(\pi|l,X\right)\right)$ denotes the entropy of feasible paths of input and target sequences. The maximum conditional entropy reduces the influence of the positive feedback of error information on the model training search, and the loss value calculated by Eq. (11) and the loss function value of the attention mechanism can be calculated according to Eq. (3) to obtain the loss value of the hybrid Improve CTC-Prob-Sparse Attention model.

$$Loss_{ctc}^{'}=Loss_{ctc}-\alpha H\left(p\left(\varpi|l,X\right)\right) \tag{11}$$

C. *CTC-AM Speech Recognition Framework Optimized by Fusion and Improved ABC Algorithm*

After the optimization of CTC loss function and AM, in order to realize the adaptive recognition of speech recognition model, the study introduces ABC algorithm to optimize the Improve CTC-Prob-Sparse Attention hybrid speech recognition model. ABC is a swarm intelligence global optimization algorithm that draws on the honey harvesting behavior of honeybee colonies. ABC algorithm divides honeybees into hiring bees, following bees and scout bees, and considers feasible solutions as food sources for bees. The hired bees search for new honey sources based on the old honey source information and determine whether to update the solution based on the evaluated objective function value. The following bee joins the honey source search process based on the information shared by the hiring bee; the scouting bee’s task is to randomly select a new honey source in the whole solution space to try when the hiring bee does not make further progress, and the algorithm flow is shown in Fig. 5. The ABC algorithm is chosen for the study to further improve the global search ability and adaptive capability of the hybrid speech recognition model.

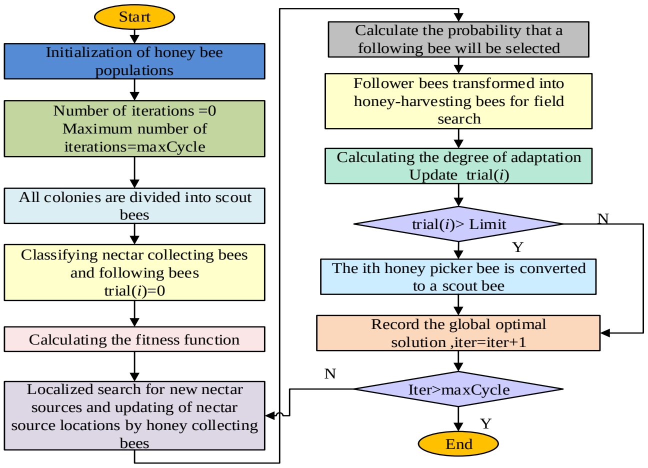

Start
Initialization of honey bee populations
Number of iterations =0 Maximum number of iterations=maxCycle
All colonies are divided into scout bees
Classifying nectar collecting bees and following bees trial(i)=0
Calculating the fitness function
Localized search for new nectar sources and updating of nectar source locations by honey collecting bees
Calculate the probability that a following bee will be selected
Follower bees transformed into honey-harvesting bees for field search
Calculating the degree of adaptation Update trial(i)
trial(i)> Limit
N
Y
The ith honey picker bee is converted to a scout bee
Record the global optimal solution ,iter=iter+1
Iter>maxCycle
N
Y
End

Fig. 5. Schematic diagram of ABC algorithm process.

The total number of bees is N_s, the population size of hired bees is N_e, the population of following bees is N_u, and the search space is defined as S. The number of honey sources is equal to the number of hired bees, the search feasible solution for the i th hired bee X_i^j is computed in Eq. (12), $X_{\max}^j, X_{\min}^j$ denotes the maximum and minimum values of the j dimensional components of the honey sources, $j \in \{1,2,..,D\}$, D denote the individual vector dimensions, respectively.

$$X_i^j = X_{\min}^j + rand\left(0,1\right)\left(X_{\max}^j - X_{\min}^j\right) \tag{12}$$

The benefit degree value of the honey source $fitness_i$ is calculated in Eq. (13), where f_i represents the objective function of the optimization problem.

$$fitness_i = \begin{cases} \dfrac{1}{1+f_i} \\ 1+abs\left(f_i\right) \end{cases} \tag{13}$$

The nectar source $fitness_i$ determines the probability that the following bee will be selected P_i, which is calculated as shown in Eq. (14).

$$P_i = \frac{f_i}{\sum_{n=1}^{N_e} f_n} \tag{14}$$

The search position generation calculation for hired bees is shown in Eq. (15), and the nectar source update occurs when the fitness value of the new search position is larger. However, when the position of the nectar source has not been updated, the hired bees of this nectar source are converted to scout bees, and the position updating process is the same as Eq. (12).

$$\begin{gathered} new\_X_i^j = X_i^j + rand\left[-1,1\right]\left(X_i^j - X_k^j\right) \\ j \in \{1,2,...,D\} \quad k \in \{1,2,...,N_e\} \end{gathered} \tag{15}$$

However, the ABC algorithm still has some application defects, ABC has a good global search ability in the hiring bee, following bee, and scouting bee stages, but the local search ability in different stages is weakened. In this regard, the study introduces the Adaptive Double Search (ADS) strategy, which can improve the convergence of the ABC algorithm, and the improved ADSABC workflow is shown in Fig. 6.

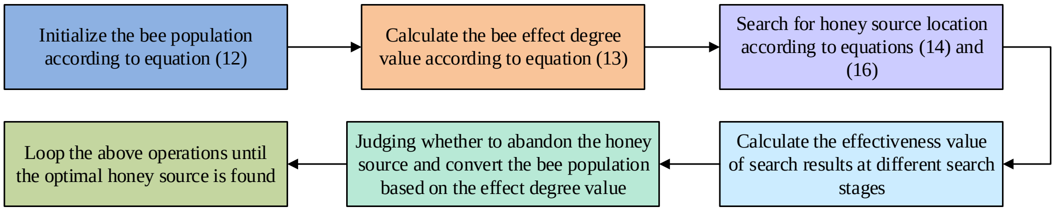

Initialize the bee population according to equation (12)
Calculate the bee effect degree value according to equation (13)
Search for honey source location according to equations (14) and (16)
Calculate the effectiveness value of search results at different search stages
Judging whether to abandon the honey source and convert the bee population based on the effect degree value
Loop the above operations until the optimal honey source is found

Fig. 6. Improved ADSABC workflow.

The search process of ADS is shown in Eq. (16), in which $x_{i,j}$, $x_{r,j}$ and $v_{i,j}$ denote the positions before and after the search of the hired bees following the bees, respectively; $x_{i,Global}$ denotes the global optimal nectar position under the current number of iterations; $x_{best,j}$ denotes the j elements of the global optimal position; α and μ denote the neighborhood search coefficients; and γ denotes the Kersey's variability factor, which helps the following bees to jump out of the local optimum. $\gamma = \tan\left[\left(\xi - 0.5\right)\pi\right]$ The following are used as the random numbers: ξ is the random number, $\xi \in [0,1]$; $\psi\left(t\right)$ is the adaptive adjustment factor, t is the number of cycles, and n is the number of iterations.

$$\begin{cases} v_{i,j} = x_{i,j} + \alpha^{n+1} rand\left(x_{best,j} - x_{i,j}\right) + \left(1-\psi\left(t\right)\right)\beta^{n+1}\left(x_{i,Global} - x_{i,j}\right) \\ v_{i,j} = x_{r,j} + \mu^{n+1} rand\left(x_{best,j} - x_{i,j} + \gamma\right) + \left(1-\psi\left(t\right)\right)\eta^{n+1}\left(x_{r,Global} - x_{r,j}\right) \\ \psi\left(t\right) = 1 - rand^{\left(1-{t}/{MaxCycle}\right)^2} \end{cases} \tag{16}$$

IV. Performance and Application Analysis of Speech Recognition Technology Based on Multimodal Information

In order to test the effectiveness of the research-designed speech recognition algorithm in multimodal human-computer interaction systems, the research designed two parts of performance test experiments, namely, the performance test of the improved ABC algorithm and the performance and application effect analysis experiments of the Improve CTC-Prob-Sparse Attention speech recognition framework.

A. Performance Test of Improved Swarm Algorithm

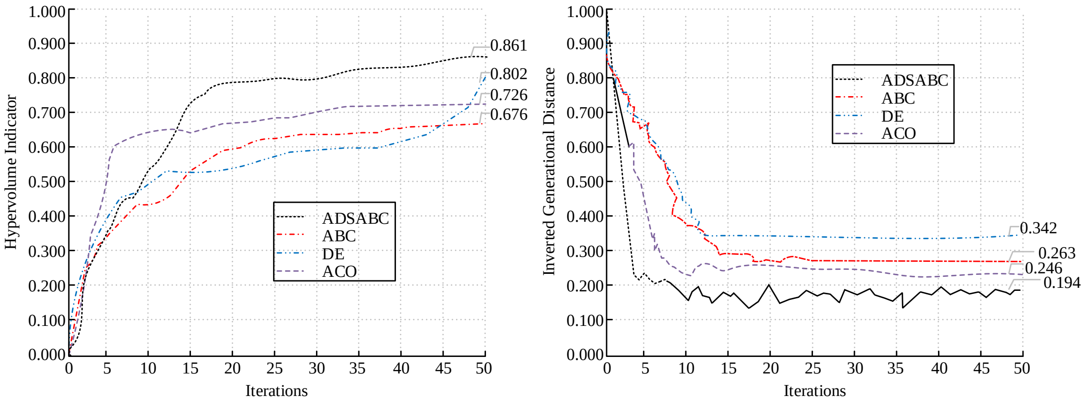
The performance of the ADSABC algorithm designed for the study is first analyzed by selecting the traditional ABC algorithm, Differential Evolution Algorithm (DE) and Ant Colony Optimization (ACO) algorithms for comparison, which are all based on Java language implementation. The performance and convergence of the optimization algorithms are analyzed by choosing the single-peak functions: The Sphere function, the Schwefel function, and the multi-peak functions, the Rastrigin function, the Griewank function, and the Ackley function.

Forty independent optimization experiments were set up to evaluate the convergence of different optimization algorithms from four angles, and the statistical results of the experiments are shown in Table I. Table I shows that the ADSABC algorithm has the best convergence performance for solving different test functions. The optimization value of the ADSABC algorithm for solving single-peak and multi-peak functions is smaller than that of the other algorithms, and the optimal solution can be found in all 40 independent experiments with a convergence rate of 100%. The ADSABC algorithm has the smallest number of solution iterations and the fastest convergence speed on average.

Hypervolume Indicator (HV) was selected to be associated with the Inverted Generational Distance Inverted Generational Distance (IGD) are chosen as the evaluation indexes of the algorithm, and the experimental results are shown in Fig. 7. HV is used to measure the size of the solution set occupied by the algorithm in the target space, and the larger the HV, the better the diversity and uniformity of the solution set is, and the IGD mainly focuses on the distance between the algorithm-generated solution set and the real optimal solution set, and the smaller the IGD, the better the performance of the algorithm. The smaller the IGD index, the better the performance of the algorithm. IGD and HV can jointly evaluate the iterative process and search effect of the algorithm, and the experimental results are shown in Fig. 7. As it can be seen in Fig. 7(a), the HV index of ADSABC algorithm takes the largest value, converges to 0.861 with the increase of iteration number, and the diversity and uniformity of the solution set are good. As seen in Fig. 7(b), the IGD curve of the ADSABC algorithm converges around the minimum value of 0.194, which is significantly different from other algorithms. It can be seen that the solution performance of the improved ABC algorithm of the study is good.

TABLE I. COMPARISON OF CONVERGENCE PERFORMANCE OF DIFFERENT ALGORITHMS

Test function	Algorithm	Average convergence value	Frequency of convergence	Minimum number of iterations	Average time (s)
Sphere	ABC	1.53E-6	29	461	2.91
	ADSABC	2.826E-16	40	215	1.94
	DE	1.947E-9	16	367	3.64
	ACO	3.672E-11	27	406	4.09
Schwefel	ABC	6.462E-10	21	403	3.08
	ADSABC	1.723E-13	40	216	2.16
	DE	9.012E-10	29	310	4.61
	ACO	9.306E-8	31	403	5.06
Rastrigin	ABC	5.77E-6	31	343	3.26
	ADSABC	9.283E-11	40	271	2.54
	DE	8.735E-9	29	302	4.16
	ACO	1.374E-8	36	425	6.17
Griewank	ABC	1.565E-7	35	461	11.66
	ADSABC	2.853E-12	40	329	9.54
	DE	5.563E-8	29	464	11.65
	ACO	2.618E-9	31	506	10.36
Ackley	ABC	7.042E-6	26	349	9.54
	ADSABC	3.983E-12	40	321	8.32
	DE	3.786E-10	30	406	11.58
	ACO	2.853E-9	25	496	11.22



(a) HV

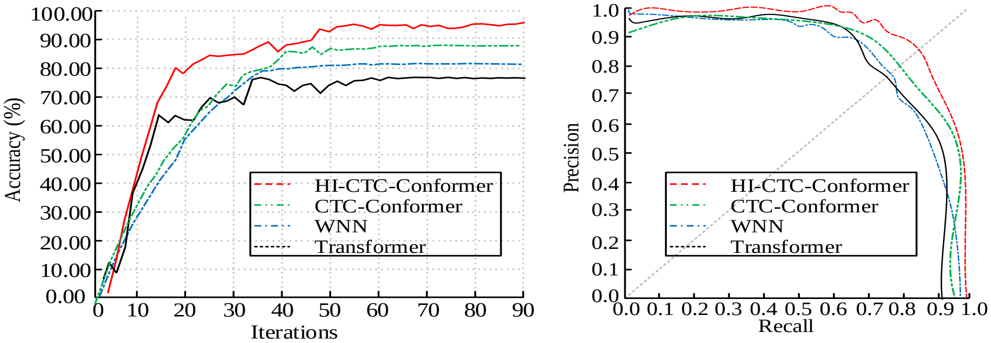
(b) IGD

Fig. 7. Comparison of HV and IGD for different optimization algorithms.

B. *Hybrid Improved Speech Recognition Model Performance Testing and Application Analysis*

Based on Windows 7 operating system, hardware environment Intel Core i7, Intel Q270 series chipset, memory is 32 GB, hard disk space is 8 TB capacity, based on Python implementation programming. The acoustic features input to the model are 80-dimensional Fbank features combined with 3-dimensional fundamental frequency features. The hybrid improved model (HI-CTC-Conformer) designed in the study is compared and analyzed with the CTC-Conformer model before improvement, Wavelet Neural Network (WNN), and Transformer model.

LibriSpeech, Fisher, Mozilla, VoxPopuli, and AN4 datasets are selected as the experimental dataset, and the experimental analyzed data is selected to be divided into training and test sets in the ratio of 9:1. The accuracy and precision-recall curve (PR) results of different speech recognition models are shown in Fig. 8. As seen in Fig. 8(a), the HI-CTC-Conformer model designed for the study has the highest recognition accuracy curve, with a maximum accuracy of 94.54%; compared to the other models, the HI-CTC-Conformer model is more accurate in its recognition results. As can be seen in Fig. 8(b), the PR curve of HI-CTC-Conformer model is in the upper rightmost part of the coordinate axis, and the Average Precision (AP), which represents the area of the PR curve, takes the largest value, and the recall of HI-CTC-Conformer model can reach 0.84 when the precision rate is 90%. In the same experimental environment, the accuracy and PR curve meticulously and comprehensively validate the overall excellent performance of the HI-CTC-Conformer model.



(a) Accuracy

(b) PR

Fig. 8. Comparison of accuracy and PR curve of different recognition models.

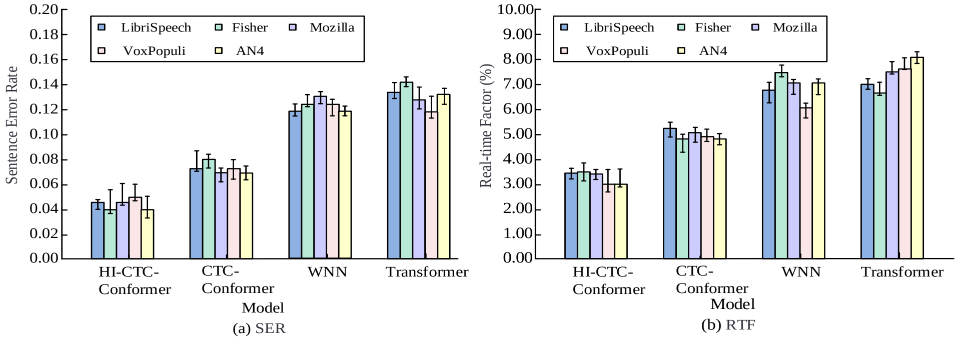
TABLE II. RECOGNITION ERROR RATES OF DIFFERENT RECOGNITION MODELS

Model	Index	Index	LibriSpeech	Fisher	Mozilla	VoxPopuli	AN4
HI-CTC-Conformer	Test	WER	0.041	0.036	0.046	0.044	0.036
		CER	0.049	0.047	0.037	0.054	0.047
	Training	WER	0.043	0.043	0.044	0.043	0.039
		CER	0.041	0.046	0.044	0.054	0.054
CTC-Conformer	Test	WER	0.066	0.063	0.064	0.061	0.068
		CER	0.065	0.060	0.067	0.064	0.065
	Training	WER	0.063	0.066	0.074	0.071	0.063
		CER	0.054	0.067	0.065	0.075	0.075
WNN	Test	WER	0.073	0.069	0.068	0.073	0.064
		CER	0.060	0.073	0.060	0.074	0.078
	Training	WER	0.073	0.084	0.085	0.081	0.072
		CER	0.082	0.074	0.084	0.078	0.073
Transformer	Test	WER	0.086	0.083	0.084	0.087	0.088
		CER	0.091	0.097	0.086	0.088	0.091
	Training	WER	0.131	0.103	0.094	0.105	0.108
		CER	0.144	0.123	0.114	0.095	0.091

Word Error Rate (WER) and Character Error Rate (CER) of speech recognition are selected as the evaluation indexes of different models, and the experimental results are shown in Table II. As seen in Table II, the WER and CER of the HI-CTC-Conformer model are significantly lower than those of the other three models on different datasets, with the lowest WER of 0.036 and the lowest CER of 0.037. In contrast, the WER of the Transformer model reaches the highest of 0.131 and the CER reaches the highest of 0.144. The WER and CER are calculated as the edit distance between the recognition results output by the system and the standard reference text, which can indicate the degree of inaccuracy of the speech recognition model. The difference between the two is that CER is more sample than WER, and the editing operations at character level are more fine-grained than those at word level. It can be seen that the model designed in the study has a more significant improvement in recognition accuracy compared to various types of baseline models.

The statistics of the Sentence Error Rate (SER) and Real-time Factor (RTF) metrics of the model are shown in Fig. 9. As can be seen in Fig. 9(a), on different datasets, the SER of the HI-CTC-Conformer model designed for the study takes the lowest level, and the median level of the SER is under 0.05, while the SER of the other three models takes the value above 0.06, and the highest value reaches 0.14. The SER denotes the editing distance between the sentence outputted by the system and the reference text, and it can be seen that the overall accuracy of the HI-CTC-Conformer model is at a high level. As seen in Fig. 9(b), the RTF of the HI-CTC-Conformer model is also at the lowest level, with values taken from different datasets under 4.00%. The RTF measures the decoding speed of the speech recognition model as the ratio of the recognition time to the speech duration, and is used to evaluate the real-time performance of the system. Comprehensively, it can be seen that the HI-CTC-Conformer model has achieved a good balance between RTF and SER metrics, and the system can have both better real-time performance and recognition accuracy.

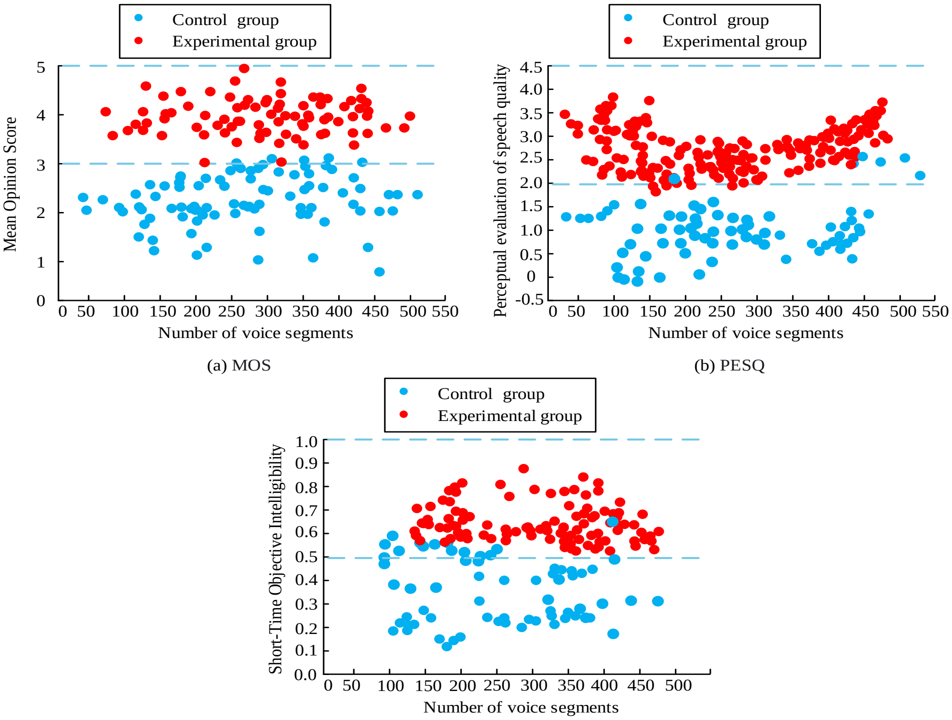
Finally, several subjects were recruited to analyze the application effect of the research-designed speech recognition model based on the multimodal human-computer interaction system with the subjective evaluation index Mean Opinion Score (MOS), the objective measurement index Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI), and the experimental results are shown in Fig. 10. Quality (PESQ) and Short-Time Objective Intelligibility (STOI) to analyze the application effect, and the experimental results are shown in Fig. 10. As can be seen in Fig. 10, the difference between the experimental group and the control group scores of the recovered scoring results is obvious, and the MOS scores of the design results of the study are higher than 3, indicating that the overall quality of the method is available and high. The PESQ and STOI scores are in the range of 2-4.5 and 0.5-1.0, respectively, and the scores are at a high level, and the quality of the speech recognition is high.



(a) SER

(b) RTF

Fig. 9. Comparison of SER and RTF metrics for different speech recognition models.



(a) MOS

(b) PESQ

(c) STOI

Fig. 10. Application effect analysis of speech recognition model.

V. Discussion

MMHCI is a technology that has gradually emerged with the advancement of AI technology, using multiple perceptual modalities such as vision, sound, and touch to enhance and optimize the human-computer interaction experience. Compared with single-modal interaction, MMHCI is more natural, efficient, and intelligent, and has been widely used in the fields of smart home, intelligent healthcare, and intelligent transportation. Through the fusion of multimodal information, the HCI system can obtain more precise semantic information, which improves the accuracy and robustness of the system. Speech recognition technology is an important part of MMHCI, which can convert speech signals into text or commands to realize efficient communication between humans and computers. Meanwhile, speech recognition can be combined with other modalities to form a more comprehensive and intelligent interaction. In the process of summarizing the existing research work, it was found that researchers in study [6], study [7], study [8], and study [9] mainly used neural network and deep learning techniques to construct speech recognition technology, and their optimal performance achieved a vocabulary sentence recognition accuracy of 89.15%, a word error rate of 10.56%, and a character error rate of 3.61%. However, this type of research lacks targeted improvement analysis of speech recognition techniques in specific tasks. In addition, Mavropoulos et al. also applied speech recognition technology to MMHCI [12], but did not carry out performance improvement and adaptation optimization studies for speech recognition technology. The application of existing speech recognition techniques to MMHCI still faces performance and application challenges.

In this regard, the study designed a basic speech recognition framework based on the improved CTC algorithm and attention mechanism, and introduced the ABC algorithm in the intelligent optimization algorithm to improve and optimize the framework in depth. The research results show that the method outperforms other speech recognition models in terms of accuracy, precision, and recall evaluation metrics, with a maximum accuracy of 94.54%. The minimum values of error rates at the character, word, and sentence levels are 0.037, 0.036, and 0.035, respectively, ensuring higher recognition accuracy while weighing the real-time rate. The performance metrics take a significant advantage over existing work. The design of the study improves the accuracy and robustness of MMHCI speech recognition and achieves innovative and adaptive optimization of the algorithm. By improving the CTC algorithm, AM, it provides a new framework for speech recognition model, which provides a valuable reference and reference for subsequent research. Meanwhile, the introduction of intelligent optimization algorithms into the field of speech recognition improves the performance of the speech recognition model and assists the model in adapting to the complex and changing application scenarios in real interaction.

Therefore, in practical applications, the speech recognition framework can provide users with a more natural and smooth interaction experience, which significantly improves user satisfaction and convenience. At the same time, the algorithmic innovation allows MMHCI's speech recognition technology to be applied to more complex scenarios, which promotes cross-fertilization of multimodal information processing, deep learning, optimization algorithms, and other fields.

In future research work, researchers can further explore how to more effectively fuse information from different modalities, and utilize different modal information to achieve the complementarity and association of multiple information. Moreover, the fusion of more techniques can be further attempted to improve the efficiency and performance of the model. In this way, MMHCI can realize wider application and popularization.

VI. Conclusion

With the rapid development of intelligent technology, speech recognition technology has been increasingly used in human-computer interaction. In order to enhance the application effect of multimodal interaction system, the study designed a speech recognition framework based on improved CTC algorithm and attention mechanism.

The experimental results show that, the improved ABC algorithm has better convergence performance on different test functions, and the convergence value, convergence number and convergence rate are better than other algorithms. Meanwhile, the diversity and homogeneity of the solution set are better, with the maximum HV of 0.861 and the minimum IGD of 0.194. The maximum accuracy of the HI-CTC-Conformer model is 94.54%, the area of the precision vs. recall curve is the largest, and the model's recall is up to 0.84 when the precision rate is 90%. Compared with the baseline model, the recognition accuracy and efficiency of this model are significantly improved because of the low recognition error rate of "word", "character" and "sentence". In the multimodal interaction system, the MOS scores, PESQ and STOI scores of the experimental group are in the range of 3-5, 2-4.5, and 0.5-1.0, respectively, and the application effect is superior.

The study improved the accuracy and comprehension of speech recognition and provided a more natural and convenient interaction experience. However, the study did not involve speech synthesis and expression analysis for multimodal interaction systems, which can be a future research direction for multimodal information analysis in the field of human-computer interaction.

Acknowledgment

This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2024C01108.

References

[1] J. Zhang, S. Wang, W. He, J. Li, Z. Cao, and B. Wei, "Projected augmented reality assembly assistance system supporting multi-modal interaction," Int J Adv Manuf Tech, vol. 123, no. 3, pp. 1353-1367, November 2022.

[2] E. Y. Oh, and D. Song, "Developmental research on an interactive application for language speaking practice using speech recognition technology," Etr&D-Educ Tech Res, vol. 69, no. 2, pp. 861-884, April 2021.

[3] A. Moin, F. Aadil, Z. Ali, and D. Kang, "motion recognition framework using multiple modalities for an effective human‑computer interaction," J Supercomput, vol. 79, no. 8, pp. 9320-9349, May 2023.

[4] A. S. Dhanjal, and W. Singh, “A comprehensive survey on automatic speech recognition using neural networks,” Multimed Tools Appl, vol. 83, no. 8, pp. 23367-23412, March 2024.

[5] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, and G. W. Burr, “An analog-AI chip for energy-efficient speech recognition and transcription,” Nature, vol. 620, no. 7975, pp. 768-775, August 2023.

[6] A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, “Automatic speech recognition method based on deep learning approaches for Uzbek language,” Sensors-Basel, vol. 22, no. 10, pp. 3683-3705, May 2022.

[7] S. Dua, S. S. Kumar, Y. Albagory, R. Ramalingam, A. Dumka, R. Singh, and A. S. AlGhamdi, “Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network,” Appl Sci-Basel, vol. 12, no. 12, pp. 6223-6235, June 2022.

[8] I. Świetlicka, W. Kuniszyk-Jóźkowiak,and M. Świetlicki, “Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network,” Sensors-Basel, vol. 22, no. 1, pp. 321-336, January 2022.

[9] S. Reza, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, “A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model,” Expert Syst Appl, vol. 215, no. 4, pp. 119293-119304, April 2023.

[10] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, “Deep learning for intelligent human‑computer interaction,” Appl Sci-Basel, vol. 12, no. 22, pp. 11457-11484, November 2022.

[11] S. Pan, “Design of intelligent robot control system based on human‑computer interaction,” Int J Syst Assur Eng, vol. 14, no. 2, pp. 558-567, April 2023.

[12] T. Mavropoulos, S. Symeonidis, A. Tsanousa, P. Giannakeris, M. Rousi, E. Kamateri, and I. Kompatsiaris, “Smart integration of sensors, computer vision and knowledge representation for intelligent monitoring and verbal human-computer interaction,” J Intell Inf Syst, vol. 57, no. 2, pp. 321-345, June 2023.

[13] R. Liu, Q. Liu, H. Zhu, H. and M. Cao, “Multistage deep transfer learning for EmIoT-Enabled Human‑Computer interaction,” IEEE Internet Things, vol. 9, no. 16, pp. 15128-15137, August 2022.

[14] D. Ran, W. Yingli, and Q. Haoxin, “Artificial intelligence speech recognition model for correcting spoken English teaching,” J Intell Fuzzy Syst, vol. 40, no. 2, pp. 3513-3524, February 2021.

[15] Y. Wu, and J. Li, “Multi-modal emotion identification fusing facial expression and EEG,” Multimed Tools Appl, vol. 82, no. 7, pp. 10901-10919, September 2023.

[16] U. Maniscalco, P. Storniolo, and A. Messina, “Bidirectional multi-modal signs of checking human-robot engagement and interaction,” Int J Soc Robot, vol. 14, no. 5, pp. 1295-1309, April 2022.

[17] L. Jia, X. Zhou, and C. Xue, “Non-trajectory-based gesture recognition in human-computer interaction based on hand skeleton data,” Multimed Tools Appl, vol. 81, no. 15, pp. 20509-20539, March 2022.

[18] G. Doras, Y. Teytaut, and A. Roebel, “A linear memory CTC-based algorithm for text-to-voice alignment of very long audio recordings,” Appl Sci-Basel, vol. 13, no. 3, pp. 1854-1879, January 2023.

[19] P. Ma, S. Petridis, and M. Pantic, “Visual speech recognition for multiple languages in the wild,” Nat Mach Intell, vol. 4, no. 11, pp. 930-939, October 2022.

[20] P. Preethi, and H. R. Mamatha, “Region-based convolutional neural network for segmenting text in epigraphical images,” AIA, vol. 1, no. 2, pp. 119-127, January 2023.

Application of Fuzzy Decision Support System Based on GNN in Anomaly Detection and Incident Response Service of Intelligent Security

Tao Chen[1]*, Xiaoqian Wu[2]
School of Public Basics, Anhui Medical College, Hefei 230032, China[1]
School of Public Health and Health Management, Anhui Medical College, Hefei 230032, China[2]

***Abstract*—This paper introduces a fuzzy decision support system (FDSS) based on a graph neural network (GNN) for anomaly detection and intelligent security. The primary aim is to develop a robust system capable of accurately identifying anomalies and providing timely incident response services. GNNs are utilized to capture the complex relationships and features between nodes in graph data, learning the embedded representation of each node through information transfer and aggregation mechanisms, which encapsulate the structural information of the graph. The FDSS leverages these features to construct a fuzzy rule base and perform fuzzy inference, generating decision suggestions that enhance the system's adaptability and robustness in dealing with uncertain data. The challenges addressed include the need for efficient anomaly detection in large-scale surveillance networks, the requirement for fast response times during emergencies, and the necessity for scalable and adaptable systems. Experimental results demonstrate that the GNN-based FDSS surpasses other methods in terms of anomaly detection accuracy, incident response service efficiency, system processing capacity, and model generalization ability. Compared to traditional statistical methods, machine learning models, and deep learning models, the proposed system maintains high precision and recall rates, processes data more efficiently, and adapts well to new datasets.**

***Keywords*—GNN; fuzzy decision support system; intelligent security; anomaly detection; incident response service**

I. Introduction

In the information age of the 21st century, social public security has become one of the core elements of national governance and urban development. With the acceleration of urbanization process, intelligent security system as an important technical means to maintain social stability and order, its intelligent, automatic level of improvement is particularly critical. Traditional security systems rely mainly on manual monitoring and simple video analysis, which is not only time-consuming, but also difficult to effectively respond to large-scale and complex scenes. In recent years, with the rapid development of artificial intelligence technology, especially the rise of deep learning and graph neural networks (GNN), a new technical path has been provided for the intelligent upgrading of intelligent security systems [1].

As shown in Fig. 1, the intelligent security system is a comprehensive security solution integrating modern technologies such as artificial intelligence, Internet of Things and advanced image recognition technology. It can not only monitor and warn of potential threats in real-time, but also automatically analyze behaviors, identify individuals, and even predict security events through cameras, sensors, access control systems and other devices, with powerful data analysis and management platforms. From home to enterprise and public facilities, intelligent security provides a series of functions including video surveillance, intrusion alarm and access control, which significantly improves the efficiency and accuracy of security prevention, while reducing manpower dependence and realizing intelligent management and rapid response.

Anomaly detection, as one of the core functions of intelligent security, aims to identify and warn against abnormal behaviors or potential threats in real-time, such as intrusion, violence and so on. However, anomaly detection algorithms are often characterized by diversity, concealment and strong environmental dependence, which require high accuracy and robustness of anomaly detection algorithms. In addition, once an abnormality is found, how to quickly and accurately start the Incident Response Service mechanism to prevent the situation from deteriorating is also a key problem that the intelligent security system must solve [2].

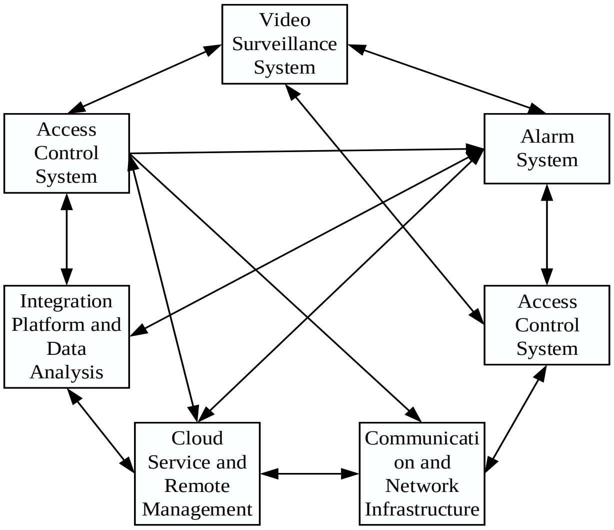

Video Surveillance System
Access Control System
Alarm System
Integration Platform and Data Analysis
Access Control System
Cloud Service and Remote Management
Communicati on and Network Infrastructure

Fig. 1. Framework diagram of intelligent security system.

*Corresponding Authors

In the field of intelligent security, significant progress has been made in the research of anomaly detection technology. Traditional statistical methods and machine learning models such as support vector machines and random forests are widely used, but these methods have obvious limitations when dealing with high-dimensional and unstructured data. In recent years, deep learning-based methods, especially convolutional neural networks (CNN) and recurrent neural networks (RNN), have demonstrated superior performance in image and video anomaly detection. However, these methods often ignore complex relationships between data, especially in large-scale surveillance networks, which are critical to accurately understanding the global situation [3].

This paper aims to fill the gaps mentioned above and proposes a fuzzy decision support system (FDSS) based on a graph neural network for anomaly detection and Incident Response Service in intelligent security. Specific research contents include: (1) According to the characteristics of intelligent security data, a graph neural network model suitable for large-scale surveillance networks is designed to extract space-time features and relationship features effectively. (2) Constructing a fuzzy rule base based on GNN output, using fuzzy logic to deal with uncertainty in monitoring data, improving robustness and adaptability of decision-making. (3) Design a set of Incident Response Service strategies linked with abnormal detection results to ensure a timely and effective start of the plan and reduce risks. (4) The performance of the proposed system in terms of anomaly detection accuracy, response time and resource consumption is verified by real data sets, and compared with existing methods [4].

This paper aims to fill the gaps mentioned above and proposes a fuzzy decision support system (FDSS) based on a graph neural network for anomaly detection and Incident Response Service in intelligent security. Specific research contents include the following. First, according to the characteristics of intelligent security data, a graph neural network model suitable for large-scale surveillance networks is designed to effectively extract spatiotemporal and relational features. Second, a fuzzy rule base is constructed based on GNN output, utilizing fuzzy logic to address uncertainty in monitoring data, thereby enhancing the robustness and adaptability of decision-making. Third, a set of Incident Response Service strategies linked with anomaly detection results is designed to ensure timely and effective initiation of plans and reduce risks. Finally, the performance of the proposed system in terms of anomaly detection accuracy, response time, and resource consumption is verified using real datasets and compared with existing methods.

In the remainder of this paper, we first describe the experimental environment and data set used in our study in Section II. We then detail the experimental design and methodology in Section III, outlining the steps involved in data preprocessing, model construction, training, and optimization. It discusses the specific data preprocessing steps taken to ensure the quality and efficiency of model training. The model training and optimization processes are elaborated in Section IV, including the strategies employed for learning rate adjustment and preventing overfitting. In Section V, we present the experimental results, comparing the performance of our GNN-based FDSS with other methods in terms of anomaly detection, incident response service efficiency, system processing ability, and model generalization. Finally, Section VI concludes the paper by summarizing the key findings and suggesting directions for future research.

II. Related Work

A. Neural Networks

At the forefront of intelligent security, which is related to public safety and urban management, Graph Neural Networks (GNN) are gradually showing their unique value and transformation potential. GNN not only revolutionizes the processing of multi-source information such as surveillance video and sensor data, but also promotes the depth and breadth of environmental understanding of security systems through its ability to operate directly on complex network structure data [5].

Recent research and application cases reveal how GNN opens up new possibilities in the field of intelligent security. On the one hand, GNN can effectively extract and integrate spatiotemporal features in video surveillance, and significantly enhance the accuracy and robustness of abnormal behavior recognition by learning complex relationship patterns between nodes, such as pedestrian behavior interaction and vehicle flow trends [6]. On the other hand, by constructing scene graphs and applying GNN, researchers successfully utilize spatial layout and dynamic interaction information between objects to improve the detection accuracy of abnormal events and maintain high performance even under complex and changeable environmental conditions [7]. GNN is also used to optimize resource allocation and event prediction for large-scale surveillance networks. By learning the correlation between monitoring points, GNN assists decision support systems in dynamically adjusting monitoring resources to ensure dense coverage of critical areas while reducing unnecessary waste of resources [8]. This methodological innovation not only strengthens the active defense capability of the security system, but also provides a more refined solution for smart city management.

It is worth noting that GNN fusion with traditional methods has also become a research hotspot, such as combining convolutional neural networks (CNN) and recurrent neural networks (RNN) to further improve the learning ability of spatiotemporal features, or integrating with fuzzy logic, reinforcement learning and other technologies to deal with more complex decision problems and dynamic response strategies [9].

Despite this, GNN applications in intelligent security are still in a rapid development stage, facing many challenges, such as efficient processing of large-scale graph data, interpretability of models, and generalization of cross-domain applications. Future research needs to continue to explore algorithm optimization, system integration, and deep integration with actual application scenarios to give full play to GNN's potential in intelligent security.

B. The Role of Fuzzy Decision Support System (FDSS) in Uncertainty Processing

Fuzzy Decision Support System (FDSS) plays an indispensable role in dealing with the uncertainty challenges inherent in intelligent security, and its influence is increasing day by day. FDSS's core strength lies in its ability to navigate situations that are ambiguous and difficult to quantify precisely, which is a common problem in the field of intelligent security, especially in the task of identifying abnormal behavior. By introducing fuzzy logic, FDSS can provide a flexible and powerful framework to adapt to and resolve complex and changeable security environments.

Marking an important milestone, they creatively integrated fuzzy logic with video surveillance systems to develop a system that efficiently identified fuzzy behavior patterns at the edges. This achievement significantly improves the response speed and recognition accuracy of the system to abnormal activities in complex scenarios, paving the way for fuzzy logic in the field of intelligent security applications [10].

By constructing and optimizing the fuzzy rule base carefully, they not only accelerate the decision-making process of Incident Response Service, but also greatly enhance the flexibility and adaptability of the decision-making mechanism [11]. This research proves that FDSS can make a reasonable judgment quickly according to the fuzzy rules set in advance when facing an emergency, effectively guide the implementation of emergency measures, reduce the decision-making delay, and fully reflect the broad prospects of fuzzy logic in improving the emergency response capability of the intelligent security system. FDSS has also demonstrated unique value in promoting transparency and interpretability in decision-making processes. It allows decision-makers to understand how the system handles uncertainty according to fuzzy rules, and can provide a reasonable decision-making basis even in the case of incomplete or conflicting information. In addition, FDSS enhances the comprehensiveness and reliability of decisions by integrating fuzzy information from different sources, such as multimodal sensor data, which is critical to building a robust intelligent security ecosystem.

C. Latest Development of Anomaly Detection and Incident Response Service Technology

In the field of intelligent security, the latest advances in anomaly detection and Incident Response Service technology reveal a profound transformation from traditional methods to intelligence and automation, especially under the catalysis of deep learning, which is undergoing an unprecedented innovation.

Anomaly detection technology in intelligent security systems has gradually moved from traditional statistical methods that rely on manual design features to automatic feature learning based on machine learning, and finally jumped to a new height of deep learning. Deep learning techniques, especially the introduction of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have revolutionized the identification of abnormal behavior in video surveillance. CNN, with its powerful ability in image recognition, can efficiently extract key visual features from video frames, while RNN captures dynamic behavior patterns in video sequences through its time series analysis ability. Together, the accuracy and efficiency of anomaly detection are significantly improved. Although deep learning has made remarkable achievements in anomaly detection, it still faces challenges in the face of increasingly complex and changeable monitoring environment, especially the processing of high-dimensional spatiotemporal data and the understanding of complex scene relationships. This requires higher-level model architectures such as graph neural networks (GNNs) and the integration of spatiotemporal attention mechanisms to better capture and understand the interactions between nodes and temporal dynamics in surveillance networks [12].

In terms of Incident Response Service, the focus has shifted from pure after-action to predictive maintenance and preparedness. Modern intelligent security systems aim to minimize damage by integrating predictive models, analyzing anomaly detection results in real time, and quickly formulating and activating the most appropriate response strategy [13]. This includes, but is not limited to, using machine learning algorithms to predict the likelihood and severity of abnormal events, dynamically adjusting response levels in conjunction with methods such as fuzzy logic or decision trees, and remotely scheduling resources through IoT technology for immediate intervention.

While these studies have made important progress in individual aspects of smart security, several key challenges and research gaps remain. Firstly, how to effectively integrate GNN's strong relationship learning ability and fuzzy logic's uncertainty processing advantage to construct an integrated system that can accurately identify anomalies and flexibly respond is an unexplored field. Second, existing methods tend to focus on specific types of anomaly detection and lack solutions that are widely applicable in complex and variable environments. In addition, system performance evaluation, especially resource consumption and response efficiency in real-world scenarios, also requires more attention.

To sum up, this study intends to design fuzzy decision support system based on GNN to make up for the shortcomings of anomaly detection and Incident Response Service in the current intelligent security field, and promote the development of intelligent security technology to a higher level by integrating spatiotemporal feature learning, fuzzy logic decision and efficient Incident Response Service mechanism [14].

III. Relevant Theoretical Basis

A. Graph Neural Network (GNN) Basics

Graph Neural Networks (GNNs) are advanced deep learning models designed to process graph data and capture complex structural information and relationship features between nodes in graphs. The core idea of GNN is to learn the embedded representations of each node through iterative propagation and aggregation of node features, which can contain the location information, neighborhood features and structural context of the node in the graph.

GAT is a computational model that exists in software implementations for working with graph data structures, and as such it is a virtual algorithm in computer science. It is not a physical entity, but a model of an algorithm implemented in a programming language and run on a computer.

The input to the GNN model is a graph G=(V,E), where V is the set of nodes and E is the set of edges. Each node is usually accompanied by a feature vector representing the initial feature information of the node. Edges can also carry eigenvectors that characterize relationships between nodes. The goal of GNN is to learn a mapping function f that maps nodes to a new feature space containing information about the graph structure. Where d and are the dimensions of the node and edge features, respectively, and are the dimensions of the output embedding [15].

GNN's working principle can be summarized as two core steps: information transfer and aggregation. Node characteristics are updated step by step through multi-layer iteration. In each iteration, each node generates a message vector according to its own characteristics and the characteristics of its neighbors through a message transfer function. This process can be expressed as: where is the message received by node v at the lth layer, is the neighbor node set of node v, and is the characteristic representation of node v at the previous layer. The received messages need to be aggregated to generate a new feature representation of the node. Commonly used aggregation functions are summation, average, maximum, etc. This process is described as in study [16].

In order to learn deeper graph structure features, GNN usually designs multilayer structures. With each additional layer, the process of information propagation and aggregation repeats over a wider neighborhood, allowing the model to capture structural information over greater distances. To introduce nonlinearity, nonlinear activation functions such as ReLU are often used after aggregation to enhance the expressiveness of the model.

The flexibility of GNN framework is reflected in the choice of message passing and aggregation functions, and different designs can cope with different types of graph data and task requirements. For example, Graph Convolutional Network (GCN) uses graph convolution as an aggregation function, and Graph Attention Network (GAT) introduces an attention mechanism to dynamically adjust the contribution weights of neighbor nodes.

To sum up, GNN gradually extracts high-level feature representations of nodes while retaining graph structure information through carefully designed information dissemination and aggregation mechanisms, providing a powerful tool for machine learning tasks on graph data [17].

B. Fuzzy Decision Support System (FDSS)

Fuzzy Decision Support System (FDSS) is a kind of decision support system based on fuzzy set theory, which can deal with fuzzy or uncertain problems. Fuzzy sets allow an element to have a real membership between 0 and 1, unlike traditional sets where elements either belong completely (membership 1) or do not belong at all (membership 0). Let the universe U be a nonempty set, and the fuzzy set A defined on the universe U can be described by membership functions of fuzzy sets, denoted by. For any element x in the domain of discourse, denotes the degree to which x belongs to fuzzy set A. The membership function quantifies the degree of membership of element x to fuzzy set A, and the closer its value is to 1, the higher the degree of belonging of x to A, and the closer it is to 0, the lower the degree of belonging [18].

These operations preserve the properties of fuzzy sets, i.e., the membership of elements to the set is continuous and can take any value between 0 and 1. In fuzzy decision support systems, fuzzy rules are often used to express decision logic. The general form of fuzzy rule is "if condition, then conclusion", where condition and conclusion are expressions of fuzzy set.

For example, a fuzzy rule might be written as follows:

"If the input is 'very hot'(high membership), the output is 'turn on the power air-conditioning'(also high membership)."

Fuzzy reasoning is the core part of fuzzy decision support system, Mamdani model or Takagi-Sugeno-Kang (TSK) model is usually used. Mamdani model transforms input fuzzy information into output fuzzy decision through fuzzification, inference, clipping and defuzzification. The TSK model combines fuzzy logic and multivariate regression analysis, using linear or nonlinear functions to map directly from input fuzzy sets to output real values.

Fuzzy decision support systems use these concepts and operations to deal with fuzzy or uncertain decision problems in the real world, such as expert systems, pattern recognition, control system design, etc. [19].

C. Anomaly Detection Theory

Anomaly detection in intelligent security system is the key technology to maintain public safety. It detects and warns abnormal behavior or event in time by analyzing video surveillance and sensor data, and then triggers Incident Response Service mechanism. Anomaly detection techniques can be divided into three categories: statistical methods, machine learning methods and deep learning methods. In the field of smart security, these methods are widely used to identify unusual patterns of activity. Statistical methods define boundaries of normal behavior based on statistical properties of the data, beyond which exceptions are considered. For example, a detection method based on Z-score. Where X is the observed value, is the mean value, and is the standard deviation. When| Z| is greater than a certain threshold, it is considered abnormal. This method is simple and intuitive, but it is weak when dealing with high-dimensional data and complex patterns. Support Vector Machines (SVM) maximize the spacing of normal data in anomaly detection by constructing a boundary, such as One-Class SVM [1].Where w is the normal vector of the classification hyperplane and is the slack variable that controls the proportion of outliers. ξ_i This method can deal with nonlinear problems well, but the cost of parameter selection and training is high. The specific workflow is shown in Fig. 2 [20, 21].

AutoEncoder (AE) and generative adversarial networks (GAN), identifies anomalies by learning representations of data [2].

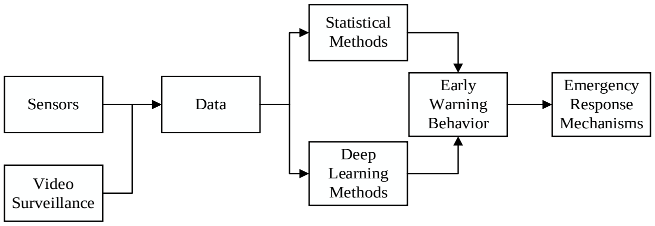


Fig. 2. Workflow.

Intelligent security system integrates a variety of sensors and video surveillance, anomaly detection applications in this field need to solve the real-time and accuracy problems in complex scenarios. In video surveillance, anomaly detection models based on deep learning, such as those based on 3D convolutional neural networks (3D-CNN) [3], are able to capture spatiotemporal dynamic features. Where y is the predictive label, X is the video segment, and X is the model parameter. θ Normal behavior is identified by training the model, and abnormal behavior is identified as a negative class by uncertainty or reconstruction error in the model output. Intelligent security systems often use multimodal data fusion, such as video and sound [4], to improve the robustness of anomaly detection. Where h is the fusion feature, v is the video feature, a is the audio feature, and a is the fusion function parameter. Multimodal fusion enhances adaptability to complex environmental changes.

Once an abnormality is detected, the intelligent security system shall immediately trigger an Incident Response Service, including but not limited to alarming, invoking resources, taking isolation measures, etc. Response strategy design needs to be combined with fuzzy logic or decision trees to achieve fast and effective action. $P(a \mid o) = \frac{1}{1 + exp(-w^T \cdot f(o))}$ where, is the probability of taking action a, given observation o, w is the weight vector, and f(o) is the function that converts the observation into a feature vector [22].

IV. Model Construction of Fuzzy Decision Support System based on GNN

A. System Architecture Design

This section will introduce the architecture design of fuzzy decision support system based on a graph neural network (GNN) in detail, including four key modules. Data preprocessing, GNN feature extraction, fuzzy rule base establishment, decision support and Incident Response Service.

Data preprocessing is the cornerstone of any data analytics model. In the GNN context, this step involves transforming the raw data into a graph structure, including definitions of nodes and edges, and possibly feature assignments. If the data is time series or sequence data, sliding window technology can be considered to extract time segments, each segment is defined as a node, and the dependency relationship between adjacent segments constitutes an edge [6]. Data normalization or normalization is also an important step in this phase to ensure stability of GNN training process, as shown in Eq. (1) [22].

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

GNN learns node characteristics on the graph through message passing mechanism, and for each node v_i, its characteristic representation $h_i^{(l)}$ is iteratively updated to at level 1, as shown in Eq. (2) [23].

$$h_i^{(l)} = \sigma\left(W^{(l)} h_i^{(l-1)} + \sum_{j \in \mathrm{N}_i} A_{ij} W^{(l)} h_j^{(l-1)} \right) \tag{2}$$

where is the activation function, is the inter-layer weight matrix, is the neighbor node set of a node, is the element of the adjacency matrix, and reflects the relationship strength between nodes. Based on the features extracted from GNN, a fuzzy rule base is constructed to support the subsequent fuzzy inference. Each rule can be formalized as: IF (Feature 1 is fuzzy set A) AND (Feature 2 is fuzzy set B) THEN (Decision is fuzzy set C) for example, the rule "If traffic flow is high and crowd density is high, there is a risk of congestion" can be converted to a fuzzy rule. The membership function of a fuzzy set, such as a triangular or Gaussian distribution, quantifies the degree to which an eigenvalue belongs to a particular fuzzy set [24].

Based on GNN features and fuzzy rule base, fuzzy inference is performed to generate decision suggestions. Fuzzy reasoning usually includes three steps: fuzzification, reasoning and defuzzification. In the reasoning process, the principle of maximum membership degree is applied to select the most consistent decision, and its formula is shown in Eq. (3).

$$u_c = \frac{\prod_{i=1}^{n} u_i^{w_i}}{\sum_{c=1}^{C} \prod_{i=1}^{n} u_i^{w_i}} \tag{3}$$

where, is the total membership of decision, is the membership of the ith feature under the corresponding decision, and $\(w_i\)$ is the weight of the feature, reflecting its importance in the decision [25, 26].

B. Algorithm Design and Implementation Optimization

In order to improve the generalization ability of the model and the ability to capture complex relationships, we will deeply customize the GNN model and introduce advanced graph learning components. For example, GraphSAGE model [7] is used for node feature aggregation, which realizes efficient graph feature learning by sampling neighbor nodes and aggregating their features. The formula can be expressed as, and its formula is shown in Eq. (4).

$$h_i^{(l+1)} = \sigma\left(W^{(l)} \cdot CONCAT\left(h_i^{(l)}, \mathrm{AGGREGATE}\left(\{ h_j^{(l)} \mid j \in \mathrm{N}_i \} \right) \right) \right) \tag{4}$$

Among them, the function is the aggregation operation on the features of neighboring nodes, such as average pooling, maximum pooling, etc. CONCAT represents the feature splicing operation to enrich the representation information of nodes [27].

In order to make GNN learning process more suitable for fuzzy decision requirements, we propose an integration strategy that embeds fuzzy logic directly into GNN training cycles. Specifically, in the reverse propagation process, the learning of features conforming to preset fuzzy rules is enhanced by adaptively adjusting the weight of the loss function, and the formula is shown in Eq. (5).

$$L_{\text{integrated}} = L_{\text{GNN}} + \alpha \cdot L_{\text{fuzziness}} \tag{5}$$

Here, is the standard GNN loss, which quantifies the consistency of the learned features with the fuzzy rule set, and is a dynamic tuning factor that adjusts automatically based on training progress and model performance.

In terms of anomaly scoring, we will use reinforcement learning methods [8] to dynamically adjust threshold settings to suit the anomaly sensitivity requirements of different scenarios. Specifically, the thresholding problem is modeled as a Markov Decision Process (MDP), where state s contains the current anomaly score distribution, action a is the direction and magnitude of the threshold adjustment, and reward r reflects the adjusted system performance improvement. Through interaction with environment, threshold strategy is optimized continuously to achieve optimal anomaly detection effect. The formula is given in Eq. (6) [28].

$$R_t = \sum_{k=t}^{t+T} \gamma^{k-t} r_k \tag{6}$$

Where is the discounted future reward starting at time t, the discount factor, and T is the number of periods the reward is considering?

Through the above-mentioned deeply customized GNN model, advanced fuzzy inference integration strategy, and dynamic threshold adjustment method, the algorithm design proposed in this section not only greatly enhances the professionalism and practicality of the model, but also improves the robustness and adaptive ability of the system in complex decision environments, providing solid technical support for the actual deployment of fuzzy decision support systems.

V. Experimental evaluation

A. Experimental Environment and Data Set Introduction

This chapter details the infrastructure configuration of the experiment and the characteristics of the data set used, laying a solid foundation for subsequent experimental design. The experimental environment is built on an advanced cloud computing platform equipped with NVIDIA Tesla V100 GPUs and equipped with high-speed network interconnection to ensure efficient data transmission and parallel computing capabilities. The memory configuration is 256GB, enough to handle the immediate processing needs of large-scale data sets. For data sets, the widely recognized MNIST handwritten digit data set and CIFAR-10 image classification data set were selected [29].

B. Experimental Design

This section provides an in-depth explanation of the overall architecture and core strategy of the experiment. The experimental design follows the modularization principle and is divided into four main stages: data preprocessing, model construction, training and evaluation. Among them, the model construction part uses convolutional neural networks (CNN) in deep learning, specifically LeNet-5 model for MNIST dataset, and more complex ResNet-18 model for CIFAR-10 dataset, aiming to explore the relationship between model performance and data complexity through different network depths and structural complexity. Eq. (7) shows the forward propagation calculation process of a general convolutional layer, where f represents the filter, x is the input feature map, * represents the convolution operation, and F is a nonlinear activation function, such as ReLU as shown in Eq. (7) [30].

$$y = \mathrm{F}\ (f * x + b) \tag{7}$$

C. Data Preprocessing Steps

Data preprocessing is a key step to ensure the quality and efficiency of model training, which mainly includes data cleaning, standardization, enhancement and division. Data cleaning removes invalid or mislabeled samples to ensure the purity of the data set. Normalization scales the input data to the same range, typically a distribution with a mean of 0 and a standard deviation of 1, using Eq. (8) to improve model convergence speed and stability.

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{8}$$

Data enhancement increases sample diversity through rotation, inversion, clipping, etc., reduces overfitting risk and enhances generalization ability of models. Finally, the data is randomly divided into training, validation and test sets, typically 70%, 15%, 15% to ensure fairness of model evaluation on independent test sets [31].

D. Model Training and Optimization

During the model training phase, we employ a stochastic gradient descent (SGD) optimizer and introduce momentum terms to accelerate convergence and reduce oscillations, as shown in Eq. (9), where is the learning rate, is the current parameter, is the gradient, and is the momentum cumulative variable.

$$v_t = \beta v_{t-1} + (1-\beta) g_t \tag{9}$$

$$\theta_t = \theta_{t-1} - \eta v_t \tag{10}$$

At the same time, learning rate decay strategy and early stopping method are applied to dynamically adjust learning rate and prevent overfitting. Model evaluation uses cross-entropy loss function combined with precision, recall and other evaluation indicators to ensure the performance of the model on classification tasks. The model training and optimization process is shown in Fig. 3.

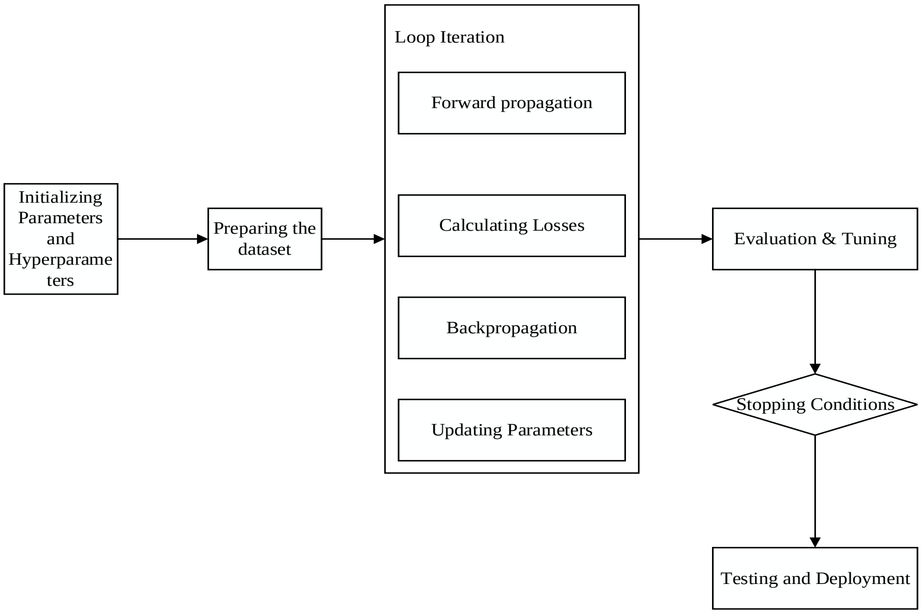


Fig. 3. Model training and optimization.

E. Experimental Results

Table I shows the performance evaluation metrics of different anomaly detection methods on selected data sets. The GNN-based fuzzy decision support system performed best in terms of F1 score and AUC, with 0.90 and 0.95, respectively, higher than other methods.

Table II summarizes the average, shortest and longest response times of GNN-based fuzzy decision support systems under different emergency scenarios. For example, in emergency evacuation drills, the average response time was 3.5 seconds, the shortest response time was 2.8 seconds, and the longest response time was 4.2 seconds. These data show that the fuzzy decision support system based on GNN can quickly start Incident Response Service in practical application, and the average response time is lower than the industry standard. This may be attributed to GNN's efficient computing power and ability to handle exceptions quickly.

Table III compares the GNN-based system with two other systems (Systems A and B) for different load pressures. Under high load conditions, GNN-based systems can handle 120 events/hour, while systems A and B can handle only 50 and 60 events/hour, respectively. This indicates that GNN-based systems have higher processing efficiency and stability, and can effectively cope with a large number of events. This may be due to GNN's ability to process complex data relationships quickly, thus improving the processing power of the system.

TABLE I. ANOMALY DETECTION PERFORMANCE EVALUATION

Method	Precision	Recall	F1 score	AUC
Fuzzy Decision Support System Based on GNN	0.92	0.88	0.90	0.95
traditional statistical methods	0.85	0.9	0.867	0.92
Machine Learning Models (Isolation Forest)	0.88	0.82	0.85	0.91
Deep Learning Model (Autoencoder AE)	0.9	0.86	0.88	0.93

TABLE II. INCIDENT RESPONSE SERVICE TIME STATISTICS

Scene	Average Response Time (sec)	Minimum Response Time (Seconds)	Maximum Response Time (Seconds)
emergency evacuation drill	3.5	2.8	4.2
fire warning	4.1	3.7	4.6
Traffic accident response	3.2	2.9	3.8

TABLE III. SYSTEM EFFICIENCY TEST (UNIT: EVENTS / HOUR)

Load Pressure	System A	System B	Fuzzy Decision Support System Based on GNN
low	120	150	180
in	80	100	160
high	50	60	120

TABLE IV. EFFICIENCY COMPARISON WITH OTHER METHODS

Method	Training Time (Hours)	Detection Time (Ms/Event)	Overall Efficiency Score (1-10)
Fuzzy Decision Support System Based on GNN	20	30	8.5
traditional statistical methods	-	10	7
machine learning model	15	25	6.5
deep learning models	40	50	5

Table IV assesses the differences in training time and detection efficiency between the different methods, as well as a composite efficiency score. The GNN-based system had a longer training time of 20 hours, but a detection time of only 30 ms/event, with an overall efficiency score of 8.5. This shows that GNN-based systems have advantages in long-term operation, which can quickly and accurately identify abnormal situations and provide timely support for decision makers. This may be because GNN is able to efficiently learn and capture features in the data, thereby improving detection speed and accuracy.

Fig. 4 shows how well each model generalizes across different datasets. The performance of fuzzy decision support system based on GNN on dataset A, dataset B and dataset C is 0.92, 0.87 and 0.93 respectively, which is better than other methods.

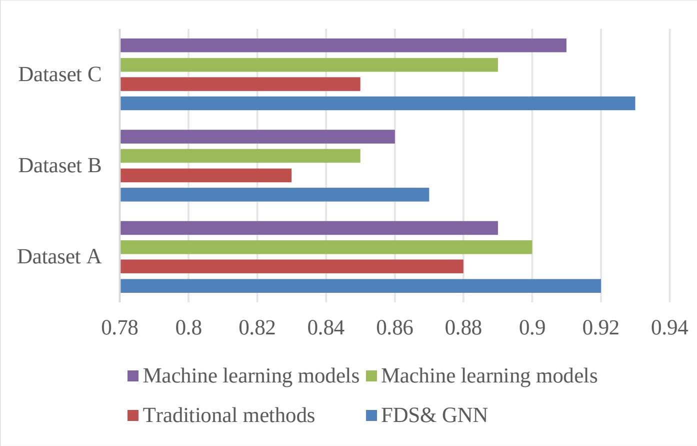


Fig. 4. Comparison of the generalization ability of anomaly detection models.

Experimental results show that the fuzzy decision support system based on GNN outperforms other methods in anomaly detection performance, Incident Response Service efficiency, system processing ability and model generalization ability. The system maintains high precision and recall and has strong adaptability to new data. It can identify anomalies quickly and accurately, and provide timely and reliable support for decision-making. Overall, fuzzy decision support system based on GNN is an efficient and stable anomaly detection solution.

F. Discussion

The experimental results presented in the previous sections demonstrate the effectiveness of the fuzzy decision support system (FDSS) based on Graph Neural Networks (GNNs) in multiple aspects. However, a deeper examination of the findings provides valuable insights into the system's strengths and potential areas for improvement.

Firstly, the superior performance of the GNN-based FDSS in terms of F1 score and AUC, as shown in Table I, suggests that the system is highly adept at balancing precision and recall, making it particularly suitable for anomaly detection tasks where both false positives and false negatives need to be minimized. The ability of GNNs to capture the complex relationships within graph data contributes to this enhanced performance.

Secondly, the Incident Response Service times recorded in Table II indicate that the system not only identifies anomalies accurately but also responds promptly. The average response time of 3.5 seconds for emergency evacuation drills, for instance, highlights the system's capability to initiate actions swiftly, which is crucial in emergency scenarios. This responsiveness can be attributed to the efficient computational framework of the GNN-based system.

Thirdly, the system's processing efficiency, as outlined in Table III, shows that it can handle a significantly higher number of events per hour compared to Systems A and B, especially under high load conditions. This robustness and scalability are critical for real-world applications where the volume of incoming data can fluctuate widely.

Moreover, the efficiency comparison in Table IV reveals that despite a longer training period, the GNN-based system achieves faster detection times and a higher overall efficiency score. This implies that the initial investment in training time pays off in the form of quicker and more accurate detections, which is beneficial for operational efficiency.

Finally, the generalization ability of the system, as depicted in Fig. 4, indicates that the FDSS based on GNN can maintain high performance across different datasets. This adaptability is essential for deploying the system in diverse environments where data characteristics may vary.

While the results are promising, there are potential areas for further investigation. Future work could focus on optimizing the training phase to reduce the initial time required, exploring hybrid models that combine the strengths of GNNs with other techniques, and conducting more extensive testing on varied datasets to further validate the system's generalization capabilities.

In summary, the experimental results underscore the robustness and efficiency of the GNN-based FDSS, positioning it as a powerful tool for anomaly detection and intelligent security applications. Its ability to handle large volumes of data, respond quickly, and generalize well makes it a valuable addition to the field of anomaly detection systems.

VI. Conclusion

In this study, we have proposed a fuzzy decision support system (FDSS) based on Graph Neural Networks (GNNs) for anomaly detection and intelligent security applications. Extensive experimental evaluations demonstrate the superior performance of our system compared to traditional statistical methods, machine learning models, and deep learning models. Our results show that the GNN-based FDSS achieved the highest F1 score and AUC on the selected datasets, highlighting its effectiveness in accurately identifying anomalies. Furthermore, the system demonstrated consistently fast response times in various emergency scenarios, underscoring its capability to initiate incident response services promptly and effectively. In terms of system processing efficiency, the GNN-based system managed a significantly higher number of events per hour under high load conditions, outperforming alternative systems. Evaluations also revealed that despite a longer training period, the GNN-based system achieved rapid detection times and a high overall efficiency score. Additionally, the system exhibited strong generalization ability across different datasets, demonstrating robustness and adaptability. These results confirm the reliability and efficiency of the GNN-based FDSS, making it a viable solution for anomaly detection in complex decision-making environments.

In terms of future work, efforts will focus on several key areas to further enhance the capabilities of the fuzzy decision support system (FDSS) based on Graph Neural Networks (GNNs). One direction involves optimizing the training process to reduce the initial time required, potentially through the use of more advanced optimization algorithms or distributed computing frameworks. Another area of interest is the development of hybrid models that integrate the strengths of GNNs with other machine learning techniques, such as reinforcement learning, to improve the system's adaptability and decision-making capabilities. Additionally, there is a need for more extensive testing across a broader range of datasets and real-world scenarios to further validate the system's generalization and robustness. Lastly, exploring the integration of user feedback mechanisms could help refine the fuzzy rule base, making the system even more responsive to evolving security threats and user-specific requirements. These enhancements aim to solidify the position of the GNN-based FDSS as a leading solution in anomaly detection and intelligent security applications.

References

[1] F. Louati, F. B. Ktata, and I. Amous, “An Intelligent Security System Using Enhanced Anomaly-Based Detection Scheme,” Computer Journal, vol. 2024, p. 14, January 2024.

[2] S. B. Han, Q. H. Wu, and Y. Yang, “Machine learning for Internet of things anomaly detection under low-quality data,” International Journal of Distributed Sensor Networks, vol. 18, no. 10, p. 13, October 2022.

[3] A. Alharbi, A. H. Seh, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, “Analyzing the Impact of Cyber Security Related Attributes for Intrusion Detection Systems,” Sustainability, vol. 13, no. 22, p. 19, November 2021.

[4] J. Duan, “Deep learning anomaly detection in AI-powered intelligent power distribution systems,” Frontiers in Energy Research, vol. 12, p. 17, March 2024.

[5] S. M. Nagarajan, G. G. Deverajan, A. K. Bashir, R. P. Mahapatra, and M. S. Al-Numay, “IADF-CPS: Intelligent Anomaly Detection Framework towards Cyber Physical Systems,” Computer Communications, vol. 188, pp. 81–9, September 2022.

[6] V. Moshkin, D. Kurilo, and N. Yarushkina, “Integration of Fuzzy Ontologies and Neural Networks in the Detection of Time Series Anomalies,” Mathematics, vol. 11, no. 5, p. 13, May 2023.

[7] J. H. Jeong, H. H. Jung, Y. H. Choi, S. H. Park, and M. S. Kim, “Intelligent Complementary Multi-Modal Fusion for Anomaly Surveillance and Security System,” Sensors, vol. 23, no. 22, p. 16, November 2023.

[8] L. Cui, Y. Y. Qu, G. Xie, D. Z. Zeng, R. D. Li, S. G. Shen, and S. Yu, “Security and Privacy-Enhanced Federated Learning for Anomaly Detection in IoT Infrastructures,” IEEE Transactions on Industrial Informatics, vol. 18, no. 5, pp. 3492–500, May 2022.

[9] R. Sarno, F. Sinaga, and K. R. Sungkono, “Anomaly detection in business processes using process mining and fuzzy association rule learning,” Journal of Big Data, vol. 7, no. 1, p. 19, February 2020.

[10] R. Afzal and R. K. Murugesan, “Rule-Based Anomaly Detection Model with Stateful Correlation Enhancing Mobile Network Security,” Intelligent Automation and Soft Computing, vol. 31, no. 3, pp. 1825–41, March 2022.

[11] M. N. Gao, L. F. Wu, Q. Li, and W. Chen, “Anomaly traffic detection in IoT security using graph neural networks,” Journal of Information Security and Applications, vol. 76, p. 10, March 2023.

[12] L. Y. Qi, Y. H. Yang, X. K. Zhou, W. Rafique, and J. H. Ma, “Fast Anomaly Identification Based on Multiaspect Data Streams for Intelligent Intrusion Detection Toward Secure Industry 4.0,” IEEE Transactions on Industrial Informatics, vol. 18, no. 9, pp. 6503–11, September 2022.

[13] S. H. Almotiri, “Integrated Fuzzy Based Computational Mechanism for the Selection of Effective Malicious Traffic Detection Approach,” IEEE Access, vol. 9, pp. 10751–64, February 2021.

[14] D. Ge, Y. H. Cheng, S. S. Cao, Y. M. Ma, and Y. W. Wu, “An enhanced abnormal information expression spatiotemporal model for anomaly detection in multivariate time-series,” Complex & Intelligent Systems, vol. 10, no. 2, pp. 2937–50, February 2024.

[15] H. Wang, Y. Y. Zhang, Y. J. Liu, Y. J. Liu, F. L. Liu, H. Y. Zhang, and B. Xing, “ASAD: Adaptive Seasonality Anomaly Detection Algorithm under Intricate KPI Profiles,” Applied Sciences-Basel, vol. 12, no. 12, p. 18, June 2022.

[16] S. Y. LU, K. Wang, Y. L. Wei, H. R. Liu, Q. L. Fan, and B. L. Wang, “GNN-based Advanced Feature Integration for ICS Anomaly Detection,” ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 6, p. 32, June 2023.

[17] M. Semerci, A. T. Cemgil, and B. Sankur, “An intelligent cyber security system against DDoS attacks in SIP networks,” Computer Networks, vol. 136, pp. 137–54, August 2018.

[18] N. Berjab, H. H. Le, and H. Yokota, “Recovering Missing Data via Top-k Repeated Patterns for Fuzzy-Based Abnormal Node Detection in Sensor Networks,” IEEE Access, vol. 10, pp. 61046–64, July 2022.

[19] T. Qin, B. Wang, R. Y. Chen, Z. Y. Qin, and L. Wang, “IMLADS: Intelligent Maintenance and Lightweight Anomaly Detection System for Internet of Things,” Sensors, vol. 19, no. 4, p. 19, April 2019.

[20] Z. Sun, Q. K. Peng, X. Mou, Y. Wang, and T. Han, “An artificial intelligence-based real-time monitoring framework for time series,” Journal of Intelligent & Fuzzy Systems, vol. 40, no. 6, pp. 10401–15, June 2021.

[21] M. Ahmed, “Intelligent Big Data Summarization for Rare Anomaly Detection,” IEEE Access, vol. 7, pp. 68669–77, July 2019.

[22] P. S. Kumar and L. Parthiban, “Scalable Anomaly Detection for Large-Scale Heterogeneous Data in Cloud Using Optimal Elliptic Curve Cryptography and Gaussian Kernel Fuzzy C-Means Clustering,” Journal of Circuits Systems and Computers, vol. 29, no. 5, p. 38, May 2020.

[23] M. Alanazi and A. Aljuhani, “Anomaly Detection for Internet of Things Cyberattacks,” CMC-Computers Materials & Continua, vol. 72, no. 1, pp. 261–79, January 2022.

[24] F. Al-Obeidat and E. S. M. El-Alfy, “Hybrid multicriteria fuzzy classification of network traffic patterns, anomalies, and protocols,” Personal and Ubiquitous Computing, vol. 23, no. 5-6, pp. 777–91, June 2019.

[25] K. D. Gupta, K. Singhal, D. K. Sharma, N. Sharma, and S. Malebary, “Fuzzy Controller-empowered Autoencoder Framework for anomaly detection in Cyber Physical Systems,” Computers & Electrical Engineering, vol. 108, p. 13, March 2023.

[26] G. Sharma, A. K. Kapil. Intrusion Detection and Prevention Framework Using Data Mining Techniques for Financial Sector. Acta Informatica Malaysia. vol. 5, no. 2, pp. 58-61. 2021.

[27] C. Wang, “IoT anomaly detection method in intelligent manufacturing industry based on trusted evaluation,” International Journal of Advanced Manufacturing Technology, vol. 107, no. 3-4, pp. 993–1005, February 2020.

[28] F. Abdullah and A. Jalal, “Semantic Segmentation Based Crowd Tracking and Anomaly Detection via Neuro-fuzzy Classifier in Smart Surveillance System,” Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 2173–90, February 2023.

[29] X. L. Wang, C. Fidge, G. Nourbakhsh, E. Foo, Z. Jadidi, and C. Li, “Anomaly Detection for Insider Attacks From Untrusted Intelligent Electronic Devices in Substation Automation Systems,” IEEE Access, vol. 10, pp. 6629–49, February 2022.

[30] K. Agrawal, T. Alladi, A. Agrawal, V. Chamola, and A. Benslimane, “NovelADS: A Novel Anomaly Detection System for Intra-Vehicular Networks,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 22596–606, November 2022.

[31] M. Masdari and H. Khezri, “Towards fuzzy anomaly detection-based security: a comprehensive review,” Fuzzy Optimization and Decision Making, vol. 20, no. 1, pp. 1–49, January 2021.

Using Combined Data Encryption and Trusted Network Methods to Improve the Network Security of the Internet of Things Applications

Yudan Zhao
Network and Information Center, Guangzhou Open University, Guangzhou, 510000, China

***Abstract*—With the integration of big data and artificial intelligence, the Internet of Things has rapidly developed as the foundation for collecting data. The data collected by Internet of Things devices is mostly sensitive information, but limited resources can easily lead to data leakage. Therefore, this study adopts a combination of data encryption and trusted networks to improve the network security of the Internet of Things. This study proposes an Internet of Things network security system based on an improved SM9 encryption algorithm and iTLS security protocol. The system uses a key generation center to generate and distribute keys to complete information encryption and decryption, and identity authentication is carried out through dynamic keys. The results indicated that the total time for key generation, encryption, and decryption based on the SM9-iTLS network security system was 3.63 seconds. The total time for key generation, signature, and signature verification in the system was 3.65 seconds, which is better than other Internet of Things network security systems, and it also had better network resource occupancy and latency than other systems. The Internet of Things network security system based on improved SM9-iTLS can not only improve the security of information transmission among Internet of Things devices but also optimize the efficiency of information transmission. The research results have a certain promoting effect on developing the Internet of Things information security field.**

Keywords—Security protocols; Encryption; Internet of Things; Resource constraints; SM9

I. Introduction

The Internet of Things (IoT) brings great convenience to people's daily lives, enterprise production and supply, etc. However, IoT devices usually use ordinary software and hardware configurations, so they are prone to exposing significant data and network security risks in open environments [1-3]. There are numerous IoT devices, but these devices have limited resources and insufficient network security protection capabilities, leading to a large number of network attacks and information leakage problems. Therefore, improving the network security protection capability of the IoT under limited resources is the main direction of research. The use of intrusion detection systems as a network security method can detect and prevent potential attacks, but with an increase in IoT devices, the difficulty of device maintenance and maintenance becomes greater, and the cost is higher [4-6]. Data encryption (DE) is a low-cost security solution that can prevent device data from being accessed or cracked by intruders, protecting the security of IoT devices when transmitting data [7-8]. Security protocols (SPs) in trusted networks can establish security mechanisms such as identity verification, key exchange, encrypted communication, auditing, and logging in network communication, effectively ensuring the security of network communication [9]. DE and SPs are widely used to ensure network security. Many scholars and experts have conducted relevant research on DE, SPs, and IoT network security (IoT-NS). Leigh proposed an encryption algorithm based on NIST-AES to address the issue of low security of confidential data in third-party HPC systems. The running time of using the NIST-AES algorithm was basically the same as that of using an ordinary encryption algorithm, and the data security of the HPC system has been improved by 34.8%. Therefore, the NIST-AES algorithm solved the data security problem of third-party HPC systems [10]. To address the security issues of voice data in centralized cloud storage, Zhang and Zhao proposed a distributed voice DE storage scheme based on IPFS and CP-HABE. This solution had reliability, security, and scalability. DE could solve data security issues in distributed communication [11]. To address the issue of IoT devices being vulnerable to attacks due to limited resources and being at the edge of the network, He et al. proposed an anonymous and lightweight authentication and key exchange protocol for IoT devices. This protocol had higher security than other SPs, and lightweight protocol deployment on IoT was feasible [12]. Mvah et al. proposed a self-executing SP based on Nash equilibrium to address the issue of attackers using artificial intelligence for ARP spoofing attacks. The simulation results showed that compared with other methods, this method could better prevent, detect, and recover ARP spoofing attacks. This protocol could solve the problem of ARP spoofing attacks and effectively ensure network security [13]. Rathee et al. designed a dynamic Pub/Sub communication mechanism to address communication trust and security issues during message publishing and subscription processes in environmental intelligence. This mechanism was more secure than traditional security measures [14]. Chowdhury et al. developed a network security tool based on device fingerprints to address the vulnerability of IoT devices to Mirai botnet and spoofing attacks in communication. The recognition accuracy of this scheme for the UNSW dataset was 99.81%, and the tool could enhance network security in heterogeneous network environments [15].

The above research indicates that in the case where traditional intrusion detection systems have application limitations, many scholars have conducted relevant research on DE, SPs, and IoT-NS. However, there is a paucity of research exploring the integration of DE and SPs for IoT-NS. Based on existing research, it is known that both DE algorithms and SPs have the effect of strengthening network security protection. Therefore, this study combines the two to construct a lightweight and high-security IoT-NS protection model. This study designs an improved SM9 encryption algorithm and iTLS SP for IoT network communication and constructs an IoT communication system with dual layer security protection of communication encryption and identity verification. The innovation of the research lies in the cancellation of certificate authentication for both encryption algorithms and SPs. The encryption algorithm uses a private key generator (PKG) to generate and distribute private keys and completes encryption and decryption through private key calculation. The SP completes the handshake protocol through dynamic password calculation. This improved encryption and identity authentication mechanism not only enhances network security but also increases communication speed and occupies fewer resources.

The research content mainly consists of five sections. Section I is the introduction, which analyzes the research achievements in IoT-NS and briefly describes the proposed IoT-NS protection model. Section II is based on the algorithm and mechanism of the improved SM9-iTLS IoT-NS system. Section III tests the research model. Section IV discusses the experimental results. Section V summarizes the research findings.

II. Methods and Materials

The SM9 encryption algorithm uses highly secure asymmetric encryption technology, but the certificate mechanism of SM9 is prone to vulnerabilities. This study proposes an SM9 algorithm for certificate free encryption and signature. In addition, a lightweight TLS communication protocol has been introduced and optimized, using dynamic keys in the handshake protocol to enhance the security of IoT information transmission through identity authentication. This study combines the improved SM9 encryption algorithm with iTLS SP to construct a secure and low-latency IoT network system [16].

A. Design of Improved SM9 Encryption Algorithm

The identity-based cryptography (IBC) system is a cryptographic technique that facilitates the deployment of asymmetric cryptographic systems. SM9 is a cryptographic algorithm in IBC. SM9 mainly includes encryption, key encapsulation, key exchange, and digital signature. The logic of SM9 utilizes the Abel group discrete logarithmic difficulty formed by the points of elliptic curves over a finite field to achieve encryption, decryption, and digital signature [17-18]. The cracking difficulty of the 3072-bit factorization algorithm is equivalent to that of the 256-bit elliptic curve algorithm, indicating that the SM9 encryption algorithm has sufficiently high security. The PKG of SM9 only generates decryption private keys and uses identity information as the public key, which is beneficial for forming lightweight network encryption. However, this form of encryption that completely entrusts the key to a third party is vulnerable to network attacks. Therefore, this study improves SM9 by designing a certificate-free encryption and signature-based SM9 encryption algorithm. In improving SM9, user public keys are no longer verified through certificates but are generated and distributed using a key generation center, which can improve the network security of IoT devices [19]. In the certificate free encryption SM9 algorithm, the encryption process is completed using two parts: a public key and user identity, while the user decryption operation is decrypted using the generated key. The PKG in improved SM9 calculates the decryption private key for the encryption task. The calculation formula for decrypting the private key is Eq. (1).

$$sk_{SM9} = s(Hash(ID \| hid) + s)^{-1} P_2 \tag{1}$$

In Eq. (1), sk_{SM9} is the decryption private key, and P_2 is the generator of the addition group G_2. PKG calculates and decrypts the private key, and the complete encryption and decryption process is shown in Fig. 1.

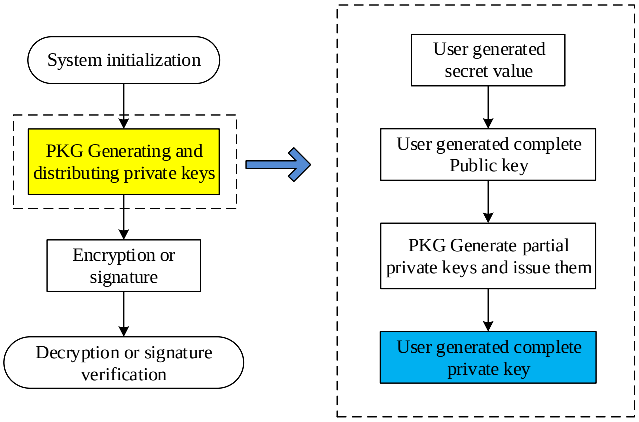

System initialization
PKG Generating and distributing private keys
Encryption or signature
Decryption or signature verification
User generated secret value
User generated complete Public key
PKG Generate partial private keys and issue them
User generated complete private key

Fig. 1. The process of PKG calculation and decryption of private keys and complete encryption / decryption.

In Fig. 1, the improved SM9 first performs system initialization, using multiplication group G_1, addition group G_2, and addition group G_2 of prime number q order, and then generates the distributed private key sk_{SM9} by PKG. The improved SM9 is encrypted based on the private key sk_{SM9}, and the encryptor completes the encryption through user ID and the public key calculated by ID. The user calculates the decrypted private key through the private key sk_{SM9} for decryption. The calculation formula for decrypting sk_{SM9} by the user is Eq. (2).

$$sk_{SM9-CLE} = x_{ID}^{-1} s(Hash(ID \| PK \| hid) + s)^{-1} P_2 \tag{2}$$

In Eq. (2), $sk_{SM9-CLE}$ is the decryption private key calculated by the user. P_2 is the generator of the additive group G_2. $Hash$ is the decryption private key generated by the user. x_{ID} is a random value. hid represents the user's key. ID represents the user's identity. s is the primary signature private key of PKG. PKG calculates the public key in the encrypted information based on the user's identity, and the calculation of the public key is Eq. (3).

$$PK = (x_{ID} P_1, x_{ID} P_{pub-e}) \tag{3}$$

In Eq. (3), PK is the public key, which is completely public, and P_1 is the generator of the additive group G_1. After receiving the public key, the sender encrypts the ciphertext based on the public key and user identity. After receiving the ciphertext, the user decrypts it through calculation to obtain the plaintext. The most important thing in decryption is to verify whether the packaging information of the encrypted calculation is consistent with the packaging information of the decrypted calculation. If they are not consistent, decryption cannot be performed. The expression to determine whether the encapsulated information is consistent is Eq. (4).

$$\begin{aligned} w' &= e(C_1, SK) \\ &= e(rx_{ID}(h_{ID}+s)P_1, x_{ID}{}^{-1}s(h_{ID}+s)^{-1}P_2) \\ &= e(P_1,P_2)^{r(h_{ID}+s)x_{ID}x_{ID}{}^{-1}}s(h_{ID}+s)^{-1} \\ &= e(P_1,P_2)^{rs} \\ &= g^r \\ &= w \end{aligned} \tag{4}$$

In Eq. (4), w and w' are the encapsulated information for encryption and decryption calculations. C_1 is ciphertext. The improved SM9 algorithm without certificate encryption has an encapsulation mechanism that can meet the IND-CCA security definition and avoid internal and external attacks in terms of security. In addition, when internal and external adversaries have attack advantages, the algorithm can use the advantages of the adversary to solve difficult problems, indicating that the improved SM9 with certificate free encryption has extremely high security. The data integrity and source verification in IoT data transmission are improved through certificate free signature algorithms, which maintain the computational speed of the SM9 encryption algorithm while ensuring the security of DE to better build lightweight IoT. The calculation of the private key of the certificate free signature algorithm is Eq. (5).

$$sk_{SM9} = s(Hash(ID \| hid)+s)^{-1}P_1 \tag{5}$$

In Eq. (5), PKG uses generator P_1 to calculate the private key. The formula for decrypting sk_{SM9} by the user is Eq. (6).

$$sk_{SM9-CLS} = x_{ID}^{-1}s(Hash(ID \| PK \| hid)+s)^{-1}P_1 \tag{6}$$

In Eq. (6), the user uses the generator P_1 to calculate the decryption private key. The signer signs by transmitting information and the decryption private key $sk_{SM9-CLE}$ calculated by the user. The verifier verifies the signature using the public key and user identity calculated by PKG, and the calculation expression of the public key is Eq. (7).

$$PK = (x_{CD}P_2, x_{ID}P_{pub-s}) \tag{7}$$

The most important thing in signature verification is to verify whether the bilinear pairing results are consistent. If they are the same, it means the signature verification is successful. The expression to determine whether the bilinear pairing results are consistent is Eq. (8).

$$\begin{aligned} w &= u*t \\ &= e(S,P)*g^h \\ &= e(P_1,P_2)^{\frac{l*x*s(h1+s)}{(h1+s)x}}*e(P_1,P_2)^h \\ &= e(P_1,P_2)^{l*s+h*s} \\ &= e(P_1,P_2)^{(l+h)s} \\ &= e(P_1,P_2)^{rs} \\ &= g^r \end{aligned} \tag{8}$$

In Eq. (8), the digital signatures are (h,S), $u=e(S,P)$, $t=g^h$, and $g=e(P_1,P_{pub-s})$. In terms of security, the improved SM9 without certificate signature adopts a step-by-step protocol to complete the verification. The flowchart for proving distribution conventions is shown in Fig. 2.

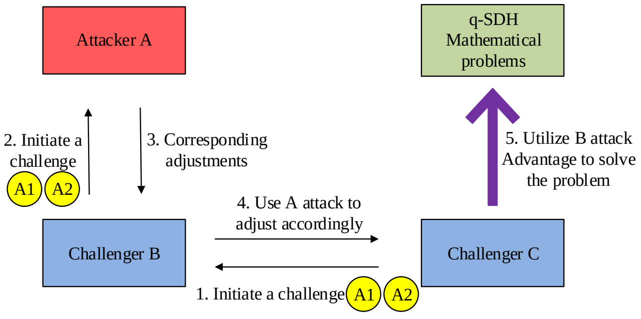

Attacker A
q-SDH Mathematical problems
2. Initiate a challenge
A1
A2
3. Corresponding adjustments
5. Utilize B attack Advantage to solve the problem
4. Use A attack to adjust accordingly
Challenger B
Challenger C
1. Initiate a challenge
A1
A2

Fig. 2. The process of proving distribution conventions.

In Fig. 2, the distribution specification sets an attacker A, as well as challenger B of A and challenger C of B. C launches an attack on B, B launches an attack on A, and A responds to B's attack. B utilizes A's response to challenge C, while C utilizes B's attack advantage to solve the q-SDH problem. The improved SM9 algorithm reduces the proof of unforgeability to solving the q-SDH problem to demonstrate the strong security of the algorithm. During the attack, the hash function constructed attack models for two attackers A1 and A2. Due to the forging signatures that can allow attackers to solve q-SDH mathematical problems, which are unsolvable, A1 cannot complete the attack. Similarly, forging signatures can solve the BDHI problem, which is also unsolvable, and A2's attack is also not feasible. Therefore, the improved SM9 without certificate signature is secure and reliable. The formula for forging signatures for A1 and A2 is Eq. (9).

$$\begin{cases} h = H_2(M \| w)S = (r-h)\dfrac{s}{x(h_*+s)}P_1 \\ h' = H_2'(M \| w)S' = (r-h')\dfrac{s}{x(h_*+s)}P_1 \end{cases} \tag{9}$$

In Eq. (9), x is a random unknown number. Challenger C in A2 calculates the solution to the BDHI problem as $e(P_1,P_2)^{1/x}$. The overall structure diagram of the improved certificate free encryption and signature SM9 algorithm is shown in Fig. 3.

In Fig. 3, the improved SM9 encryption algorithm key system structure includes five layers: business processing layer, password service layer, user key layer, PKM key layer, and data layer. This study improves the yellow section, including encryption and signature algorithms, key generation methods, and PKM calculation methods.

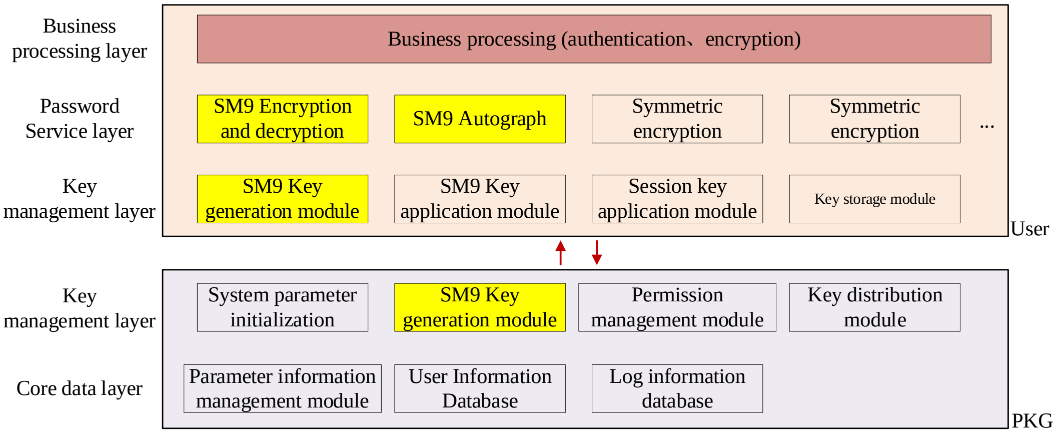

Business processing layer
Business processing (authentication、encryption)
Password Service layer
SM9 Encryption and decryption
SM9 Autograph
Symmetric encryption
Symmetric encryption
...
Key management layer
SM9 Key generation module
SM9 Key application module
Session key application module
Key storage module
User
Key management layer
System parameter initialization
SM9 Key generation module
Permission management module
Key distribution module
Core data layer
Parameter information management module
User Information Database
Log information database
PKG

Fig. 3. The overall structure of IoT SM9 key management system.

B. Construction of IoT-NS System Based on Improved SM9-iTLS

DE focuses on data protection, even if attackers obtain data, they cannot interpret the content of the data. The SP in a trusted network focuses on the rules of communication between devices, and communication can only be established based on these rules. This study integrates DE and SP to jointly build a more secure IoT communication network. IoT devices typically have low protection capabilities and are prone to interception and eavesdropping. This study introduces a lightweight TLS protocol to enhance network security [20]. The standard TLS protocol necessitates the authentication of certificates. In the case of a considerable number of IoT devices with limited computing capabilities, it requires a greater bandwidth, memory, and computational resources, which consequently increases network latency. Therefore, this study designs an improved iTLS protocol to ensure communication efficiency and further increase network security. The device establishes ecure communication throsugh a handshake protocol, and the handshake protocol of iTLS is shown in Fig. 4.

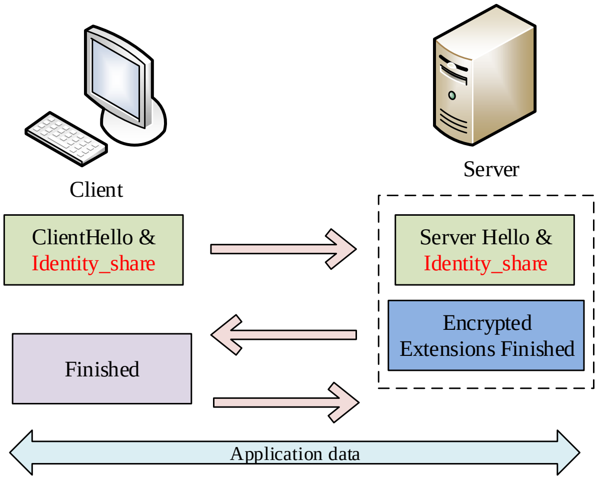


Fig. 4. Completed handshake protocol process of iTLS.

In Fig. 4, an Identity share is added between the client and server to perform key exchange, and a shared key is established through this exchange mechanism. The server extends through Identity share and can provide multiple key exchange parameters, from which the server selects a key for negotiation. This approach increases the flexibility and security of the protocol. In addition, the entire handshake process of iTLS does not require certificate authentication, reducing a significant amount of authentication time. The calculation of the shared key by the server is Eq. (10).

$$shared\_secret = yEK_C \,\|\, e(EK_C + H(ID_C), yP_{pub} + sk_S) \tag{10}$$

In Eq. (10), ID_C represents the identity of the client. EK_C is the temporary public key of the client. P is the generator. H is a hash function. sk_S is the private key generated by KGC on the server. y is the temporary private key of the server. The handshake key in the shared key is obtained using a key derivation function, and the formula for the key derivation function is Eq. (11).

$$handshake\_secret = \text{HKDF-Extract}(ek, shared\_secret) \tag{11}$$

In Eq. (11), HKDF is the key derived function. The client uses the server ID_S and key provided by ServerHello to calculate the shared key, and the expression for the shared key is Eq. (12).

$$shared\_secret = xEK_S \,\|\, e(xP_{pub} + sk_C, EK_S + H(ID_S)) \tag{12}$$

In Eq. (12), EK_S is the temporary public key of the server, and sk_C is the private key generated by KGC for the client. This study aims to enhance protocol compatibility and reduce latency. A compatibility design is adopted to enable iTLS to establish a connection with TLS1.3. The 0-RTT mode of TLS1.3 is introduced, and dynamic keys (IDEK) are used in the 0-RTT mode to improve security. The expression for calculating IDEK on the client side is Eq. (13).

$$IDEK_C = \text{HKDF-Extract}(0, e(sk_C, H(ID_S))) \quad (13)$$

In Eq. (13), the client uses server identities ID_S and sk_C to calculate $IDEK_C$. The formula for calculating IDEK on the server is Eq. (14).

$$IDEK_S = \text{HKDF-Extract}(0, e(H(ID_C), sk_S)) \quad (14)$$

Both parties perform a 0-RTT handshake using the generated dynamic key to complete data decryption. The compatibility design of ITLS's 0-RTT handshake is shown in Fig. 5.

In Fig. 5, the protocol enhances compatibility with TLS1.3 by adding key_share extension and handshake negotiation with TLS1.3. In the 0-RTT mode, the client has added the early_data extension to prove that it will carry an Application data *. The protocol calculates dynamic passwords to protect the security of 0-RTT data, and the server determines whether to perform a handshake based on the properties of bilinear pairs. The bidirectional pair expression is Eq. (15).

$$e(sk_C, H(ID_S)) = e(H(ID_C), H(ID_S))^s = e(H(ID_C), sk_S) \quad (15)$$

The server and client complete handshake negotiation through the early_data extension of Eq. (15), achieving secure connection of IoT devices. This study improves the SM9 algorithm and TLS protocol, integrating the encryption algorithm and SP to construct an IoT-NS system based on the improved SM9-iTLS, as shown in Fig. 6.

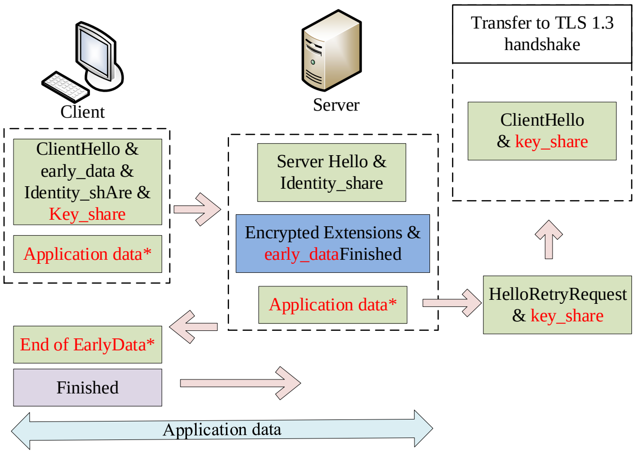

Client
Server
Transfer to TLS 1.3 handshake
ClientHello & early_data & Identity_shAre & Key_share
Application data*
Server Hello & Identity_share
Encrypted Extensions & early_dataFinished
Application data*
ClientHello & key_share
HelloRetryRequest & key_share
End of EarlyData*
Finished
Application data

Fig. 5. Compatibility design of ITLS's 0-RTT handshake mode.

In Fig. 6, the IoT device is connected to the device service enterprise and transmits communication data. The improved SM9 key system distributes keys to IoT devices and device service enterprises. During the communication process, IoT devices use SM9 keys to negotiate symmetric keys and then encrypt data based on the symmetric keys to decrypt information for device service enterprises. Using the iTLS protocol for device authentication in communication further enhances the security of information transmission. The SM9 key system includes a registration center and a key generation center PKG, mainly responsible for user registration and SM9 key generation and distribution. The IoT network system utilizes SM9 encryption technology and iTLS protocol to achieve identity authentication and encrypted data exchange.

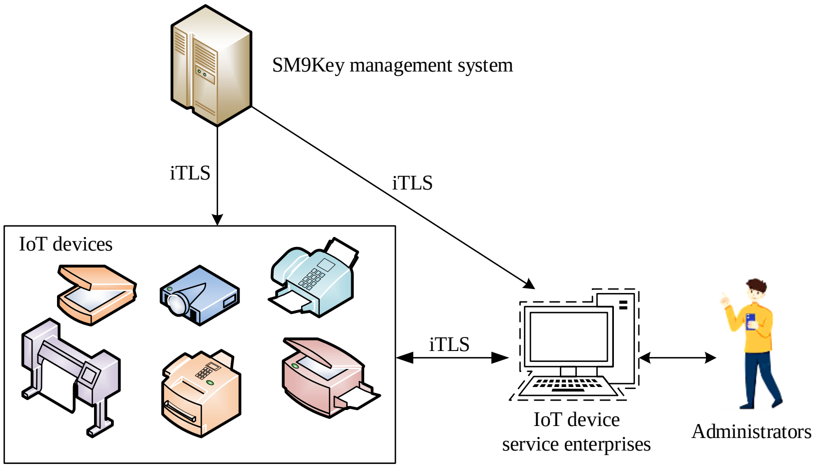


Fig. 6. IoT security management system based on improved SM9 encryption algorithm and iTLS SP.

III. Results

To verify the performance of the system, relevant experiments were conducted in this study. The experiment first conducted a comparative experiment on the improved SM9 encryption algorithm to test its computational efficiency. Then, comparative experiments were conducted on iTLS to test the network latency under iTLS SP. Finally, experimental analysis of the IoT-NS system based on improved SM9-iTLS showed that the proposed IoT-NS system not only enhanced security, but also could respond quickly under the improved encryption algorithm and SP, without causing IoT communication burden.

A. *Experimental Environment and Parameter Settings*

This study used the MIRACL library to test the improved SM9 encryption algorithm. The test code was written on ESP32, and the size of the encrypted information was selected as 120 bits. The elliptic curve of SM9 selected 256 bits to construct an addition group, with two addition groups of 64 and 128 bytes in length. 120 repeated tests were conducted to calculate the average running time of the algorithm. This study implemented SP iTLS experimental analysis in the WolfSSL library written in C language, with security levels of 112 and 128 bit. All experiments selected TLS_AES_128_CCM_SHA256 as the hash function and symmetric encryption cipher suite. The computer used in the experiment had 16GB of memory and an Intel i7 8700 processor. Table I shows the detailed parameter settings for each encryption algorithm in ESP32.

The experiment selected RSA3072, SM9, and improved SM9 for comparative testing, and TLS1.3 and iTLS with RSA, ECC, and 0-RTT authentication modes were used for comparative testing. A comparative experiment was conducted between the IoT-NS system based on ECC-iTLS, RSA-TLS, and ElGamal TLS and SM9 iTLS. The evaluation indicators used encryption, decryption, signature, and signature verification time as the computational efficiency evaluation indicators of encryption algorithms. Communication connection delay was used as an indicator of SP connection efficiency. Traffic overhead was utilized as an evaluation indicator for network resource occupancy.

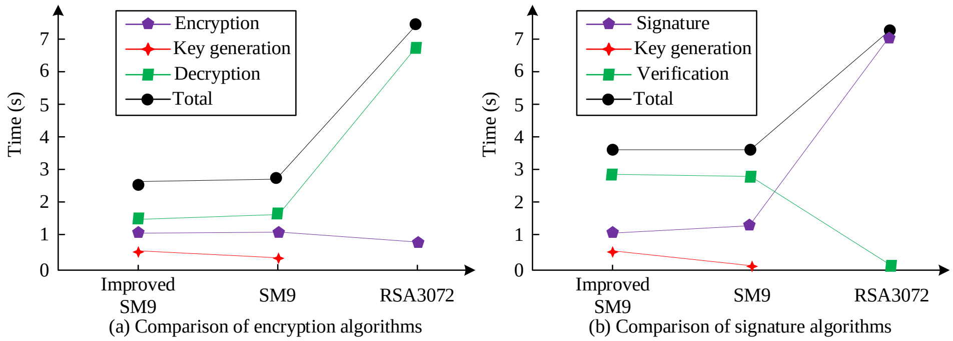
TABLE I. DETAILED INFORMATION ON PARAMETER SETTINGS FOR ALGORITHMS

/	Key escrow	Certificate	(Encryption algorithm) Public key	(Signature algorithm) Public key	(Encryption algorithm) Private key	(Signature algorithm) Private key
Improved SM9	No	No	ID+128	ID+256	128	64
RSA3072	No	Yes	384	384	384	384
SM9	Yes	No	ID	ID	128	64

B. Analysis of Improved SM9 Computing Efficiency and iTLS Communication Delay

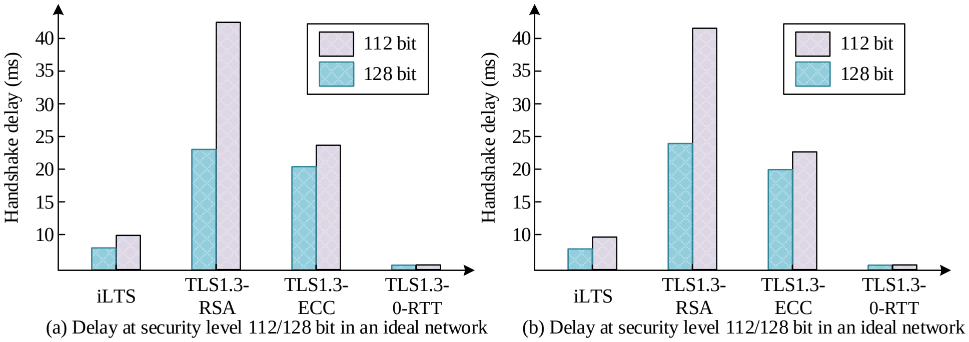
The experiment first compares the improved SM9 algorithm. The calculation time results for the complete encryption solution, encryption process, and complete signature and verification process are shown in Fig. 7.

In Fig. 7 (a), the key generation time of the research algorithm is 0.39s, the encryption time is 1.02s, the decryption time is 1.30s, and the total time is 2.71s. The encryption time of SM9 is lower than that of the improved SM9, but the decryption time and key time are longer, with a total time of 2.74s, which is higher than SM9. The key generation time of RSA3072 is lower than that of the research algorithm, but the decryption time is 6.78s, and the total calculation time is 7.31s. In Fig. 7 (b), the total time for key generation, signature, and signature verification of the research algorithm is 3.74s, and the total time for SM9 algorithm is 3.76s. The signature verification time of RSA3072 is the lowest, with a minimum value of 0.06s, but the key generation time takes 6.88s, resulting in a total time of 6.94s. The experiment sets up an ideal network with zero delay and no packet loss under wireless broadband, and tests the handshake delay of iTLS protocol and TLS1.3 protocol at security levels of 112-bit and 128-bit, as shown in Fig. 8.



(a) Comparison of encryption algorithms (b) Comparison of signature algorithms

Fig. 7. Comparison of encryption and signature verification times of algorithms in ESP32.



(a) Delay at security level 112/128 bit in an ideal network (b) Delay at security level 112/128 bit in an ideal network

Fig. 8. Full handshake delay for iTLS and TLS 1.3 under ideal network conditions.

In Fig. 8 (a), when the security levels are 112-bit and 128-bit, the TLS1.3 protocol delay in the 0-RTT mode is the lowest, with the lowest values of 1.4ms and 1.5ms. The delay of iTLS is 6.4ms and 9.9ms respectively, the delay of TLS1.3-RSA is 23.3ms and 43.9ms, and the delay of TLS1.3-ECC is 21.5ms and 23.7ms. Fig. 8 (b) shows the results of the second group of experiments. The delay of iLTS is 6.7ms at 112-bit security level and 9.7ms at 128-bit security level. The delay of iLTS is slightly increased compared to TLS1.3-0-RTT protocol, which is also improved by TLS, but it is still at a low latency level.

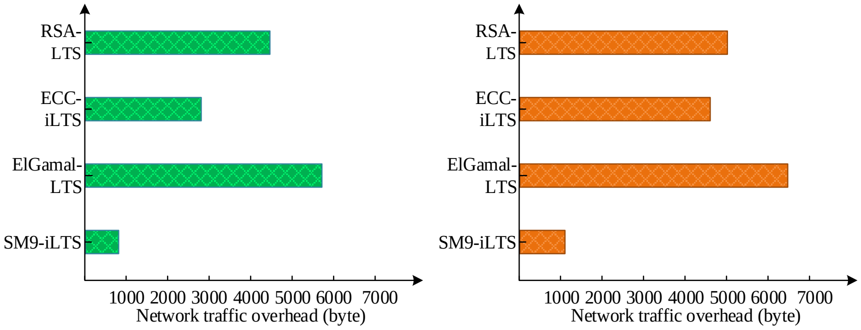
C. Performance Analysis of IoT-NS System Based on Improved SM9-iTLS

This study compares the improved IoT-NS system of SM9-iTLS with the aforementioned IoT security system. The network traffic overhead of each system under 112-bit and 128-bit security levels is shown in Fig. 9.

In Fig. 9 (a), when the security level is 112-bit, the network traffic overhead of the research system is the lowest, with a minimum value of 794 bytes, ElGamal-TLS of 5731 bytes, ECC-iTLS of 2788 bytes, and RSA-TLS of 4312 bytes. In Fig. 9 (b), when the security level is 128-bit, the network traffic overhead of the research system is the lowest, with a minimum value of 1127 bytes, ElGamal-TLS of 6395 bytes, ECC-iTLS of 4623 bytes, and RSA-TLS of 5018 bytes. In summary, the proposed IoT-NS system has the lowest network traffic overhead in tests at different security levels. The experimental results of the complete encryption and decryption time and signature verification time of each system are shown in Fig. 10.

In Fig. 10 (a), the total time for key generation, encryption, and decryption of SM9-iTLS is 3.63 seconds. The total time for ElGamal-TLS is 9.27, ECC-iTLS is 5.33 seconds, and RSA-TLS is 7.31 seconds. In Fig. 10 (b), the total time for key generation, signature, and verification of SM9-iTLS is 3.65 seconds. The total time of ElGamal-TLS is 10.97 seconds, ECC-iTLS is 9.27 seconds, and RSA-TLS is 11.37 seconds. Therefore, the IoT-NS system based on improved SM9-iTLS can ensure good efficiency in information encryption, decryption, and identity authentication during secure communication. To further test and study the performance of the system, comparative experiments are conducted on system communication delay under different network delays, network bandwidth, and packet loss rates, as shown in Fig. 11.

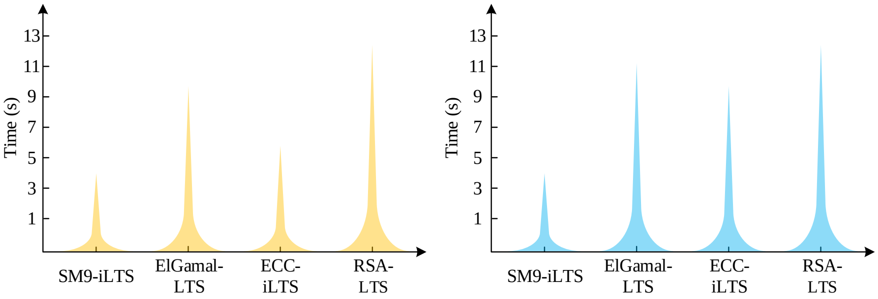
Fig. 11 (a) shows the experimental results under a network delay of 1-256ms. Under different network delays, SM9-iTLS has the lowest communication delay, while ElGamal-TLS has the highest communication delay. Fig. 11 (b) shows the experimental results under different network packet loss rates. When the network packet loss rate is below 15%, the communication delay of each system remains around 0, but as the packet loss rate increases, the communication delay continues to increase, especially ECC-iTLS, which shows exponential growth. When the packet loss rate is 25%, the communication delay of SM9 iTLS is 317ms, RSA-TLS is 5879ms, ElGamal-TLS is 1752ms, and ECC-iTLS is 11674ms. In Fig. 11 (c), under different network bandwidths, the communication delay of SM9-iTLS is the lowest, while ECC-iTLS has the highest communication delay. In summary, under different network quality communication environments, the SM9-iTLS IoT-NS system has the fastest response speed and the least burden on communication.



(a) 112-bit Indicates the network traffic cost of security registration

(b) 128-bit Indicates the network traffic cost of security registration

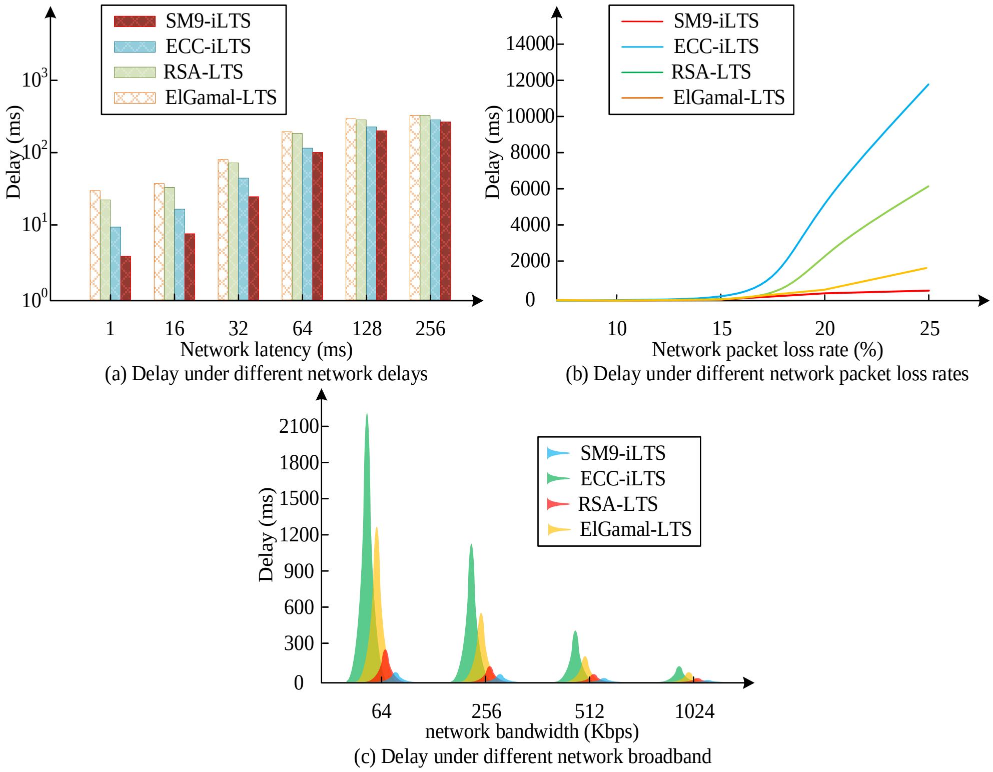
Fig. 9. Network traffic overhead for secure registration of 112 bit and 128 bit.



(a) Total time to complete encryption and decryption

(b) Total time to complete signature and verification

Fig. 10. The total travel time results of various model comparison experiments.



(a) Delay under different network delays

(b) Delay under different network packet loss rates

(c) Delay under different network broadband

Fig. 11. Communication latency of various systems under different latency, bandwidth, and packet loss rates.

IV. Discussion

This study compared and analyzed the computational efficiency and communication delay of the improved SM9 encryption algorithm and iTLS SP, and conducted comparative tests on the performance of the IoT-NS system based on the improved SM9-iTLS. The results of this study showed that the SM9 encryption algorithm took a total of 2.71s for key generation, encryption, and decryption, while the total time for key generation, signature, and verification was 3.74s, which is more efficient than other encryption algorithms. This result was similar to the research conclusion of Jing's team on improving the addition and multiplication sets of the SM9 algorithm [21]. The improved SM9 encryption algorithm not only had higher security but also improved computational efficiency. In the comparative experiment of iTLS, at the security levels of 112 and 128 bits, the latency of iTLS was 6.4ms and 9.9ms, slightly higher than TLS1.3-0-RTT, but at a low latency level. This result was similar to the conclusion of Zhang's team in designing SP based on La-TLS network [22]. Therefore, the improved SP could further enhance network security without increasing communication burden. Finally, in the performance analysis of the IoT-NS system based on improved SM9-iTLS, the network traffic overhead of SM9-iTLS was 794 and 1127 bytes at security levels of 112 and 128 bits, respectively. The total time for key generation, encryption, and decryption was 3.63s. The total time for generating, signing, and verifying dynamic keys was 3.65s. When the packet loss rate was 25%, the communication delay of SM9-iTLS was 317ms. Under different conditions, the network traffic overhead, encryption and decryption time, and latency of the research system were superior to other systems. This was similar to the conclusion obtained by Wang's team in the lightweight communication network based on SM9 and TLS improvements designed in 2024 [23]. This result indicated that the IoT-NS system based on improved SM9-iTLS not only enhanced the security capability of IoT device communication but also enhanced the efficiency of network communication.

V. Conclusion

The development of digitalization has brought about an increasing number of IoT devices, but these devices often have weaker self-protection capabilities. This study established a lightweight IoT-NS system by improving the SM9 encryption algorithm and combining iTLS SP with IoT devices. The experimental data demonstrated that the IoT-NS system, based on an enhanced SM9-iTLS, exhibited notable optimization in network resource utilization, computational efficiency, and communication delay, while simultaneously enhancing network security. The research system exhibited superior performance compared to other IoT-NS systems. Therefore, in situations where IoT device resources were limited, using the SM9 encryption algorithm to generate and distribute keys using PKG and the iTLS communication protocol to use dynamic keys for identity authentication, could improve the security of IoT networks. In practical environments, there are significant differences in the configuration of IoT devices. Some devices have strong computing power, but most devices have low computing power. In special environments, the time cost may be high and cannot meet the communication needs. Therefore, improving the hardware environment is the direction of later research. For example, despite the SM9 algorithm's robust security, the public key appends a sequence of inconsequential digits to the initial one, necessitating additional storage space and communication overhead. This is a problem that exists in certificate free systems, and further research can be conducted to improve it. The KGC of a single trust domain network is set by the network manager. However, in multi-trust domain networks, a single KGC greatly limits the scalability of iTLS. Future research will focus on studying cross domain authentication schemes based on identity passwords.

ACKNOWLEDGMENT

The research is supported by The 2023 Research Fund Project of Guangzhou Open University, titled "Research on the Model of Building a Digital University to Support the High Quality Development of Guangzhou Open University", with reference number: 2023KYZXYB09.

REFERENCES

[1] Qian J, Li H, Huo Y, Xing X. Empowering IoT security: an innovative handover-driven node selection approach to tackle conscious mobile eavesdropping. International Journal of Sensor Networks (2), 2024, 44(2):84-98. circuit for iot security. IET Circuits, Devices & Systems, 2022, 16(1):40-52.

[2] Ipseeta Nanda, Rajesh De. THE STATE OF THE ART IN ECO-FRIENDLY IOT. Information Management and Computer Science, 2022, 5(1):18-22.

[3] Ahmad Muhammad Thantawi, Sri Astuti Indriyati. Conceptual Design Impacts in New Normal Era: The Use of Artificial Intelligence (AI) And Internet Of Things (IOT) (Case Studies: Class Room And Restaurant). Acta Informatica Malaysia. 2022; 6(2): 39-42.

[4] Maseno E M, Wang Z, Liu F. Intrusion Detection System in IoT Based on GA-ELM Hybrid Method. Journal of Advances in Information Technology, 2023, 14(4):625-629.

[5] Marzouk R, Alrowais F, Negm N, Alkhonaini M A, Hamza M A, Rizwanullah M, Yaseen I, Motwakel A. Hybrid deep learning enabled intrusion detection in clustered iiot environment. Computers, Materials & Continua, 2022, 1(8):3763-3775.

[6] Kiruba D G, Benita J. A Survey of Secured Cluster Head: SCH based Routing Scheme for IOT based Mobile Wireless Sensor Network.ECS transactions, 2022, 107(1):16725-16745.

[7] Yadav K, Jain A, Alharbi Y, Alferaidi A, Alkwai L M, Ahmed N M O S, Hamad S A S. A secure data transmission and efficient data balancing approach for 5g-based iot data using uudis-ecc and lsrhs-cnn algorithms.

[8] Senthilkumar M, Murugan BS. Enhancing The Security of An Organization From Shadow Iot Devices Using Blow-Fish Encryption Standard. Acta Informatica Malaysia. 2022; 6(1): 22-24.

[9] Rani D, Tripathi S. Design of blockchain-based authentication and key agreement protocol for health data sharing in cooperative hospital network. Journal of supercomputing, 2024, 80(2):2681-2717.

[10] Lapworth L. Parallel encryption of input and output data for HPC applications. International Journal of High Performance Computing Applications, 2022, 36(2):231-250.

[11] Zhang Q, Zhao Z. Distributed storage scheme for encryption speech data based on blockchain and IPFS. Journal of supercomputing, 2023, 79(1):897-923.

[12] He D, Cai Y, Zhu S, Zhao Z, Chan S, Guizani M. A lightweight authentication and key exchange protocol with anonymity for iot. IEEE transactions on wireless communications, 2023, 22(11):7862-7872.

[13] Mvah F, Tchendji V K, Djamegni C T, Anwar A H, Tosh D K, Kamhoua C. Gatebasep: game theory-based security protocol against arp spoofing attacks in software-defined networks. International Journal of Information Security(1), 2024, 23(1):373-387.

[14] Rathee G, Kerrache C A, Calafate C T. An Ambient Intelligence approach to provide secure and trusted Pub/Sub messaging systems in IoT environments. Computer networks, 2022, 218(9):1-9.

[15] Chowdhury R R, Idris A C, Abas P E. Identifying SH-IoT devices from network traffic characteristics using random forest classifier. Wireless networks, 2024, 30(1):405-419.

[16] Jinghua Z. Special Issue on Machine Learning and Big Data Analytics for IoT Security and Privacy (SPIoT2022). Neural computing & applications, 2024, 36(5):2119-2120.

[17] Mohanrasu S S, Udhayakumar K, Priyanka T M C, Gowrisankar A, Banerjee S, Rakkiyappan, R. Event-triggered impulsive controller design for synchronization of delayed chaotic neural networks and its fractal reconstruction: an application to image encryption. Applied mathematical modelling, 2023, 115(1):490-512.

[18] Wu S T. An Application of Keystream Using Cellular Automata for Image Encryption in IoT. Journal of Internet Technology, 2023, 24(1):149-162.

[19] Zhang Y, Wu Q, Wang P, Wen L, Luan Z, Gu C. Tvd-pb logic circuit based on camouflaging

[20] Yadav A K, Misra M, Pandey P K, Braeken A, Liyange M. An improved and provably secure symmetric-key based 5g-aka protocol. Computer networks, 2022, 218(9):1-13.

[21] Jing S, Yang X, Feng Y, Liu X, Hao F, Yang Z. Hardware Implementation of SM9 Fast Algorithm Based on FPGA. Atlantis Press, 2022, 12(27):797-803.

[22] Xinglong Z, Qingfeng C, Yuting L I. LaTLS:A Lattice-Based TLS Proxy Protocol. Chinese Journal of Electronics, 2022, 31(2)313-321.

[23] Wang D, Dong L, Tang H, Gu J, Liu Z, You X. SDN Security Channel Constructed Using SM9. International Symposium on Digital Forensics and Security, 2024, 4(12):1-5.

Application and Effectiveness of Improving Retrieval Systems Based on User Understanding in Smart Archive Management Systems

Chao Yan
General Administration Office, YanCheng Polytechnic College, Yancheng, 224005, China

Abstract—**In traditional archive management systems, keyword-based retrieval systems often fail to meet users' personalized and precise retrieval needs. To solve this problem, a knowledge graph is first constructed using bidirectional long short-term memory networks and conditional random fields and combined with user understanding-based semantic retrieval to obtain an improved personalized retrieval system. The research results show that the improved personalized retrieval system has significantly better retrieval accuracy and recall rate than traditional retrieval systems. The improved personalized retrieval system has retrieval accuracy rates of 90.24%, 89.65%, 87.52%, 96.33%, and 95.18% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, respectively, and recall rates of 89.35%, 91.57%, 89.34%, 97.54%, and 96.63%, respectively. Applying it to the smart archive management system, the accuracy of archive retrieval, personalized recommendation accuracy, response time, and user satisfaction are significantly better than conventional management systems. The improvement and introduction of personalized retrieval systems based on user understanding and knowledge graphs have achieved significant results.**

Keywords—Knowledge graph; user understanding; retrieval system; smart archive management; BiLSTM-CRF

I. INTRODUCTION

The rapid development of the information age has led to increasing attention to the Smart Archive Management System (SAMS). SAMS is a system that utilizes modern information technologies such as big data, cloud computing, the Internet of Things, artificial intelligence, etc. to intelligently manage archive information. It not only includes digital archival information input, storage, and retrieval, but also includes the protection, analysis, and utilization of archival digital resources, as well as online services and interactive functions. This system aims to improve the efficiency and quality of archive management while facilitating users to quickly and accurately obtain the required archive information. SAMS is mainly used by government agencies, enterprises, educational institutions, research institutions, archive workers, and the general public. It is responsible for managing public archives and providing public services, managing enterprise documents and commercial archives, managing academic and administrative archives of schools, and querying and utilizing archive resources. Traditional archive management methods often rely on paper documents and manual operations, which have problems such as low management efficiency and difficulties in information sharing. In the context of intelligence and automation, providing a more accurate, intelligent, and personalized retrieval system has become an important challenge for smart archive management systems. Traditional retrieval systems use keyword matching, which often makes it difficult to accurately understand the user's query intent and needs. Because relying solely on keyword matching, the system cannot understand the true intention behind the user's query, resulting in the returned results may not be accurate enough or do not meet the user's expectations. And traditional systems lack the ability for personalization. This method does not consider user behavior patterns and preferences, and cannot provide targeted search results, which cannot meet the personalized needs of different users. SAMS has replaced the old AMS because it provides a more efficient and comprehensive management approach. SAMS can improve efficiency, automate processing processes, reduce manual operations, and improve work efficiency. Meanwhile, digital archive information can be quickly retrieved, improving the speed of information acquisition. Cloud storage technology enables remote access and sharing of archival information, advanced data security technologies provide better protection of archival information, and data analysis tools can help decision-makers better understand archival information and make wiser decisions. In the context of intelligence and automation, providing a more accurate, intelligent, and personalized retrieval system has become an important challenge for SAMS [1-2]. Traditional retrieval systems use keyword matching, which often makes it difficult to accurately understand users' query intentions and needs. Therefore, providing personalized and accurate retrieval services has become particularly important for meeting the needs of users. User behavior mainly focuses on the operational behavior of users in the system. User preferences refer to the degree of user preference for the system or information, while semantic understanding is a deep understanding of user input or needs. Choosing to model user behavior is to improve the operational process and efficiency of the system. Choosing to model user preferences is for personalized recommendations and customized services. Choosing to model semantic understanding is to more accurately understand user needs and provide accurate retrieval results. The improvement of retrieval systems based on user understanding can provide users with more adaptive and personalized retrieval results by deeply understanding their query intentions and information needs, as well as modeling and analyzing their behavior patterns and preferences [3-5]. This method can be improved from multiple perspectives, including modeling user behavior, user preferences, and semantic understanding.

Knowledge graph is a structured knowledge representation method that describes the relationships between entities in the real world. Knowledge graph can provide richer and more accurate semantic information, providing strong support for personalized retrieval. By matching users' query intentions with entities and relationships in the knowledge graph, the accuracy and effectiveness of retrieval can be effectively improved [6-7]. The study utilizes bidirectional long short-term memory networks and conditional random fields to construct a knowledge graph, and combines it with user understanding-based semantic retrieval to obtain an improved personalized retrieval system. We hypothesize that combining user understanding with knowledge graph can better understand users' real needs and provide more accurate and personalized archive retrieval services. Applying the improved personalized retrieval system to SAMS will have more impact. The innovation of this study is to use Bidirectional Long Short Term Memory Network (BiLSTM) and Conditional Random Field (CRF) to construct a knowledge graph, namely BiLSTM-CRF. This system can better capture the semantic associations between queries and archives, thus achieving more intelligent and accurate retrieval results. Research has shown that the improved personalized retrieval system has significantly better retrieval accuracy and recall than traditional retrieval systems. Although significant progress has been made in many aspects of existing archive management systems and data processing frameworks, there are still several unresolved research gaps. Firstly, most existing systems lack flexibility and adaptability to meet different user needs. The current system is usually designed to be fixed and difficult to adjust according to the specific needs of different user groups, resulting in poor efficiency and effectiveness in handling cross-domain and diverse data. This limitation is particularly evident in application scenarios that require cross institutional collaboration and cross-platform integration. Secondly, there is a lack of in-depth research on the management of data bias and the handling of dynamic content changes. Although deviation detection has been preliminarily applied in some data analysis fields, its application in data archiving and long-term management is still insufficient. In addition, how to efficiently detect and adapt to changes in archived content to ensure the freshness and relevance of data remains an urgent challenge that needs to be addressed. This study aims to fill these research gaps and propose a flexible and adaptive archive management system framework that can not only dynamically adjust according to the needs of different user groups, but also effectively handle potential biases and content changes in data. By introducing advanced machine learning algorithms and dynamic update mechanisms, this study provides a new solution to improve the accuracy and fairness of data management. This will not only help improve the performance and practicality of existing systems but also promote the development of data archiving management and facilitate cross-industry and interdisciplinary data sharing and cooperation.

II. Related Work

Smart Archive Management (SAM) is a method that utilizes artificial intelligence and automation technology to improve the efficiency and quality of archive management. This technology can draw on the concepts and methods of knowledge graph and semantic retrieval to provide more intelligent and valuable archival services. Li Y proposed a cross domain recommendation method based on user preference perception and graph attention networks to improve the performance of knowledge graph recommendation systems. This method utilized a graph embedding model to obtain preference aware entity embedding, and combined preference features for personalized recommendation. It could effectively alleviate the problem of data sparsity and had good comprehensive performance, which can generate good personalized recommendation results [8]. Sui Y have constructed a model based on causal filters to improve the performance of knowledge graph question answering systems. This model utilized a data-driven approach to cause and effect interference in the relationship representation space, and disconnected other confounding factors in the knowledge embedding space through causal intervention. This method had good robustness and could effectively improve the comprehensive performance of the knowledge graph question answering system [9]. HX A et al. designed a knowledge graph model based on semantic fusion to address the issue of inaccurate semantic expression in knowledge graphs. It extracted subject entities by constructing a deep transformation model and generated candidate answer sets using dynamic candidate paths. This model could effectively solve the problem of inaccurate semantic expression in the knowledge graph, and improve the overall performance of the model [10]. Wang et al. designed a Top-k semantic aware query method based on semantic boundary perception to improve the accuracy of Knowledge Graph Star Query (KGSQ). This method utilized boundary deletion for matching with low semantic similarity, which can improve the accuracy of KGSQ [11].

User understanding can improve retrieval efficiency and accuracy, enhance user experience, and promote the development of personalized services. A deep understanding of user needs and behaviors can help improve the data analysis and processing capabilities behind retrieval systems, enabling them to provide more accurate and effective decision support, such as optimizing resource allocation and guiding policy formulation. In order to explore the impact of clarifying questions on user behavior and the ability to identify relevant information, Zou J utilized implicit interaction data and explicit user feedback to discover that high-quality clarifying questions can improve user performance and satisfaction [12].

Attention mechanisms and models such as BiLSTM can be applied to different stages of semantic retrieval tasks, such as semantic modeling correlation evaluation of queries and document sequences, to improve the accuracy and effectiveness of retrieval. Dong X et al. proposed a cross-modal graph attention method that combines recursive gated neural networks and attention mechanisms to address the semantic gap issue in data in cross-modal retrieval. This method could eliminate heterogeneous gaps between modalities, effectively solving the semantic gap problem of data in cross-modal retrieval [13]. To improve the performance of knowledge graph embedding, Dai has constructed a knowledge graph representation learning model based on generative adversarial networks, using Wasserstein distance instead of traditional divergence. Wasserstein distance could effectively solve the problem of small gradients on discrete data and improve the embedding performance of knowledge graphs [14]. Liu T and other scholars constructed a Chinese WeChat click bait dataset to detect the security of hyperlinks in online social media, and constructed a WeChat click bait detection method based on BiLSTM and multiple features, introducing attention mechanisms. This method could effectively detect bait links and had a high accuracy [15]. Hu Y et al. proposed a semantic behavior prediction method to improve the prediction accuracy of traffic participants' behavior in an autonomous vehicle. This method constructed a universal semantic representation method suitable for the driving environment and converted it into a spatiotemporal semantic map to infer the internal relationship between the two. This method had high prediction accuracy and was beneficial for solving problems under different traffic conditions [16].

In summary, many scholars have conducted research on knowledge graphs, semantic retrieval, and smart archive management, but how to combine them is an unresolved issue [17-18]. Knowledge graphs, semantic retrieval, and smart archive management are interrelated concepts and technologies, and have important application value in the fields of information management and services. Attention mechanisms, bidirectional long short-term memory networks, and semantic retrieval are interrelated and play a role in semantic modeling, correlation evaluation, and information retrieval, aiming to improve the performance and effectiveness of retrieval systems. In view of this, this study will use a bidirectional long short-term memory network to construct a knowledge graph, improve personalized retrieval systems based on user understanding and knowledge graphs, and further enhance SAMS [19-20].

III. Improvement of Retrieval Function in SAMS

To improve the performance and user experience of the retrieval system in SAMS, this study improves the personalized retrieval system based on user understanding and knowledge graph. The first step is to construct a knowledge graph and use the BiLSTM-CRF model to complete the process of entity recognition and relationship extraction in the archival field. Then, combining user understanding and knowledge graph, users' search requirements for keywords can be better understood and met from the traditional semantic retrieval level.

A. Construction of Knowledge Graph Based on BiLSTM-CRF

The knowledge graph network is a relatively structured entity semantic knowledge network used to analyze and express the objective and real world, as well as the relationships among various concepts, entities, etc. A knowledge event graph network is composed of a set of relationships between entities in different time, space, and states. This network forms a new semantic based knowledge network model by describing the interrelationships between concepts and entity concepts [21-22]. Fig. 1 shows the process of constructing a knowledge graph.

Fig. 1 shows the knowledge graph constructed in this study, In Fig. 1, the process of constructing a knowledge graph, it is necessary to first obtain data, transform the data into knowledge, then fuse the knowledge, and provide knowledge graph visualization services. Finally, apply based on knowledge graph. However, this study designs and models the ontology hierarchy, relationships, and attributes of archives. Among them, attribute structure refers to the organizational description of archive metadata and content, clarifying the characteristics and values of each archive item. By establishing a correct ontology hierarchy model, relevant knowledge can be extracted and analyzed, and clear semantic descriptions and relationship explanations can be provided for knowledge concepts within the domain. The data sources of archival knowledge include data extracted from existing business systems and other fields, and the data is divided into institutional, semi-structured, and non-manual structured.

From the knowledge framework of archive ontology language, it can be seen that entity recognition is required for the institutions, characters, files, time, location, etc. of archive ontology language. The language relationships of archive ontology mainly include synonyms, antonyms, context, subclasses, and other relationships [23]. For entity recognition, this study intends to select the BiLSTM-CRF model for extracting relational data and adopt a distributed relational data extraction technology based on remote data supervision to reduce the model's dependence on manually annotated data.

Next, construct the BiLSTM-CRF model. Recurrent Neural Network (RNN) is a type of neural network that performs temporal processing on sequence data. Time series data refers to the data collected at different time points, which reflects the situation or degree to which something, phenomenon, etc. occurs over time. Long Short-Term Memory (LSTM) networks are a new type of RNN that aims to overcome the problems of gradient loss and gradient explosion in long sequence learning. Compared with RNN, LSTM has better performance over longer time series [24]. LSTM is composed of three parts: input gate, forgetting gate, and output gate, achieving protection and control of information. Fig. 2 shows the LSTM's basic structure.

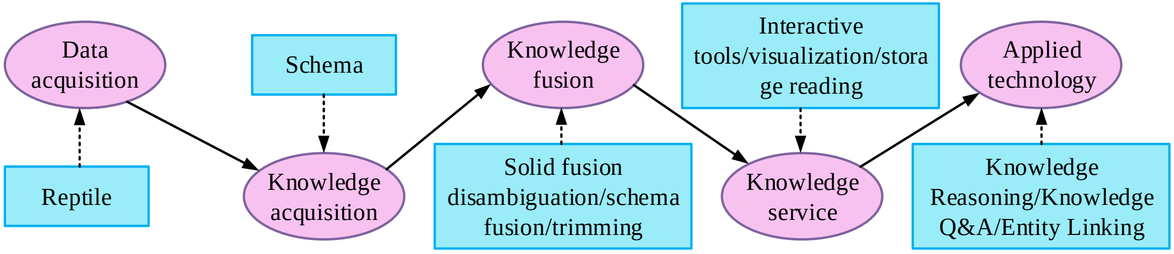

Data acquisition
Schema
Knowledge fusion
Interactive tools/visualization/storage reading
Applied technology
Reptile
Knowledge acquisition
Solid fusion disambiguation/schema fusion/trimming
Knowledge service
Knowledge Reasoning/Knowledge Q&A/Entity Linking

Fig. 1. The construction process of knowledge graph.

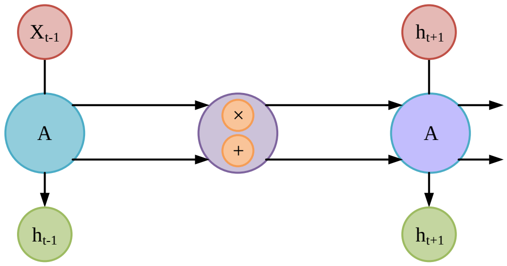


Fig. 2. The basic structure of LSTM.

BiLSTM is an RNN model that can process sequence data and capture long-term dependencies within the sequence. By running both a forward LSTM and a backward LSTM simultaneously, BiLSTM can utilize contextual information to better understand and encode input sequences. CRF refers to the conditional probability distribution modeling of another set of output variables given a set of input random variables. Its characteristic is to perform under the premise that the output variable satisfies the Markov random field [25].

The BiLSTM-CRF model is a neural network model used for sequence annotation tasks. It combines BiLSTM and CRF to simultaneously consider the dependency relationship between contextual information and labels. By utilizing the BiLSTM-CRF model, the process of entity recognition and relationship extraction in the archival field can be completed, as shown in Fig. 3.

After completing entity recognition and relationship extraction in the archival field, it is also necessary to store and retrieve the knowledge graph. Knowledge graphs are generally accessed through graph databases. After constructing and storing the enterprise archival knowledge graph in the graph database, it is necessary to combine advanced graph data retrieval technology to improve and enhance the efficiency of data query and processing in the enterprise archive knowledge graph. Its purpose is to provide support for achieving massive real-time dynamic information queries and data inference analysis.

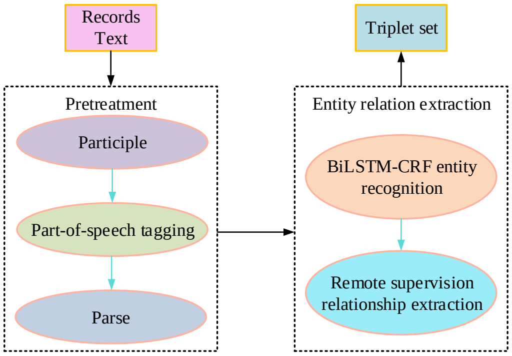

Records
Text
Triplet set
Pretreatment
Participle
Part-of-speech tagging
Parse
Entity relation extraction
BiLSTM-CRF entity recognition
Remote supervision relationship extraction

Fig. 3. Entity recognition and relationship extraction process in the field of archives.

B. Improvement of Personalized Retrieval System Based on User Understanding and Knowledge Graph

Traditional search engines mainly provide search results based on keyword matching, neglecting users' personalized needs and contextual information. However, users' search behavior and preferences are diverse, and traditional methods often find it difficult to accurately understand and meet these personalized needs. With the rapid growth of the internet scale and information content, it has become very important to better understand the semantic relationship between users and information [26]. Fig. 4 is a personalized retrieval system framework based on user understanding and knowledge graph.

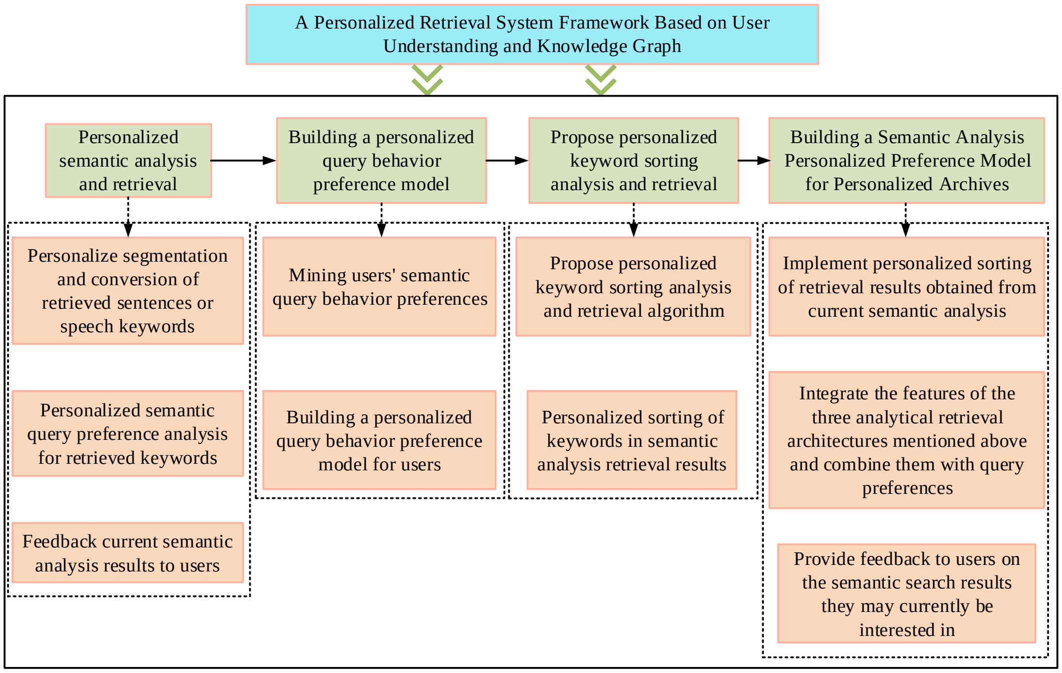

A Personalized Retrieval System Framework Based on User Understanding and Knowledge Graph
Personalized semantic analysis and retrieval
Building a personalized query behavior preference model
Propose personalized keyword sorting analysis and retrieval
Building a Semantic Analysis Personalized Preference Model for Personalized Archives
Personalize segmentation and conversion of retrieved sentences or speech keywords
Personalized semantic query preference analysis for retrieved keywords
Feedback current semantic analysis results to users
Mining users' semantic query behavior preferences
Building a personalized query behavior preference model for users
Propose personalized keyword sorting analysis and retrieval algorithm
Personalized sorting of keywords in semantic analysis retrieval results
Implement personalized sorting of retrieval results obtained from current semantic analysis
Integrate the features of the three analytical retrieval architectures mentioned above and combine them with query preferences
Provide feedback to users on the semantic search results they may currently be interested in

Fig. 4. A personalized retrieval system framework based on user understanding and knowledge graph.

From Fig. 4, it can be seen that in the personalized retrieval system framework, the first step is to conduct personalized semantic analysis retrieval. The second step is to build a personalized query behavior preference model. The third step is to propose a personalized keyword sorting analysis retrieval algorithm. The last step is to build a semantic analysis personalized preference model for personalized archives. Integrating the features of the three analytical retrieval architectures mentioned above, combining them with query preferences, to achieve personalized sorting of the retrieval results obtained from current semantic analysis, and providing feedback to users on the semantic retrieval results they may be interested in at present.

Traditional personnel file retrieval and query engines do have some limitations in providing personalized personnel file services. It is usually based on keyword matching and cannot fully understand users' personalized needs and contextual information [27]. Based on this, this study models personal information, enterprise information, and other information in the personnel information system based on user preference for personnel file query and data mining methods, achieving the retrieval of personal information in the personnel information system. Fig. 5 shows the mining framework operation process of user query preferences.

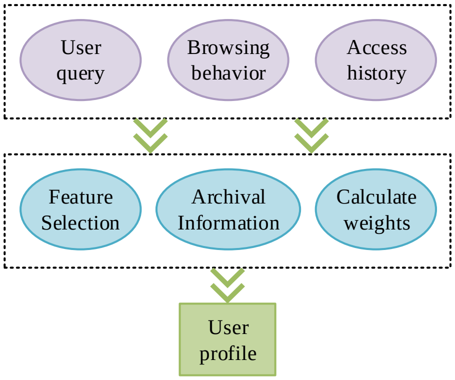


Fig. 5. The mining framework operation process of user query preferences.

In Fig. 5, the mining of user query preferences mainly includes seven steps, aiming to accurately analyze and represent user query preferences. These steps help to understand users' interests and preferences, and provide personalized search results, recommendations, and services based on these preferences.

The user query preference model is an algorithm model that describes the degree of interest of users in different types of archives. Each file object has a weight value that represents the user's level of interest in the file type. The higher the weight, the more interested the user is in this type of file [28]. This study uses probability models to compare and express user preferences for queries. It puts the file related data and file information related data from the user preference database into the user preference database, and generates a user preference probability distribution model for queries through analysis. w_i is defined as a weighting factor, which represents a user's preference for querying type i and file c_i in a short period of time. The calculation process is Eq. (1).

$$w_i = AF\left(c_i, \frac{x+1}{T}\right) \tag{1}$$

In Eq. (1), $AF\left(c_i, \frac{x+1}{T}\right)$ is used to describe the average value of file type c_i that users browse during period T. x represents the number of files browsed within T. By solving the weights of all file types in order, the user's query preference vector based on the file category can be obtained. The query preference characteristics of users are not fixed, and the required file types may also change under different job positions. So, after establishing a user's query preference model, there is a need for an algorithm that can update according to the user's query preferences. By setting an appropriate forgetting time, the problem of user query preference transfer can be solved. This study proposes a new user preference model based on forgetting factor using the Ebbinghaus memory curve. The expression is Eq. (2).

$$F(x) = \exp\left(-\frac{\log_2(t)}{f}\right) \tag{2}$$

In Eq. (2), t represents the time interval between the current date and the query preference creation date. f represents the half-life, which means that half of the time the user forgets will last for at least f days. Then, the user's query preference can be calculated based on the probability model and the user's query preference model [29].

Traditional keyword based retrieval methods cannot understand and meet users' retrieval needs for keywords at the traditional semantic level. Therefore, this study proposes a personalized retrieval method based on knowledge graph, which can better understand and meet users' search requirements for keywords from the traditional semantic retrieval level. Fig. 6 shows the semantic personalized retrieval process of archives based on knowledge graph.

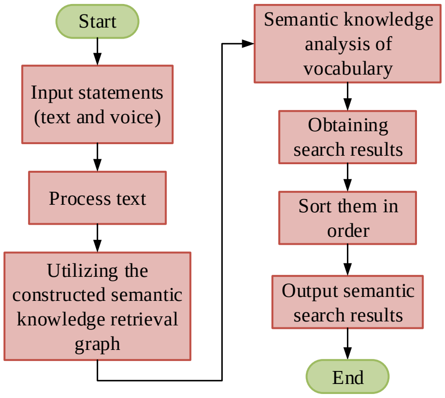

Start
Input statements (text and voice)
Process text
Utilizing the constructed semantic knowledge retrieval graph
Semantic knowledge analysis of vocabulary
Obtaining search results
Sort them in order
Output semantic search results
End

Fig. 6. The semantic personalized retrieval process of archives based on knowledge graph.

In Fig. 6, in the process of semantic personalized retrieval of archives based on knowledge graph, the first step is to input statements, including two types: text and speech. It is necessary to convert speech into text and then process the text. Afterwards, the constructed semantic knowledge retrieval graph can be used for semantic knowledge analysis of vocabulary. Then the search results are obtained and sorted to output semantic search results [30]. If the retrieval vocabulary is not included in the archive knowledge graph, semantic expansion and semantic mapping of the vocabulary are required, and the expanded retrieval vocabulary and the original retrieval vocabulary are added to synonymous concepts and segmentation specialized dictionaries respectively. If the current search term does not exist in the synonym dictionary, semantic similarity needs to be calculated, as shown in Eq. (3).

$$Sim(c,w)=\beta * Sim\_(c,w)+(1-\beta)* Sim\_2(c,w) \quad (3)$$

In Eq. (3), $Sim(c,w)$ is used to describe the minimum editing time distance between two words, calculated as shown in Eq. (4).

$$Sim_1(c,w)=1-\frac{ED(c,w)}{MLen(c,w)} \quad (4)$$

In Eq. (4), $Sim_2(c,w)$ is used to describe the distance from the editor, meaning the maximum semantic cosine length between and $\cos(e_c,e_w)$ represents the maximum semantic cosine distance between two words. Its similarity is equal to the distance between two synonyms and the embedded cosine similarity, as shown in Eq. (5).

$$Sim_2(c,w)=\cos(e_c,e_w) \quad (5)$$

In Eq. (5), e_c and e_w are embedded representations of words c and w. Then, by generating and constructing the inference graph of archival ontology knowledge, an archival knowledge inference system was constructed [31]. Through the Jena inference machine in the archive, semantic inference can be performed on the existing hierarchical knowledge relationships in the document. This enables semantic inference of the existing upper and lower knowledge relationships in archives, as well as better analysis and retrieval of the upper and lower semantics of documents. A key challenge faced by the system in the process of data archiving and management is how to handle potential biases in user data and changes in archive content. To effectively address these issues, the system has introduced a bias detection algorithm that can identify potential biases in the data. For example, in user-generated content, the system can identify anomalies or biases by analyzing data distribution and usage patterns. Through machine learning models, the system can automatically label these potential biased data and make corrections as necessary. This not only helps maintain the objectivity of data, but also improves the accuracy of data analysis results. The study selected 100 different types of individuals aged 18-45 for the experiment, all of whom were from the local archives bureau. The selected individuals were all from different professions and had significant occupational differences. The participants were divided into five groups, with 20 people in each group. Information was collected through a questionnaire survey. In the experiment data collection of this study, the network data are all from the Internet, and the files are stored separately in the database and web pages; the file data is provided by the local archives bureau, which scans paper files, converts them into PDF format electronic files, and saves them in a MySQL database.

IV. Application and Effect Analysis of Improving Retrieval System on SAMS

To verify the performance of the improved retrieval system, the performance of the knowledge graph based on BiLSTM-CRF and the retrieval performance of the preference semantic retrieval model based on user understanding are first analyzed, and the improved retrieval system was applied to the smart archive management system.

A. *Performance Analysis of Improved Personalized Retrieval System Based on User Understanding and Knowledge Graph*

In order to verify the performance of personalized retrieval systems based on user understanding and knowledge graph improvement, knowledge graph, user understanding, and improved personalized retrieval systems were analyzed in sequence. In the experimental data collection of this study, network data was sourced from the Internet, and archives were stored separately in databases and web pages. The piece of data is provided by the local archives bureau, which scans paper files, converts them into electronic files in PDF format, and saves them in a MySQL database. The file data mainly includes retrieval information for specific occupational groups. Table I shows the experimental system settings.

The experimental setup introduced in Table I can meet the experimental analysis of user understanding, knowledge graph, and retrieval system.

For the improvement of user understanding of the smart archive management system, specific occupational groups such as students, civil servants, demobilized soldiers, law enforcement personnel, and retirees were selected for application and effect analysis. These occupational groups represent users in society with different information needs and backgrounds in using archive management systems. Students need academic resources such as historical data and scientific research data, civil servants and law enforcement personnel need policy documents, laws and regulations, work files, etc. Demobilized soldiers and retirees need relevant welfare policies, personal service records, and other information. They need to obtain academic resources, research data, policy documents, welfare policies, laws and regulations, and other information from archives. Obtain information directly from users through methods such as questionnaire surveys, face-to-face interviews, and online interviews. Analyze the logs and usage records of the smart archive management system, and understand the user's usage habits and needs. Study specific cases of user groups using the system, including successful cases and problematic cases. Design representative tasks or scenarios for users from different professions to perform in a smart archive management system, such as retrieving specific information, using a certain service, etc. Collect data through system logs, user feedback, observation records, and other methods, including user operation behavior, task completion time, user satisfaction, etc. Using methods such as statistical analysis, content analysis, and behavioral analysis, analyze data from both quantitative and qualitative perspectives, identify user needs, evaluate system effectiveness, and identify system shortcomings. Firstly, to analyze the performance of the BiLSTM-CRF based knowledge graph and compare it with conventional knowledge graphs. The results are shown in Fig. 7.

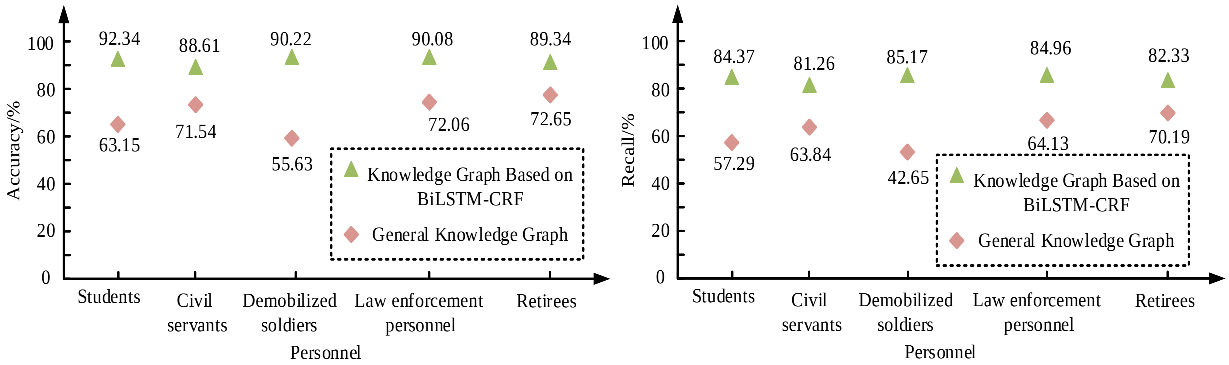
Fig. 7(a) and Fig. 7(b) respectively represent the retrieval accuracy and recall of BiLSTM-CRF based knowledge graphs and conventional knowledge graphs. The former has retrieval accuracy rates of 92.34%, 88.61%, 90.22%, 90.08%, and 89.34% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, respectively. The latter are 63.15%, 71.54%, 55.63%, 72.06%, and 72.65%, respectively. The recall rates of the former are 84.37%, 81.26%, 85.17%, 84.96%, and 82.33%, respectively, while the latter is 57.29%, 63.84%, 42.65%, 64.13%, and 70.19%, respectively. The retrieval accuracy and recall of knowledge graphs based on BiLSTM-CRF are significantly superior to conventional knowledge graphs. Next, analyzing the retrieval performance of the preference semantic retrieval model based on user understanding, and comparing it with conventional semantic retrieval models. The results are shown in Fig. 8.

Fig. 8(a) and Fig. 8(b) represent the retrieval accuracy and recall rates of the preference semantic retrieval model based on user understanding and the conventional semantic retrieval model, respectively. The former has retrieval accuracy rates of 93.64%, 90.12%, 88.47%, 95.24%, and 89.68% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, respectively. The latter are 71.23%, 70.36%, 69.53%, 68.18%, and 71.62%, respectively. The recall rates of the former are 86.55%, 91.62%, 82.18%, 84.23%, and 96.35%, respectively, while the latter are 68.37%, 71.34%, 64.38%, 71.09%, and 80.24%, respectively. The experimental results indicate that, the retrieval accuracy and recall of the preference semantic retrieval model based on user understanding are significantly superior to conventional semantic retrieval models. Then to analyze the personalized retrieval system based on user understanding and knowledge graph improvement, compare it with traditional retrieval systems, and the results are shown in Fig. 9.

In Fig. 9, compared to traditional retrieval systems, the improved personalized retrieval system has better accuracy and recall. The improved personalized retrieval system has retrieval accuracy rates of 90.24%, 89.65%, 87.52%, 96.33%, and 95.18% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, respectively. The retrieval accuracy of traditional retrieval systems is 79.63%, 78.27%, 74.57%, 80.24%, and 82.15%, respectively. The recall rates of the former are 89.35%, 91.57%, 89.34%, 97.54%, and 96.63%, respectively, while the latter is 72.14%, 74.54%, 70.68%, 79.27%, and 81.08%, respectively. The experimental results indicate that this study can improve the efficiency and user experience of smart archive management systems through user understanding-based retrieval systems, to achieve more accurate and personalized information retrieval and management.

TABLE I. SETTING OF EXPERIMENTAL SYSTEM ENVIRONMENT

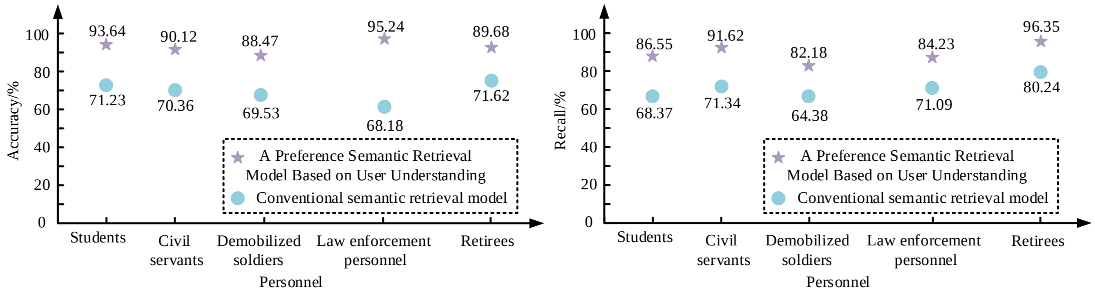
Number	Project	Type	Name
(1)	Hardware requirements	Server	Hewlett Packard Enterprise
		Cloud platform	Microsoft Azure
(2)	Software requirements	Operating system	Windows
		Database Management System	MySQL
		Development environment	Python
(3)	Knowledge Graph Construction	Build Tools	OpenKG, Protégé, etc
(4)	Algorithms and Technologies	Algorithm	BiLSTM-CRF



(a) Accuracy

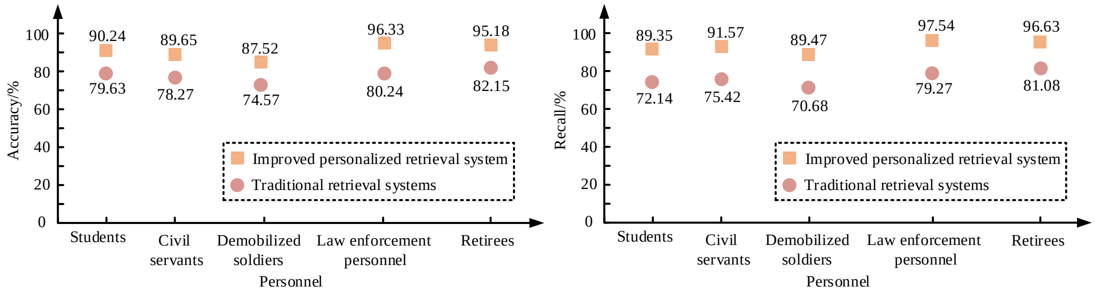
(b) Recall

Fig. 7. Performance analysis of knowledge graph based on BiLSTM-CRF.



(a) Accuracy (b) Recall

Fig. 8. Analysis of retrieval performance of preference semantic retrieval model based on user understanding.



(a) Accuracy (b) Recall

Fig. 9. Comparison of personalized vs. traditional retrieval systems, showing improvements in user understanding and knowledge graphs.

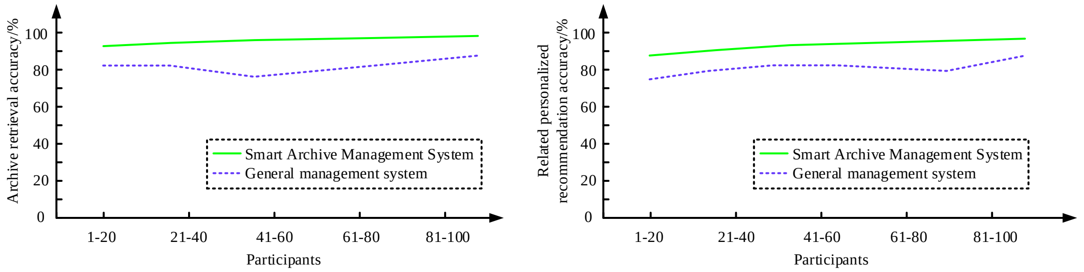
B. *Analysis of the Application Effect of Improved Retrieval System in SAMS*

In order to comprehensively evaluate the application effect of improving the retrieval system in the smart archive management system, first set goals, determine the user group participating in the evaluation, design representative retrieval tasks, and set a series of quantitative indicators to evaluate system performance. It is necessary to establish a control group in order to more accurately evaluate the improvement effect. In order to ensure the reliability of the experimental results, it is necessary to conduct experimental tests in a controlled environment and minimize the interference of external factors.

To select an archive dataset with a certain scale and diversity, and pre-process it into a format suitable for simulation analysis. 100 individuals aged 18 to 45 of different types were randomly selected to conduct experiments to analyze the application effect of improved retrieval systems in SAMS. Firstly, the accuracy of file retrieval and related personalized recommendations of the participants in the experiment were compared. The participants were divided into five groups of 20 people each, and the results are shown in Fig. 10.

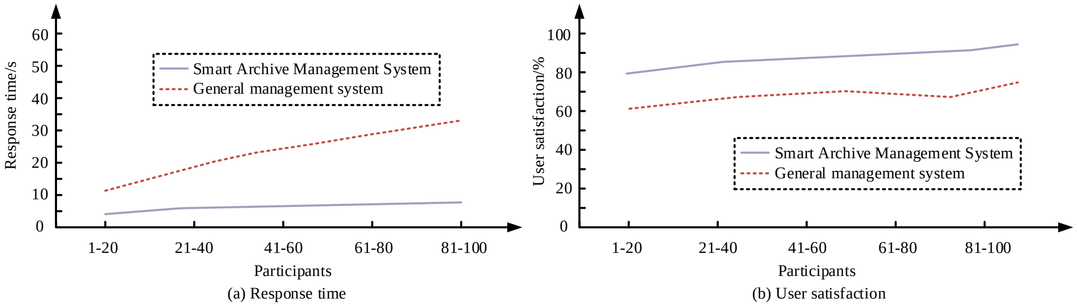
In Fig. 10, the accuracy of file retrieval and related personalized recommendations in SAMS are significantly superior to conventional management systems. The accuracy of the former is more stable and smoother, while the accuracy of the latter shows a decrease. The experimental results indicate that the proposed system model has better performance. Next, to analyze the response time and user satisfaction of different personnel towards improving the retrieval system, as shown in Fig. 11.

In Fig. 11, the response time of SAMS is significantly shorter than that of conventional management systems, while user satisfaction is significantly higher than that of conventional management systems, indicating that the former has better overall performance. The response time of SAMS and conventional management systems is up to 8 seconds and 32 seconds respectively, with the highest user satisfaction of 92.34% and 71.42%, respectively.



(a) Archive retrieval accuracy (b) Related personalized recommendation accuracy

Fig. 10. Accuracy of archive retrieval and personalized recommendation.



(a) Response time

(b) User satisfaction

Fig. 11. Response time and user satisfaction of different personnel towards improving the retrieval system.

V. Discussions

In order to improve the performance and user experience of retrieval systems, the improvement of personalized retrieval systems based on user understanding and knowledge graphs has aroused the interest of researchers and practitioners. Research combines knowledge graphs and user understanding to construct personalized retrieval systems that meet user preferences. In order to meet the personalized needs of different users, the system continuously analyzes users' behavior patterns and preferences through machine learning algorithms and user feedback mechanisms. The system can provide personalized recommendations based on the user's operational history. In order to better understand the application of the system in real-world scenarios, research considers archive management environments of different scales and types. In large multinational corporations, this system can be used to manage millions of employee files and financial records. In government agencies, the system can process historical documents spanning decades. The scalability of the system enables it to handle the transition from paper documents to electronic documents and supports complex permission management to ensure the security of sensitive data. Through distributed storage and processing technology, the system is able to effectively manage and retrieve large amounts of documents, adapting to the constantly growing amount of data. In academic institutions, the archiving and management of research data require extremely high flexibility. This system meets the diverse data management needs of researchers by supporting various data formats and custom metadata tags.

In the actual implementation process, the system will face a series of technical challenges. For example, the real-time requirements of data processing may lead to increased system performance pressure, especially in high concurrency environments. In addition, data privacy and security issues are also a key challenge, and the system needs to effectively prevent potential data leakage risks. There are differences in technology acceptance among different user groups, which may affect the promotion and use of the system. With the continuous development of big data and artificial intelligence technology, the system should introduce more intelligent deviation detection and correction mechanisms to ensure the integrity and accuracy of data. Neglecting this aspect may lead to inaccurate data analysis results and even affect the effectiveness of decision-making. Therefore, I believe that future research should place greater emphasis on identifying and correcting data biases, and developing more robust data management strategies. In addition, the discussion on system scalability and adaptability also made me realize that relying solely on existing technological means is not enough. Interdisciplinary collaboration is needed to combine knowledge and technology from different fields in order to design more efficient and intelligent archiving management systems. This is not only a technical challenge, but also a management and policy issue that requires joint efforts from all parties to promote the development of this field.

VI. Future Work Prospects

This study has achieved preliminary results in the design and development of a flexible and adaptable archive management system, but in order to cope with constantly changing user needs and technological challenges, the following key areas still need to be further explored and improved in future work.

Firstly, improving the scalability of the system is one of the key focuses of future work. With the rapid growth of data volume and user numbers, the system must be able to effectively expand to support large-scale data storage and management. Future work will focus on developing more efficient data storage and retrieval algorithms, while exploring distributed storage and computing technologies to enhance system processing power and response speed. Through these technological means, we hope to achieve seamless system expansion without sacrificing performance. Secondly, in order to better meet the needs of different user groups, we plan to optimize the user interface and interaction design. Different users have different operating habits and needs, and in-depth user research will help collect and analyze user feedback to improve the user experience of the system. By introducing designs that are more intuitive and in line with user psychological models, the aim is to improve the usability and user satisfaction of the system. Future design optimization will include reorganizing information structures, simplifying operational processes, and adding personalized settings options. Thirdly, enhancing the intelligence level of the system is another important direction for future work. Introducing artificial intelligence and machine learning technologies to enable the system to automatically learn and adapt to user behavior patterns, thereby providing personalized services.

In summary, future research will focus on improving system scalability, optimizing user experience, and enhancing intelligence, addressing data bias issues, promoting interdisciplinary collaboration, and conducting long-term evaluations. Through these efforts, we look forward to further enhancing the adaptability and practicality of the archive management system, providing new research directions for the academic community, and offering effective solutions for practical applications. These future jobs will not only contribute to technological advancements, but also bring broader impacts to society.

VII. Conclusion

Traditional archive retrieval systems are often based on keyword matching and cannot understand users' query intentions and information needs. To improve the performance and user experience of retrieval systems, the improvement of personalized retrieval systems based on user understanding and knowledge graph has attracted the interest of researchers and practitioners. Providing personalized and accurate retrieval services has become particularly important for meeting the needs of users. This study combined knowledge graph with user understanding to construct a personalized retrieval system that meets user preferences. The experimental results showed that the retrieval accuracy and recall rate of knowledge graph based on BiLSTM-CRF were significantly superior to conventional knowledge graphs. The former had retrieval accuracy rates of 92.34%, 88.61%, 90.22%, 90.08%, and 89.34% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, and recall rates of 84.37%, 81.26%, 85.17%, 84.96%, and 82.33%, respectively. The retrieval accuracy and recall rate of the preference semantic retrieval model based on user understanding were significantly superior to conventional semantic retrieval models. Compared with traditional retrieval systems, improving personalized retrieval systems had better accuracy and recall. The improved personalized retrieval system had retrieval accuracy rates of 90.24%, 89.65%, 87.52%, 96.33%, and 95.18% for students, civil servants, demobilized soldiers, law enforcement personnel, and retirees, respectively. Compared with conventional management systems, SAMS had more comprehensive performance, with a maximum response time of 8 seconds and 32 seconds, and the highest user satisfaction of 92.34% and 71.42%, respectively. This indicates that the improved personalized retrieval system has good performance and has good application effects in SAMS.

A SAMS can integrate information from multiple data sources and connect them in a graphical structure. This type of link helps to discover the relationships between data, providing a more comprehensive and comprehensive perspective. Through semantic technology, SAMSs can achieve intelligent search and question answering functions. Users can ask questions through natural language, and the system will understand the meaning of the questions and find relevant information in the graph network, thereby providing accurate answers. A SAMS can help recommendation systems better understand user needs and interests. By analyzing user behavior and preferences, combined with information from the knowledge graph, recommendation systems can provide more personalized and accurate recommendation results. A SAMS can help computers understand and analyze unstructured data such as text and images. By mapping this data to a knowledge graph network, the system can understand the meaning of the data and extract useful information from it.

There are still some shortcomings in this study, such as a small sample size. Subsequent studies will expand the sample size to verify the stability of research accuracy and other indicators.

References

[1] Xu H, Huang C, Wang D. Enhancing semantic image retrieval with limited labeled examples via deep learning. Knowledge-Based Systems, 2019, 163(JAN.1):252-266.

[2] Feiyan Zhang. Thinking on The Information Construction and Standardized Management of University Construction Engineering Archives. Acta Informatica Malaysia. 2023; 7(2): 105-107.

[3] Almousa M, Benlamri R, Khoury R. Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet. Knowledge-based systems, 2021, 212(Jan.5):106565.1-106565.19.

[4] Saleh Ahmed Jalal Siam, Mubashshir Bin Mahbub. Investigating User Perceptions of Ai Technology and Its Ethical Implications on Employment Dynamics and Bias. Acta Informatica Malaysia. 2024; 8(1): 22-25.

[5] Girang Permata Gusti, Hilda. The Potential Use of Quick Response (Qr) Codes in Mobile Banking: An Analysis of Implementation and Its Impact on User Experience. Malaysian E Commerce Journal. 2023; 7 (1): 54-57.

[6] Huang W, Mao Y, Yang Z, L Zhu, J Long. Relation classification via knowledge graph enhanced transformer encoder. Knowledge-based systems, 2020, 206(Oct.28):106321.1-106321.10.

[7] Tuo Shi, Danyang Li, Qi Zhang. Research On the Construction of A Knowledge Graph Of Interaction Risk Between Home Invasion Theft Offenders And Victims Based On Information Extraction Technology. Acta Informatica Malaysia. 2024; 8(1): 01-04.

[8] Li Y, Hou L, Li J. Preference-aware Graph Attention Networks for Cross-Domain Recommendations with Collaborative Knowledge Graph. ACM transactions on information systems, 2023, 41(3):80.1-80.26.

[9] Sui Y, Feng S, Zhang H, J Cao, L Hu, N Zhu. Causality-aware Enhanced Model for Multi-hop Question Answering over Knowledge Graphs. Knowledge-based systems, 2022, 250(Aug.17):1-16.

[10] HX A, SW B, MT A, LW A, XLA C. Knowledge Graph Question Answering with semantic oriented fusion model. Knowledge-Based Systems, 2021, 221(Jun.7):106954.1-106954.10.

[11] Wang Y, Xu X, Hong Q, J Jin, T Wu. Top- k star queries on knowledge graphs through semantic-aware bounding match scores. Knowledge-Based Systems, 2020, 213(2):106655.1-106655.13.

[12] Zou J, et al. (2023), “Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification”, ACM transactions on information systems, Vol. 41 No. 1, pp.1.1-1.25.

[13] Dong X, Zhang H, Dong X, X Lu. Iterative graph attention memory network for cross-modal retrieval. Knowledge-Based Systems, 2021, 226(6):107138.1-107138.12.

[14] Dai Y, Wang S, Chen X, C Xu, W Guo. Generative adversarial networks based on Wasserstein distance for knowledge graph embeddings. Knowledge-Based Systems, 2019, 190(2):105165.1-105165.12.

[15] Liu T, Yu K, Wang L, X Zhang, H Zhou, X Wu. Clickbait detection on WeChat: A deep model integrating semantic and syntactic information. Knowledge-based systems, 2022, 245(Jun.7):108605.1-108605.11.

[16] Hu Y, Zhan W, Tomizuka M. Scenario-Transferable Semantic Graph Reasoning for Interaction-Aware Probabilistic Prediction. IEEE transactions on intelligent transportation systems, 2022, 23(12):23212-23230.

[17] Wozniak P A (2022). Hybrid Electric Vehicle Battery-Ultracapacitor Energy Management System Design and Optimization. Elektronika ir Elektrotechnika, Vol. 28 No. 1, pp.4-15.

[18] Kili S, Firat M (2023), zdemir.A Novel Current Condition Assessment System for Sustainable Management and Operation of Wastewater Treatment Plants.Journal of Environmental Engineering, 2023, Vol. 149 No. 6, pp.241-256.

[19] Du W, et al (2023). Sequential patent trading recommendation using knowledge-aware attentional bidirectional long short-term memory network (KBiLSTM). Journal of Information Science, Vol. 49 No. 3, pp.814-830.

[20] Jin H, Bao Z, Chang X, et al (2023). Semantic segmentation of remote sensing images based on dilated convolution and spatial-channel attention mechanism. Journal of Applied Remote Sensing. Vol. 17 No. 1, pp.016518-1- 016518-17.

[21] Efrén Rama-Maneiro, Vidal J C, Lama M. Collective disambiguation in entity linking based on topic coherence in semantic graphs. Knowledge-Based Systems, 2020, 199(Jul.8):105967.1-105967.14.

[22] Nie Z, Zheng S, Liu Y, Z Chen, S Li, K Lei, F Pan. Automating Materials Exploration with a Semantic Knowledge Graph for Li-Ion Battery Cathodes. Advanced functional materials, 2022, 32(26):2201437.1-2201437.7.

[23] Ayetiran E F, Sojka P, Novotn, Vít. EDS-MEMBED: Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses. Knowledge-Based Systems, 2021, 219(May.11):106902.1-106902.14.

[24] Fares M, Moufarrej A, Jreij E, J Tekli, W Grosky. Unsupervised word-level affect analysis and propagation in a lexical knowledge graph. Knowledge-Based Systems, 2019, 165(FEB.1):432-459.

[25] Zhu Y, Hu L, Ning N, W Zhang, B Wu. A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection. Knowledge-based systems, 2022, 249(Aug.5):108952.1-108952.14.

[26] K.-J. C, Liu Z, Lu H, J Zhang. Heterogeneous graph convolutional network with local influence. Knowledge-based systems, 2022, 236(Jan.25):107699.1-107699.10.

[27] Chen Z, Zhao X, Liao J, X Li, E Kanoulas. Temporal knowledge graph question answering via subgraph reasoning. Knowledge-based systems, 2022, 251(Sep.5):109134.1-109134.11.

[28] Rao V, Dai P, Singla S. Structural fragmentation in scene graphs. Knowledge-Based Systems, 2020, 211(Jan.9):106504.1-106504.11.

[29] Fang Y, Ren Y, Park J H. Semantic-enhanced discrete matrix factorization hashing for heterogeneous modal matching. Knowledge- Based Systems, 2020, 192(Mar.15):105381.1-105381.13.

[30] Fang Y, Luo B, Zhao T, He D, Jiang B, Liu Q. ST-SIGMA: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. CAAI Transactions on Intelligence Technology, 2022, 7(4): 744-757.

[31] Guo Y, Mustafaoglu Z, & Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2022, 2(1), 5–9.

Automatic Recognition and Labeling of Knowledge Points in Learning Test Questions Based on Deep-Walk Image Data Mining

Ying Chang, Qinghua Zhu
Beijing Polytechnic, Beijing 100176, China

***Abstract*—This paper deeply studies and discusses the application of image data mining technology based on the Deep-Walk algorithm in automatic recognition and annotation of knowledge points in learning test questions. With the rapid development of educational informatization, how to effectively mine and label the knowledge points in learning test questions from image data has become an urgent problem to be solved. In this paper, we introduce a novel approach that integrates graph embedding technology with natural language processing techniques. Initially, we leverage the Deep-Walk algorithm to embed the knowledge points present in the test question images, effectively transforming the high-dimensional image data into a low-dimensional vector representation. This transformation meticulously preserves the intricate structural information while meticulously capturing the subtle semantic nuances embedded within the image data. Subsequently, we undertake a thorough semantic analysis of these vectors, seamlessly integrating natural language processing techniques, to facilitate automated recognition with unparalleled precision. This innovative methodology not only elevates the accuracy of knowledge point recognition to new heights but also achieves semantic annotation of these points, thereby furnishing richer, more insightful data support for subsequent intelligent education applications. Through experimental verification, the proposed method has achieved remarkable results on multiple data sets, which proves its feasibility and effectiveness in practical applications. Furthermore, this paper delves into the expansive potential applications of this methodology in the realm of image data mining, encompassing areas such as online education, intelligent tutoring systems, personalized learning frameworks, and numerous other domains. As we look ahead, we aim to refine the algorithm, enhance recognition accuracy, and uncover additional application scenarios, thereby contributing significantly to the intelligent evolution of the education sector.**

***Keywords*—Deep-walk; image data mining; study test questions; knowledge point recognition**

I. Introduction

As human society progresses towards the era of intelligence, education in my country is undergoing a paradigm shift from traditional models to intelligent education. However, the pressing issue of imbalanced and inadequate education development persists. Recently, various government ministries and commissions have enacted a series of pertinent policies [1, 2] to ensure and expedite this educational transformation. These policies aim to harness the power of artificial intelligence, big data, blockchain, and 5G technology to drive innovation and expedite the evolution of digital education. They strive to unlock the latent potential of digital education, explore novel governance approaches, and seize the opportunities presented by the digital revolution for future educational transformation. Additionally, these policies promote the utilization of information technology to innovate teaching methods, develop teaching aids that align with educational needs, and leverage artificial intelligence to provide teachers with comprehensive assistance in tasks such as resource sourcing, homework grading, and online question answering. Ultimately, the goal is to establish an intelligent, efficient, and comprehensive educational analysis system.

In 2006, Hinton published an article titled "Reducing the dimension of data with neural networks" in "Science" magazine, which opened the prelude to the field of deep learning [3, 4]. Since then, research and applications based on deep learning have achieved great success in many fields, such as speech recognition, object visual recognition and target detection. In this disruptive wave of technology, the field of education has also been greatly impacted. Big data and artificial intelligence technologies can be applied to all stages of the field of education, providing support for new learning methods, improving teaching level and teaching quality, and improving Teaching efficiency and teaching effect, reducing the burden on teachers and students. A large number of intelligent education programs have been applied to actual education and teaching, enriching the forms of education and teaching at this stage. With the large-scale development of online education during the epidemic, this form of educational organization based on the Internet has further entered people's lives. Online education actively responds to the challenges brought by the global COVID-19 pandemic to the way education is organized, helping to minimize the impact of the epidemic on normal teaching order. At the same time, online education expands the supply of educational resources, reduces the differences in education levels in different regions, and further realizes educational equity [5, 6].

In the process of carrying out learning and evaluating learning, subject test questions play a vital role. By mining the deep hidden information in online test questions, it can help teachers and students build an intelligent learning environment and reduce the burden. For example, by analyzing the similarity of the hidden information in the test text, it can quickly locate the approximate question type; Using the hidden information of test text to build an intelligent question bank system, further realize a more intelligent and balanced automatic test paper generation model. Among them, the knowledge points labeling of test questions is the basic work of building an intelligent question bank system.

The labeling of knowledge points in test questions refers to the process of labeling the knowledge points used in answering test questions through a certain method. In teaching activities, knowledge points refer to the basic organizational units and transmission units that transmit teaching information, including concepts, definitions and theorems. A knowledge point is a general description of a certain concept, which usually exists as a goal that teachers and students want to achieve together. There are mainly five types of relationships between different knowledge points, namely, hierarchical relationship, dependent relationship, implication relationship, association relationship and dissociative relationship [7, 8]. Hierarchical relationship refers to the form that the upper layer contains the lower layer, and the lower layer belongs to the previous form. These knowledge points are interrelated and usually form a tree structure. According to the tree structure, different knowledge points can be described as a father-son relationship or a brother relationship.

The traditional method of labeling knowledge points of test questions is mainly manual labeling. Usually, front-line teachers with rich teaching and research experience are used as labeling personnel to label knowledge points of test questions. However, artificial knowledge point labeling methods are highly subjective, usually have cognitive biases, and the accuracy is difficult to guarantee. Furthermore, manual labeling fails to leverage the full potential of labeled test question data efficiently, as incremental test question data continues to require time-consuming manual annotation, hindering the pursuit of more efficient solutions. Consequently, amidst the exponential growth of online test question resources, the costs associated with manual knowledge point labeling—both in terms of time and human resources—have skyrocketed, underscoring the pressing need for alternative approaches.

With the application of artificial intelligence technology in various fields, some scholars use deep text mining technology to try to automatically label knowledge points and have achieved certain results. The method based on deep text mining can automatically learn and discover the potential features of test text from the test question data set and can label the appropriate knowledge points for the test questions according to the learned features. Therefore, it can better solve the problem of manual labeling costs caused by the explosive growth of online test question resources, and help automatically build an intelligent test bank system [9, 10].

Most of the existing knowledge point labeling methods of test questions are migrated from general text categorization methods. Test questions refer to a class of question texts used to test teaching effects. They have strong professionalism and certain structure and are different from texts in general fields. Therefore, methods in general fields cannot be simply transferred directly to knowledge point labeling field to complete labeling. Compared with the text in the general field, the test text contains more types of knowledge point labels, the sparsity is higher, and the workload and difficulty of the test question knowledge point labeling task are greater.

II. Overview of Related Technologies and Theories

A. *Text Preprocessing*

Text preprocessing is a necessary step in text categorization task, and its effect affects the effect of text categorization to a certain extent. Therefore, different preprocessing methods are generally adopted according to the needs of text categorization tasks. The text of test questions generally has strong specialization and structure, so it is not easy to transfer the preprocessing method of general domain text to the field of test question text processing directly. Therefore, this paper mainly carries out the following operations in terms of text preprocessing:

1) Data cleansing: Cleanse task-independent text in the original text, and design regular expressions to match and delete this part of the text.

2) Chinese word segmentation: Chinese word segmentation methods are broadly categorized into two primary approaches: character-based and word-based, based on the granularity of segmentation. Each of these methodologies carries its unique set of advantages and constraints, underscoring the importance of selecting the optimal granularity tailored precisely to the demands of a given task. Entity recognition demands a high degree of accuracy in word segmentation, as the precision of this initial step directly impacts the overall effectiveness of these downstream applications. Therefore, these tasks usually choose character-based word segmentation methods [11, 12]. In order to achieve better performance, such tasks as text categorization, sentiment analysis, and text summarization, which pay more attention to text semantic understanding, usually choose word-based word segmentation methods. The process of knowledge point identification and labeling is shown in Fig. 1.

3) De-stop words: The accuracy of text classification can be improved by removing stop words in the text. Stop words refer to the words that appear frequently in categorized texts but have little effect on helping to improve the classification effect. For example, and the land of in English text, these words can be found in almost every sentence, but they do not provide much help in the semantic understanding of the sentence. Studies have shown that in a small English paragraph, more than 50% of the words are contained in a list of 135 commonly used words, which are generally considered noise words and should be removed during the text preprocessing stage [13]. There are also many such words in Chinese text. For example, stop words such as He, Ruo, can provide very little information for text categorization tasks, but often introduce more noise information. Deleting stop words helps significantly reduce the size of the text feature space, speed up model calculation and improve the accuracy of text categorization [14].

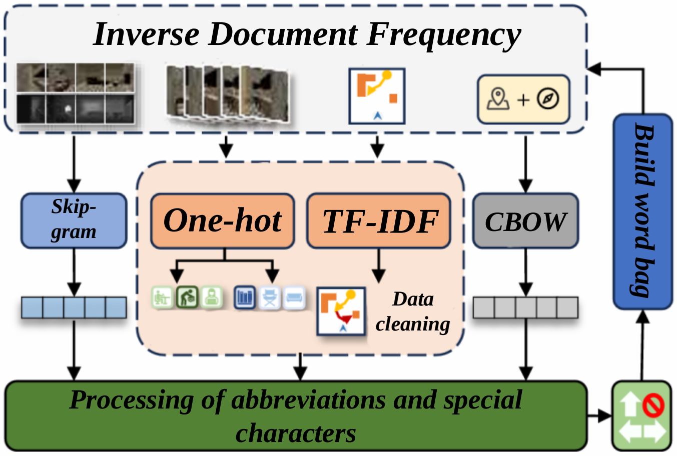

Inverse Document Frequency
Skip-gram
One-hot
TF-IDF
CBOW
Build word bag
Data cleaning
Processing of abbreviations and special characters

Fig. 1. The process of knowledge point identification and labeling.

4) Handling abbreviations and special characters: In the process of daily life and learning, some words are often set as their own abbreviations. The original meaning of these abbreviations is for the convenience of memory, but after abbreviation, part of the semantic information of the original words will be lost, which is not conducive to the learning of classification models [15]. Therefore, it is necessary to convert these abbreviated words in the processing stage to restore their original text expressions. Most text datasets will contain many unnecessary characters, such as punctuation marks and special characters. These punctuation marks and special characters are very important for human wording, sentence breaking, understanding, etc., but these are not helpful to improve the performance of classification algorithms, and will bring a lot of noise to the learning of the model. Choose to remove these punctuation and characters. The node embedding formula in the Deep Walk model is shown in Eq. (1).

$$W(d,t) = TF(d,t) * log\left(\frac{N}{df(t)}\right) \tag{1}$$

Here, N is the number of documents, and df(t) is the number of documents in the corpus that contain the word t. The first term in the equation improves the recall rate, while the second term improves the precision of the word embedding. Although TF-IDF reduces the problems caused by high-frequency words in documents to a certain extent, it also ignores the relationship between words in the text, and directly ignores the semantic information of words.

Word2Vec is a popular word embedding method that captures the relationship between words in context and embeds words into Euclidean space [16]. The Word2Vec method uses two shallow neural networks with continuous word bags (CBOW) and Skip-gram to create a vector for each word in the library. The day scale of the Skip-gram model is to maximize the probability in Eq. (2).

$$\underset{\theta}{argmax} \prod_{w \in T} [\prod_{c \in c(w)} p(c \mid w;\theta)] \tag{2}$$

Fig. 2 shows image data mining and processing process. As Fig. 2 shows, the purpose of the CBOW model and the Skip-gram model are different. While the CBOW model is tasked with finding words based on a sequence of words, the Skip-gram model is tasked with finding the words most likely to appear near a given word based on that word. Word2Vec has greatly advanced the field of natural language understanding by providing a very powerful tool for capturing similarity relationships between words in a corpus [17].

Prior to delving into Word Frequency-Inverse Document Frequency (TF-IDF), it is imperative to grasp the fundamentals of the Bag of Words (BoW) model, which serves as a cornerstone for text representation. The BoW model simplifies text by transforming documents or sentences into a concise list of word frequencies. This list is compiled during the model's construction process, where each unique word in the vocabulary is first encoded into a one-hot encoding vector. For instance, in the given scenario, assuming a vocabulary size of 19, the model would generate a 19-dimensional vector for each word, with a '1' occupying the position corresponding to the word's index in the vocabulary and all other positions set to '0'. This encoding scheme provides a straightforward yet effective way to represent textual data. Then the bag-of-words model combines word frequency as a feature representation of the document and sentence. However, the word bag model only pays attention to the feature of word frequency when collecting and constructing word bags, and ignores the semantic relationship between words, which cannot help the model learn the deep-seated semantic information of the text well [18].

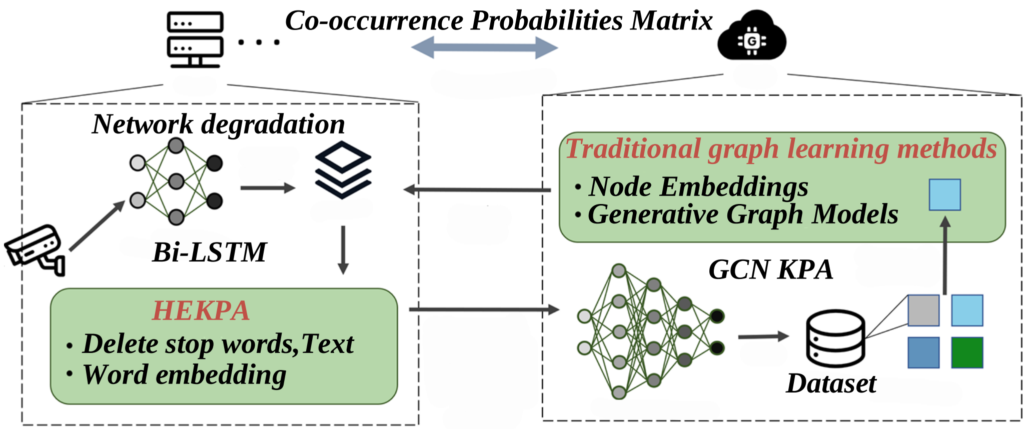


Fig. 2. Image data mining and processing process.

Inverse Document Frequency (IDF) is often used in conjunction with word frequency to reduce the negative impact of frequent words in the corpus. IDF assigns higher weight to high-frequency or low-frequency words in a document: this combination of TF and IDF is called word frequency-inverse document frequency (TF-IDF). Its word embedding is obtained by Eq. (3) as follows:

$$R_{\mathrm{err}} = \sum_{i=1}^{n} \arccos\left(\frac{\mathrm{tr}(R_{\mathrm{out}_i}^{T} R_{\mathrm{gt}_i}) - 1}{2}\right) \tag{3}$$

B. Glove

Another powerful and widely used model is Glove, which calculates word embedding by counting the number of global word co-occurrences across large corpora [19]. Before introducing the Glove model, let's introduce the latent semantic analysis algorithm based on singular value decomposition. This algorithm obtains the vector representation of words and documents by performing singular value decomposition on the word-document matrix. The Glove model combines the ideas and methods of latent semantic analysis algorithm and Word2Vec algorithm, because the author believes that both methods have certain defects. While the Latent Semantic Analysis (LSA) algorithm effectively harnesses global statistical information, its performance in word analogy tasks falls short, suggesting that there is room for enhancement in the generated vectors. In contrast, Word2Vec excels in word analogy tasks, albeit with minimal reliance on corpus statistics, underscoring its unique strengths in this domain [20]. The Glove model combines these two features together, and uses global statistical features and local contextual features of the corpus to help generate a vector representation of the text. For this reason, the Glove model introduces a Co-occurrence Probability Matrix (Co-occurrence Probability Matrix) to achieve this goal. First, the concept of a co-occurrence matrix is introduced. In the co-occurrence matrix X, the rows and columns of the matrix are words in the dictionary. Use xi,j to represent the number of times the word j appears in the context of the word i (usually a window size is set to specify the search distance of the context). The meaning of xi is the number of times the word i appears in the corpus. The co-occurrence probability matrix is obtained by counting the above two values, where the probabilities Pi,j are defined as shown in Eq. (4).

$$P_{ij} = \frac{x_{ij}}{x_i} \tag{4}$$

That is, P is the ratio of the number of times the word appears in the context of word i in the corpus to the total number of times the word i appears. Assuming i = ice, j = steam, k = solid, the feature extraction formula in image data mining is shown in Eq. (5).

$$Ratio = \frac{P_{ik}}{P_{jk}} \tag{5}$$

Using Ratio can also well reflect the relationship between i, j, and k (because the co-occurrence probability Ratio conforms to common sense), so the original author assumed that the word vector of i, j, and k generated by the Glove model can fit this Ratio after some calculation. Make the word vector obtained by Glove consistent with the co-occurrence matrix, so as to reflect the co-occurrence relationship between words, that is, the goal of Glove is defined as shown in Eq. (6).

$$f(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \tag{6}$$

C. Graph Representation Learning

As a classic data structure, Graph is widely used in various fields of natural science. It is generally believed that a Graph is a set of objects (nodes) and interactions (edges) between objects [21]. The nodes in the social network graph are usually used to represent individuals, and the edges inside are used to indicate that there is a certain connection between two people.

Based on graphs, we can analyze, understand and learn complex systems in the real world. In the past dozens of works, many high-quality large-scale graph data have emerged, such as large-scale social network graphs based on social software, knowledge graphs for general fields, Internet network device topology graphs, and so on. The appearance of these large-scale graphs has greatly promoted the development of graph technology, among which the methods based on machine learning are particularly prominent. Machine learning provides many technical means for modeling, analyzing and understanding this part of large-scale and complex graph data, helping people to further explore and discover the theory and knowledge existing in the complex system behind these large-scale graphs.

Before discussing machine learning methods applied to graphs, a formal definition of what exactly means "graph data" is needed. Formally, the graph G=(V, E) consists of a set of nodes V and a set of edges E between these nodes, and represents the edges from node a∈V to node b∈V as (a, b)∈ E. The formula for the ReLU activation function is shown in Eq. (7).

$$\mathrm{E}[D_{\theta}(x_0 + n; \sigma, c) - x_{02}^{2}] \tag{7}$$

A simple way to store a graph is through the adjacency matrix A ∈ RV, where V is the number of nodes. Each row and column of the adjacency matrix represents a specific node; A in adjacency matrix; To represent the-edge from node i to node j, if (i, j)∈E then Ai, j = 1, otherwise Ai, j = 0. The elements Ai, j in the adjacency matrix can also store any real value instead of 0 and 1. At this time, the real value stored by Ai, j is the weight of the edge (i, j)∈E.

Nodes in the graph usually also have their specific attributes or feature information (for example, profiles and pictures of users in social networks), and in most cases, the real-valued matrix F∈Rd is used to represent the nodes. Attributes or features, the order of the nodes in the real-valued matrix is consistent with the order of the rows and rows in the adjacency matrix. The vector representation of the feature is stored in the real-valued matrix, and d is the dimension of the feature vector. The operational formula for the pooling layer and the weight update formula for the fully connected layer are shown in Eq. (8) and Eq. (9).

$$D_w(x; \sigma, c) = wD_1(x; \sigma, c) + (1 - w)D_0(x; \sigma, c) \tag{8}$$

$$\hat{y}_{j,T_n} = \sum_i w_{ij}\hat{y}_{i,T_n} \tag{9}$$

Graph learning algorithms mainly have two stages of development. The first is the traditional graph learning method based on statistics, and the progressive development is the graph representation learning algorithm based on machine learning. Traditional graph learning methods are basically based on the statistical information of nodes and graphs, which requires a lot of feature engineering, so the information is limited. At the same time, the statistical information designed by hand in traditional graph learning methods is not flexible and cannot be adapted in the learning process, so it needs to be redesigned after the task is shifted. With the development of machine learning, a series of methods have emerged that can get rid of manual feature design and learn features in graphs through adaptive methods-graph representation learning. Existing graph representation learning algorithms are mainly divided into three categories: Node Embeddings, GNN, and Generative Graph Models. The research content of this paper mainly involves node embedding and graph neural networks, and does not involve graph generation models. Therefore, the following will focus on the two-graph representation learning algorithms involved in the text.

III. A Knowledge Point Annotation Model Based on Mixed Label Embedding

A. Test Question Text Data

Test questions refer to a class of question texts used to test the teaching effect, so test text is more professional than daily text, and the format of test text is usually relatively fixed, for multiple-choice questions, fill-in-the-blank questions, or answer questions. format. Generally speaking, test text data has the following two characteristics:

1) Professionalism: The text of test questions demands unwavering professionalism, stemming from their intended purpose. The descriptions within these questions must adhere to stringent standards of accuracy and clarity, with no room for ambiguity or vague expressions. When juxtaposed with everyday language, the text of test questions typically encompasses a greater abundance of subject-specific proper nouns, thereby presenting a unique challenge that to some degree complicates the migration of general domain models into this specialized context. The normalization formula and the threshold processing formula in the preprocessing are shown in Eq. (10) and Eq. (11).

$$q_i(v) = \pi(K_i R_i^T(v - t_i)) \tag{10}$$

$$c(p) = \sum_{i=1}^{N-1} I_i(p) w_i(p) \tag{11}$$

2) Structural: Most of the common test text can be divided into a certain type of question, and each type of question has its fixed format, which will lead to more meaningless symbols in the test questions, thus introducing noise to model learning, so it is necessary to remove these meaningless symbols. However, according to the specific task requirements, specific rules can also be used to extract the structure of the test questions, so that it can be used as additional information to help the model learn. Because the test text has strong specialization and a certain structure, it is not easy to transfer the text preprocessing methods in the general field to the test text for text preprocessing without modification. Fig. 3 shows the distribution map of the raw image dataset.

3) Data cleansing: The text of the test questions in the dataset used in this paper is obtained by crawling from Baidu Question Bank through crawlers. The original text in the dataset will contain some text that is irrelevant to the task. For this type of text, this paper designs regular expressions to match and delete this part of the text.

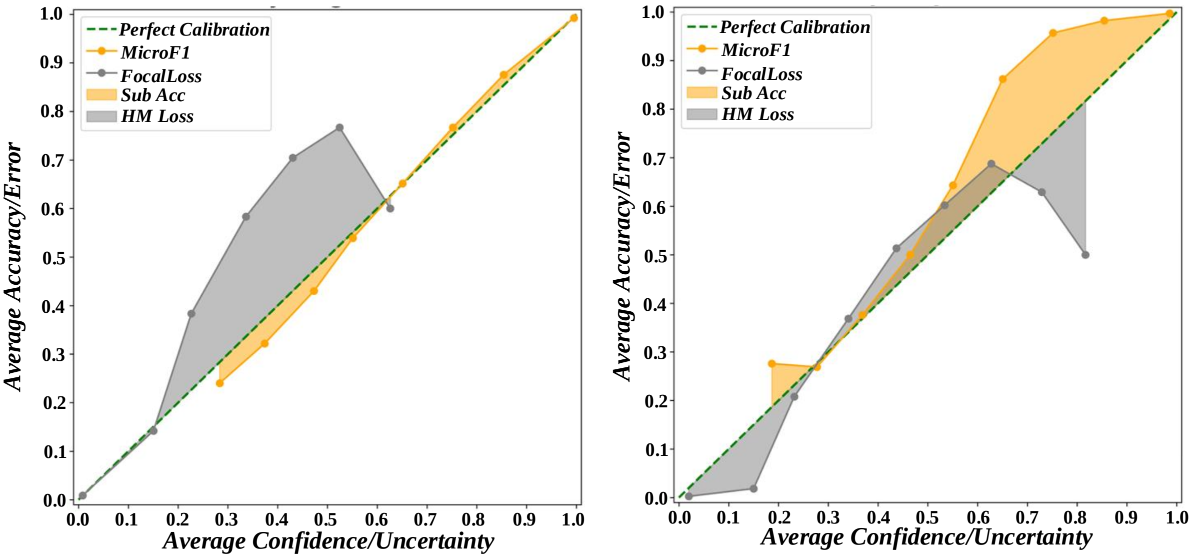


Fig. 3. Distribution map of the raw image dataset.

4) Chinese word segmentation: This paper selects word-based segmentation method to segment the text of the test questions, hoping that the model can better learn the deep-seated semantic information of the text. At present, the research of general Chinese text word segmentation has been relatively mature. There are many word segmentation tools to choose, such as NLPIR, LTP, THULAC, jieba, etc. This paper chooses jieba as a word segmentation tool to obtain word segmentation results. Due to the strong specialization of the test text, the test text usually contains a large number of proper nouns, but the original dictionaries in the existing jieba are designed for general fields, and there will be many mistakes in using the existing jieba directly to segment the test text. Therefore, this paper introduces subject dictionaries to help improve the accuracy of word segmentation and reduce the occurrence of word segmentation errors.

5) De-stop words: One of the basic methods to improve the accuracy of text categorization is to remove the stop words in the text. Stop words refer to the words that appear frequently in categorized texts but have little effect on helping to improve the classification effect. Stop words can provide very little information for text classification tasks, but they often introduce more noise information. Except stop words help to significantly reduce the size of text feature space, help to speed up model calculation and improve the accuracy of text classification. This article includes a list of 859 stop words, which contains most of the Chinese stop words, such as He, Ruo, Yu, Xi, etc.

6) Handling abbreviations and special characters: The inclusion of abbreviations and special characters did not contribute favorably to the model's ability to discern the syntactic and semantic nuances of test questions. Notably, in fields like biology, certain abbreviations of proper nouns are prevalent, designed primarily for mnemonic purposes. However, these abbreviations inherently entail the loss of a portion of the original words' semantic information, which hinders the model's learning process, as it struggles to grasp the full contextual meaning. Therefore, in the text processing stage, it is necessary to convert these abbreviated words to restore their original text expression. In the test text dataset, there are many special characters in addition to the commonly used punctuation marks. These special characters will affect the model's extraction of semantic information to a certain extent, which is not conducive to model learning. This paper will replace these special characters with blank characters in the text preprocessing stage. The gradient calculation formula and the similarity measure formula are shown in Eq. (12) and Eq. (13).

$$\left|B^{l}\right| = K^{2} * F \tag{12}$$

$$L_{ek} = \frac{1}{|P|}\sum_{p\in P}(\|\nabla sdf(p)\|_{2} - 1)^{2} \tag{13}$$

B. Model Construction

Inspired by the dictionary retrieval method, this paper proposes a two-stage automatic labeling model of knowledge points in test questions with mixed label embedding. The model is divided into a classification stage and a labeling stage. In the classification stage, the model classifies the test text into the second level of knowledge points in the knowledge point hierarchy diagram. In the labeling stage, the model obtains a label according to the results of the classification stage and further combines the node embedding and text embedding of the label according to the label-to-label knowledge points [22].

The co-occurrence relationship and the hierarchical relationship between knowledge points can be used as a priori knowledge to guide the model to label. The current method of automatic labeling of knowledge points has the problem of sparse label space. HEKPA guides the model to label knowledge points. The primary consideration is how to obtain the structural relationship.

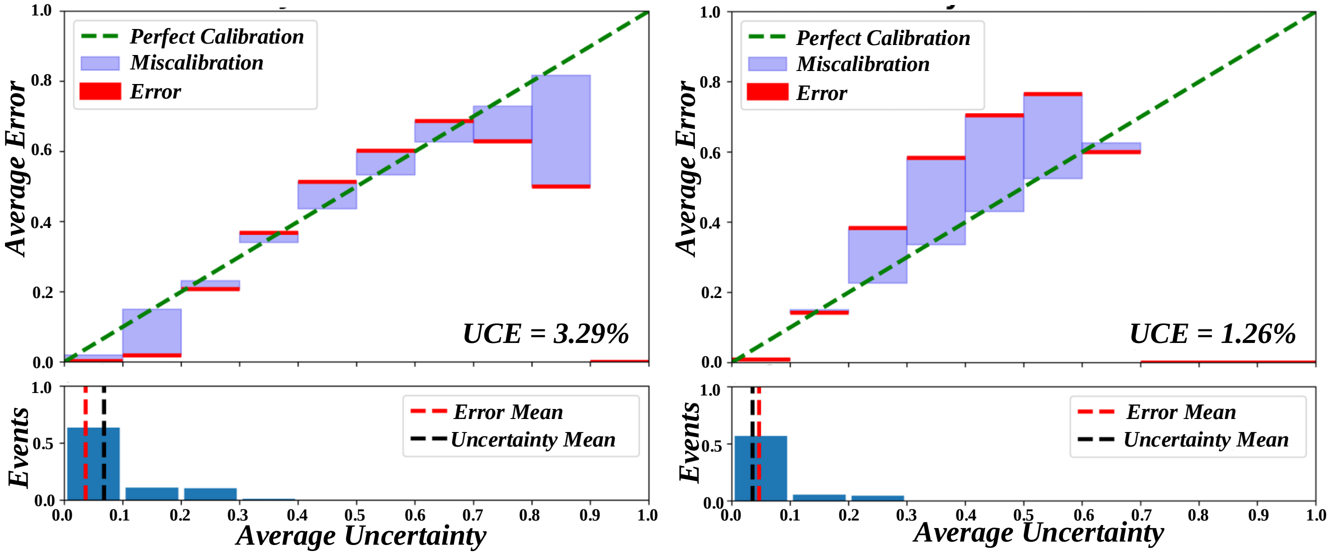


Fig. 4. Comparison diagram of the image feature extraction effect.

Fig. 4 shows comparison diagram of the image feature extraction effect. In Deep-Walk, the random walk algorithm is used to obtain the co-occurrence relationship of nodes in the graph. The results of the random walk with the vertex vi as the starting point are expressed as Wvi, = Wv1, Wv2, …, Wvi, where l is the predetermined random walk step size. After the random walk processes are obtained by using the random walk algorithm, these random walk processes are input into the Word2Vec algorithm as sentence sequences to learn the embedded representation of each node.

IV. A Knowledge Point Labeling Model Based on Graph Convolutional Neural Network

A. Model Construction

To overcome the limitations of shallow node embedding, this chapter presents an end-to-end model, GCN KPA, based on graph convolutional neural networks. This model captures the relationships between knowledge point labels. It features a feature extraction layer using Bi-LSTM to extract text features and a labeling layer that incorporates knowledge point label information using a GNN Network to accomplish labeling tasks.

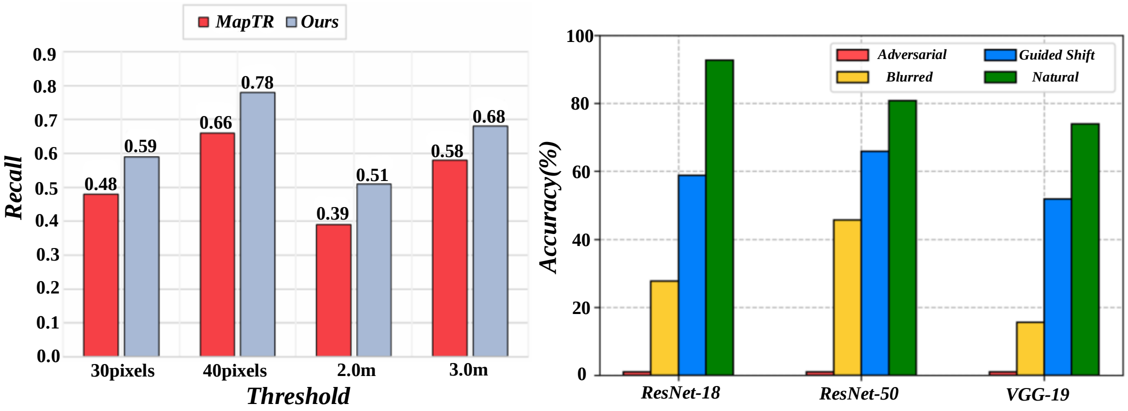


Fig. 5. Deep-walk node embedding visualization graph.

Fig. 5 displays the Deep-Walk Node embedding visualization. GCN-KPA utilizes Bi-LSTM to derive the text representation vector of test questions. Word embedding sequence X, pre-trained by a language model, is processed by Bi-LSTM to extract bidirectional text features. Eq. (14) and Eq. (15) depict random rotation and learning rate decay, respectively.

$$\vec{h}_i = \overline{LSTM}(\vec{h}_{i-1}, x_i) \tag{14}$$

$$L_{mfc} = \frac{1}{K}\sum_{k=0}^{K-1}(1-NCC_k) \tag{15}$$

After acquiring latent semantic hidden states of the text in both directions, they are concatenated to form the final hidden representation of each word. Eq. (16) and Eq. (17) exhibit the dropout layer's mechanism and image classification accuracy calculation, respectively.

$$\hat{C} = \sum_{i=1}^{M} T_i \alpha_i c_i \tag{16}$$

$$T_i = \prod_{j=1}^{i-1}(1-\alpha_j) \tag{17}$$

B. Callout Layer

The structural relationship between labels can help reduce the sparsity of label space and help guide the model to label knowledge points in test questions [23]. In this chapter, we design a classifier based on graph convolutional neural network, which usually has two parts: the eigenmatrix F ∈ Rq representing the nodes of the graph and the adjacency matrix A ∈Rq representing the edges of the graph.

Mula for the confusion matrix and the evaluation index is shown in Eq. (18).

$$G^{(l+1)} = \mathrm{ReLu}(\hat{A}G^{(l)}W^{(l)}) \tag{18}$$

However, only using a simple graph convolutional neural network as the classifier layer of the model will lead to a slower parameter update in the initialization stage, a lower learning rate of the model at the initial stage, and a model that cannot be learned through the back propagation algorithm for a period of time. GCN KPA further introduces Skip Connection to accelerate faster model initialization, the specific definition of the jump connection connection is shown in Eq. (19). The formula of the clustering algorithm in image data mining is shown in Eq. (20).

$$G_s^{(l+1)} = G^{(l+1)} + G^{(l)} \tag{19}$$

$$o = h \odot G_s \tag{20}$$

Fig. 6 shows classification of the model performance evaluation Fig. In a graph convolutional neural network. The calculation for

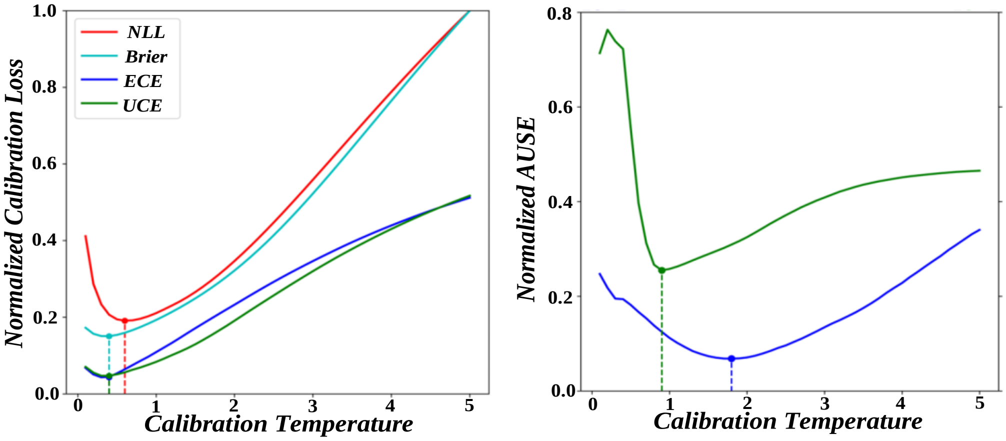


Fig. 6. Classification of the model performance evaluation.

C. Ablation Experiment

Among the three word embedding methods, TF-IDF, Glove and Word2Vec, using Word2Vec to obtain the word embedding representation sequence of text has the greatest improvement to the model, so this chapter defaults to using Word2Vec as the word embedding layer of the model to obtain the word embedding sequence of text [24]. In this section, three sets of ablation experiments were designed:

1) Feature extraction layer ablation experiment: The performance of three feature extraction layers in GCN KPA was studied and compared, with results shown in Fig. 7. Among them, Bi-LSTM performed best, followed by Text CNN, and MLP performed worst. Bi-LSTM extracts latent semantic features of text bidirectionally, hence chosen as the feature extraction layer in this model to enhance annotation accuracy. Fig. 7 displays ablation experiment results.

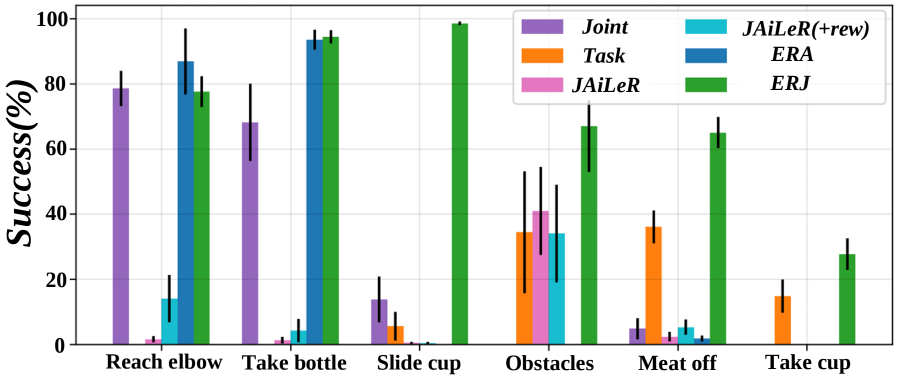


Fig. 7. Results of the ablation experiments of the feature extraction layer.

2) Classifier ablation experiment: Compare the effects of the GCNKPA separator with separate GCN separator and separate FC classifier, results are shown in Table I.

It can be seen from Table I that the model has been greatly improved, which shows that point labels introduced through the graph neural network can be very good [25]. Improve the labeling effect of the model. At the same time, the introduction of Skip Connection further improves the annotation effect of the model. And as shown in Fig. 8, introduce Skip Connection to accelerate model initialization.

TABLE I. COMPARISON OF THE CLASSIFIER EFFECTS

Classifier	Micro F1	Macro F1	H.M Loss	Sub Acc
FC	0.8640	0.7558	0.0116	0.5124
GCN	0.8817	0.8074	0.0100	0.5234
GCN KPA	0.8853	0.8206	0.0097	0.5339

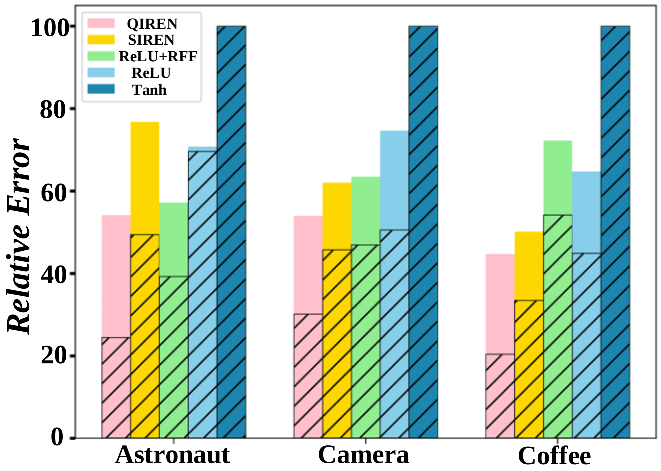


Fig. 8. The annotation effect diagram of the model.

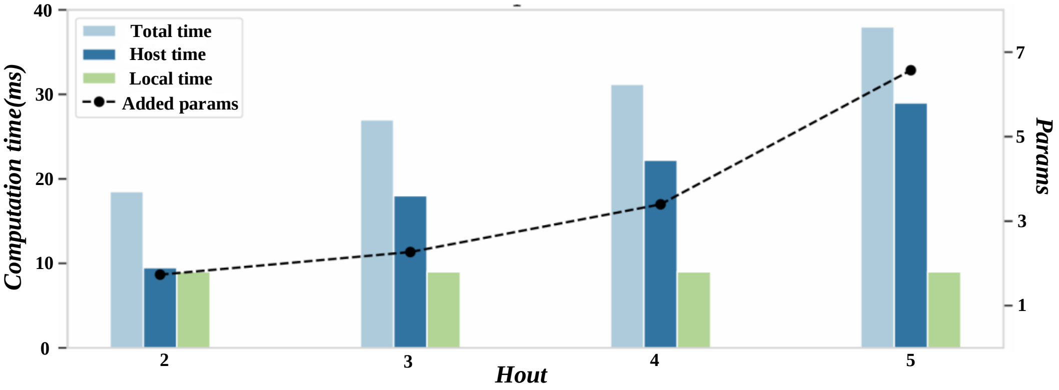


Fig. 9. Comparison diagram of knowledge point identification accuracy and annotation consistency.

Evidently, GCN KPA has demonstrated superior performance across all four evaluation metrics, with a particularly noteworthy enhancement in the MacroF1 indicator, surpassing the gains observed in the other three indicators. This underscores GCN KPA's proficiency in capturing the intricate relationships between labels, enabling it to harness this label information to guide the model's labeling process. Consequently, it mitigates the sparsity of the label space, resulting in a significant improvement in the efficacy of knowledge point labeling [26]. Fig. 9 shows comparison diagram of knowledge point identification accuracy and annotation consistency.

3) Comparison of different Loss functions: The introduction of the Focal Loss function reduces the impact of the unbalanced distribution of labels in the dataset on model learning to a certain extent, and observes the impact of different loss functions on model learning by comparing it with the BCE Loss function commonly used in multi-label text categorization. Detailed experimental results are shown in Table II.

As can be seen from Table II, after replacing BCELoss with Focal Loss, there is an improvement in the three evaluation indicators, but a decrease of 3% in Sub Acc, the most stringent evaluation indicator. This is consistent with the starting point of the Focal loss function design. Focal loss function gives different weights to different samples, reduces the weights of easy-to-classify samples, and allows categories with fewer samples to have higher weights. This makes the model pay more attention to point labels, which is reflected in the experimental results that the improvement rate of MacroF1 is greater than that of MicroF1. At the same time, because the model pays more attention to the part of the label with a small number of samples, the model is more aggressive than before, which leads to the poor performance of the model on the evaluation index Sub Acc. But from the overall experimental results, the introduction of Focal loss makes the model improve most of the evaluation indicators, and can better deal with data sets with unbalanced sample numbers, and better learn knowledge point labels with small sample numbers. Fig. 10 shows performance comparison of the model on different dataset.

TABLE II. EXPERIMENTAL RESULTS FOR THE DIFFERENT LOSS FUNCTIONS

Loss function	Micro F1	Macro F1	HM Loss	Sub Acc
BCE Loss	0.8779	0.8011	0.0106	0.5606
Focal Loss	0.8853	0.8206	0.0097	0.5339

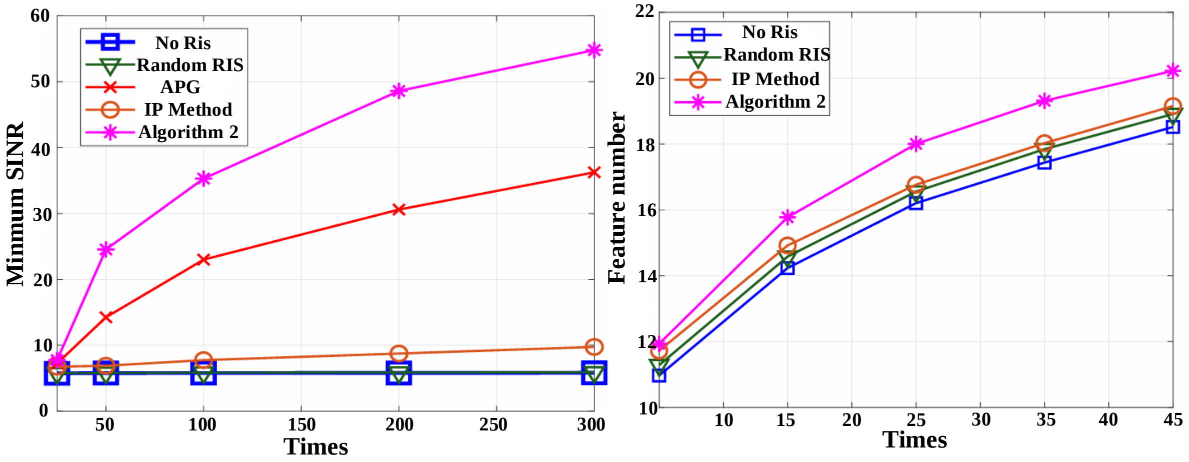


Fig. 10. Performance comparison of the model on different dataset.

V. CONCLUSION AND FUTURE WORK

Aiming at the research of automatic recognition and labeling of knowledge points in learning test questions based on Deep-Walk image data mining, a new method is proposed in this paper. This method combines graph embedding technology and advanced concepts in the field of natural language processing, and realizes the effective recognition and labeling of knowledge points in test questions in image data mining. By deeply exploring the application of Deep-Walk algorithm in graph data embedding, we successfully combine the visual features of images with the knowledge points of test questions, and realize the in-depth analysis and semantic understanding of image data.

The main conclusions of this study are as follows: First, Deep-Walk algorithm shows strong potential in the field of image data mining. It captures the topological structure of images through random walks, and then generates low-dimensional dense vector representations, which provides an effective means for the recognition of knowledge points in test questions. Secondly, combining the visual features of image data with the knowledge points of test questions cannot only improve the accuracy of recognition, but also enhance the semantic richness of labeling, providing strong support for intelligent applications in the field of education.

Anticipating the future, as image data mining technology continues to evolve, the Deep-Walk-based approach for knowledge point recognition and annotation is poised to find broader applications across diverse fields. Concurrently, we remain committed to delving deeper into the realm of algorithms and models, with the aim of enhancing recognition accuracy and refining the granularity of labeling, thereby pushing the boundaries of this technology further. In addition, we will also focus on the fusion of image data and other types of data, as well as cross-domain knowledge migration and sharing, laying a solid foundation for the realization of a wider range of intelligent educational applications.

To sum up, the research on automatic recognition and labeling of knowledge points in image data mining learning test questions based on Deep-Walk has achieved remarkable results, which provides a new idea and method for the intelligent development of education.

REFERENCES

[1] Bian, C., & Lu, S. Personalized recommendation of entertainment robots in fine arts education based on human–computer interaction and data mining. Entertainment Computing, vol. 51, pp. 100740, 2024.

[2] Cap, Q. H., Fukuda, A., Kagiwada, S., Uga, H., Iwasaki, N., & Iyatomi, H. Towards robust plant disease diagnosis with hard-sample re-mining strategy. Computers and Electronics in Agriculture, vol. 215, pp. 108375, 2023.

[3] Cerezo, R., Lara, J.-A., Azevedo, R., & Romero, C. reviewing the differences between learning analytics and educational data mining: Towards educational data science. Computers in Human Behavior, vol. 154, pp. 108155, 2024.

[4] Duque, J., Godinho, A., Moreira, J., & Vasconcelos, J. Data Science with Data Mining and Machine Learning A design science research approach. Procedia Computer Science, vol. 237, pp. 245–252, 2024.

[5] Gonzalez, L. F. P., Pivel, M. A. G., & Ruiz, D. D. A. Improving bathymetric images exploration: A data mining approach. Computers & Geosciences, vol. 54, pp. 142–147, 2013.

[6] Guo, Z., Yang, G., Wang, D., & Zhang, D. A data augmentation framework by mining structured features for fake face image detection. Computer Vision and Image Understanding, vol. 226, pp. 103587, 2023.

[7] Jiang, W., Yu, D., Xie, Z., Li, Y., Yuan, Z., & Lu, H. Trimap-guided feature mining and fusion network for natural image matting. Computer Vision and Image Understanding, vol. 230, pp. 103645, 2023.

[8] Jindal, K., & Kumar, R. A Note on “Data mining based noise diagnosis and fuzzy filter design for image processing”. Computers & Electrical Engineering, vol. 49, pp. 50–51, 2016.

[9] Kasat, N. R., & Thepade, S. D. Novel Content Based Image Classification Method Using LBG Vector Quantization Method with Bayes and Lazy Family Data Mining Classifiers. Procedia Computer Science, vol. 79, pp. 483–489, 2016.

[10] Lin, G., Wei, W., Kang, X., Liao, K., & Zhang, E. Deep graph layer information mining convolutional network. Pattern Recognition, vol. 154, pp. 110593, 2024.

[11] Liu, Y., Hu, S., Zhang, H., Dong, Q., & Liu, W. Intelligent mining methodology of product field failure data by fusing deep learning and association rules for after-sales service text. Engineering Applications of Artificial Intelligence, vol. 133, pp. 108303, 2024.

[12] Marshoodulla, S. Z., & Saha, G. A survey of data mining methodologies in the environment of IoT and its variants. Journal of Network and Computer Applications, vol. 228, pp. 103907, 2024.

[13] Raj, M. P., & Saini, J. R. A Novel Comparison of Charotar Region Wheat Variety Classification Techniques using Purely Tree-based Data Mining Algorithms. Procedia Computer Science, vol. 235, pp. 568–577, 2024.

[14] Tang, W. (2Application of support vector machine system introducing multiple submodels in data mining. Systems and Soft Computing, vol. 6, pp. 200096, 2024.

[15] Wang, C., Wang, G., Zhang, Q., Guo, P., Liu, W., & Wang, X. Eliminating and mining strategies for open-world object proposal. Neurocomputing, vol. 599, pp. 128026, 2024.

[16] Wang, Y., Wu, G., Chen, G. (Sheng), & Chai, T. Data mining based noise diagnosis and fuzzy filter design for image processing. Computers & Electrical Engineering, vol. 40(7), pp. 2038–2049, 2014.

[17] Xu, C., Lin, R., Cai, J., & Wang, S. Deep image clustering by fusing contrastive learning and neighbor relation mining. Knowledge-Based Systems, vol. 238, pp. 107967, 2022.

[18] Yang, Y., Lin, H., Guo, Z., & Jiang, J. A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis. Computers & Geosciences, vol. 3(1),pp. 20–30, 2007.

[19] Zhang, J., & Dong, L. RETRACTED: Image Monitoring and Management of Hot Tourism Destination Based on Data Mining Technology in Big Data Environment. Microprocessors and Microsystems, vol. 80, pp. 103515, 2021.

[20] Zhang, J., Gruenwald, L., & Gertz, M. VDM-RS: A visual data mining system for exploring and classifying remotely sensed images. Computers & Geosciences, vol. 35(9), pp. 1827–1836, 2009.

[21] Zhang, K., Chen, K., & Fan, B.Massive picture retrieval system based on big data image mining. Future Generation Computer Systems, vol. 121, pp. 54–58, 2021.

[22] Zhang, X., Gu, N., Chang, J., Ye, H., Lin, C., & Shen, J.Mining discriminative spatial cues for aerial image quality assessment towards big data. Signal Processing: Image Communication, vol. 80, pp. 115646, 2020.

[23] Zhang, Z., Ning, L., Liu, Z., Yang, Q., & Ding, W. Mining and reasoning of data uncertainty-induced imprecision in deep image classification. Information Fusion, vol. 96, pp. 202–213, 2023.

[24] Dol, S. M., & Jawandhiya, P. M. Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey. Engineering Applications of Artificial Intelligence, vol. 122, pp. 106071.

[25] El-Gharib, N. M., & Amyot, D. Robotic process automation using process mining—A systematic literature review. Data & Knowledge Engineering, vol. 148, pp. 102229, 2023.

[26] Wang, Z., Zhang, F., Ren, M., & Gao, D. A new multifractal-based deep learning model for text mining. Information Processing & Management, vol. 61(1), pp. 103561, 2024.

Heart-SecureCloud: A Secure Cloud-Based Hybrid DL System for Diagnosis of Heart Disease Through Transformer-Recurrent Neural Network

Talal Saad Albalawi
College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh 11432, Saudi Arabia

Abstract—**Cardiovascular disease (CVD) has rapidly increased after COVID-19. Several computerized systems have been developed in the past to diagnose CVD disease. However, the high computing expenses of deep learning (DL) models and the complexity of architectures are significant issues. Therefore, to resolve these issues, an accurate diagnosis of CVD disease is required. This paper proposes a hybrid and secure deep learning (DL) system known as Heart-SecureCloud to predict multiclass heart diseases. To develop this Heart-SecureCloud system, four major stages are makeup such as preprocessing and augmentation, feature extraction and transformation, deep learning and hyperparameter optimization, and cloud security. Advanced signal processing and augmentation technologies are applied to ECG data in the preprocessing and augmentation step to enhance data quality. In the feature extraction and transformation step, adaptive wavelet transforms, and feature scaling are used to extract and convert spectral and temporal data. The DL and hyperparameter optimization step utilize a novel hybrid transformer-recurrent neural network model, which is further optimized for accuracy and efficiency using hyperband-GA. Transfer learning refines pre-trained models using domain-specific data. The unique aspect of the Heart-SecureCloud system is its implementation through a secure cloud, which safeguards medical data with encryption and access control mechanisms. The system's efficacy is demonstrated through testing and evaluation on three publicly available datasets, such as MIT-BIH Arrhythmia MIMIC-III Waveform and PTB-ECG. The Heart-SecureCloud DL architecture achieved impressive results of 98.75% of accuracy, 98.80% of recall, 98.70% of precision, and 98.75% of F1-score. Moreover, the Heart-SecureCloud DL underscores its promise for safe medical diagnostics deployment.**

Keywords—Heart disease diagnosis; deep learning; cloud computing; feature extraction; data security; hyperparameter optimization; encryption

I. Introduction

The most common chronic disorders worldwide are cardiovascular diseases (CVDs), which have caused the most morbidity and mortality during the previous decade [1]. The WHO estimates that 17.9 million people die from CVDs yearly, 32% of all fatalities [2]. By 2030, 22.2 million individuals may die from CVDs. Over the past 30 years, CVDs have been the leading cause of death in the US, accounting for 46.2% of deaths in 2017 [3]. CVDs include congestive heart failure, coronary artery disease, congenital heart defects, cerebrovascular disease, and rheumatic heart disease [4]. Nowadays, CVD is caused by heart attacks and strokes. Early and precise prediction of CVD disease improves survival and reduces death [5]. In addition, this improvement can assist experts in treating patients faster, thanks to the potential of machine learning (ML) and deep learning (DL) methods [6]. These methods, by analyzing ECG signals, can significantly enhance our ability to combat CVDs [7].

Artificial intelligence (AI) technologies are advancing rapidly, and cloud security, along with machine learning (ML) and deep learning (DL) approaches, can now be utilized to monitor and even predict cardiovascular diseases (CVD) [8]. Cloud security, a crucial component, involves securing data, applications, and infrastructures hosted in the cloud, ensuring they are protected from unauthorized access and breaches. This is particularly important in the medical and healthcare sectors, where sensitive health data must be safeguarded. Machine learning, a branch of artificial intelligence, involves techniques that extract knowledge from data, often called predictive analytics or statistical learning. Deep learning, a subset of ML, uses neural networks with multiple layers to model complex patterns in data. These techniques, when applied to medicine, have the potential to not just revolutionize but also excite us about the future of healthcare delivery methods. Moreover, the vast amount of data generated by hospitals in a cloud environment presents significant challenges, particularly in selecting the most effective machine-learning techniques for data analysis.

Cardiovascular disease (CVD) has seen a significant rise following the Covid-19 pandemic. In response, numerous computerized systems have been developed to diagnose CVD. However, challenges such as the high computational costs of deep learning (DL) models and the complexity of their architectures remain. To address these challenges, there is a need for an accurate and efficient approach to diagnosing CVD. This paper introduces a hybrid and secure deep learning system called Heart-SecureCloud, designed to predict various types of heart diseases. This study shows multi-layered strategy to establishing a safe and efficient cloud-based deep learning system for heart disease diagnostics. In the Preprocessing and Augmentation Layer, innovative signal processing and data augmentation procedures improve medical voice record input data quality. A unique adaptive wavelet transform, and feature scaling method extracts and transforms spectral and temporal properties in the Feature Extraction and Transformation Layer.

The Hyperband-GA hybrid approach optimizes a Transformer-Recurrent Neural Network (RNN) hybrid model in the Deep Learning and Hyperparameter Optimization Layer for accuracy and efficiency. Pre-trained models are fine-tuned using domain-specific data via transfer learning. Finally, the Evaluation and Security Layer thoroughly evaluates and verifies performance metrics while protecting sensitive medical data with strong encryption and access control. Encrypting data in transit and at rest and using authentication techniques, this layer secures data processing and cloud server deployment. This study makes several significant contributions to the field of heart disease detection.

1) Novel Heart-SecureCloud DL system for effective heart disease diagnosis using advanced signal processing, feature extraction, and hybrid deep learning architectures.

2) The Hyperband-GA hybrid optimization approach improves model accuracy and computational efficiency.

3) This cloud server study integrates a thorough security layer into the cloud-based feature categorization system.

4) High model accuracy (98.75%) and comprehensive security features set a new heart disease diagnosis system benchmark.

Structure of the paper: Section II: Reviews heart disease diagnosis and prognosis approaches and their advances. Section III: Explains speech feature extraction, ML methods, picture augmentation, and data normalization. Section IV reports the experimental setup, findings, and performance evaluation of the proposed system, comparing it to alternative methods. Section V: Summarizes findings, analyzes ramifications, and offers further study.

TABLE I. A TABLE SUMMARIZING THE KEY POINTS AND COMPARISONS BASED ON THE LITERATURE REVIEW

Study	Purpose	Methodology	Results	Limitations
[14]	Detection of CAD using ECG signals	Developed AE, RBFN, SOM, and RBM models; ensemble of AE and SOM	AE: Accuracy 0.974 (MIT-BIH), 0.984 (PTB-ECG); Ensemble: Accuracy 0.984 (MIT-BIH), 0.992 (PTB-ECG)	Needs testing on larger and more imbalanced datasets
[15]	Automated diagnostic systems for CAD, MI, CHF	Developed 16-layer LSTM model	Accuracy 98.5%	Limited to classification of abnormal ECG signals
[16]	Addressing imbalanced data for detection	Developed GAN, LSTM, and ensemble GAN-LSTM models	GAN-LSTM: Accuracy 0.992 (MIT-BIH), 0.994 (PTB-ECG)	Further research needed with different ensemble models and datasets
[17]	Automated detection of ECG arrhythmia	Removed noise, extracted features, used ML and DL models	Accuracy 86.25%	Performance affected by noise in ECG signals
[18]	Distinguish normal and abnormal ECG patients	Used SVM, LR, AdaBoost; ensemble of AdaBoost and LR	Ensemble: Accuracy 0.946 (PTB-ECG), 0.921 (MIT-BIH)	Methodology can be applied to other diseases
[19]	Preprocessing, data sampling, feature extraction, classification	Used ADASYN for data sampling, GRU for feature extraction, ELM for classification	Superior in terms of accuracy, sensitivity, specificity	Needs further validation with other datasets
[20]	Simplify large data processing	Used Spark–Scala tools, evaluated with MIT-BIH datasets	GDB Tree: Accuracy 97.98% (binary), Random Forest: 98.03% (multi-class)	Limited to large-scale data processing tools
[21]	Compare 1D-CNN and SVM algorithms	Merged public ECG databases, evaluated performance	1D-CNN: Accuracy 93.07%, SVM: Accuracy 92.00%	Need for broad datasets to evaluate ML models
[22]	Recognize various cardiac arrhythmias	Developed ML-WCNN combining 1D-CNN and SWT	Superior performance with 10-fold cross-validation	Limited comparison with state-of-the-art algorithms
[23]	Improve patient prognostics of heart disease	Developed EDCNN, validated on IoMT platform	Precision up to 99.1%	Needs further clinical validation
[24]	Classify MI based on ECG signals	Developed DenseNet and CNN models	DenseNet preferred, Accuracy >95%	Requires more explainability for clinical acceptance
[25]	Determine best combination of signal information	Used raw ECG signals, entropy-based features, QRS complexes	Improved performance with combined features	Performance varies with different signal combinations
[26]	Automate detection and classification of arrhythmias	Developed 2D-CNN-LSTM model	Accuracy ≈98.7% (ARR), 99% (CHF, NSR)	Future work needed on live ECG signals
[27]	Classify CAD, MI, CHF using CNN and GaborCNN	Balanced dataset, evaluated models	High accuracy >98.5%	Needs validation with larger database
[28]	Automated MI detection using ECG signals	Developed CNN, hybrid CNN-LSTM, ensemble techniques	Ensemble: Accuracy 99.89%	Ready for clinical application
[29]	Compare transfer learning methods for ECG classification	Used ResNet50, AlexNet, SqueezeNet	Accuracy 98.8% (AlexNet)	Time-consuming with multiclassification
[30]	ECG beat classification using VGG16-based CNN	Applied SHAP for interpretability	Accuracy 100% (2-4 classes), 99.90% (5 classes)	Needs application in clinical settings
[31]	Predict arterial events using ECG recordings	Used LSTM-DBN, compared with other models	Accuracy 88.42%	Needs further validation with real-world data

II. LITERATURE REVIEW

Advanced algorithms in deep learning have improved heart disease detection systems by analyzing complicated medical data. This literature review addresses deep learning-based heart disease detection methods, their usefulness, and their obstacles.

Deep learning models, especially CNNs and LSTM networks, have improved arrhythmia diagnosis from electrocardiogram (ECG) readings. CNNs are suitable for image and signal processing because they capture spatial hierarchy. Recent advancements in deep learning have shown that Convolutional Neural Networks (CNNs) are particularly effective for image and signal processing due to their ability to capture spatial hierarchies. Studies by [9], [10] demonstrated that combining CNNs with Long Short-Term Memory (LSTM) networks enhances diagnostic accuracy by capturing spatial and temporal data. Transfer learning (TL), which fine-tunes models pre-trained on large datasets for specific, smaller datasets, has also been shown to improve model generalization and reduce computational demands. In study of [11] and [12], the authors successfully applied TL to identify heart disease from ECG data. Additionally, the potential of adaptive wavelet transformations and sparse autoencoders in feature extraction and augmentation is vast, giving us hope for the future of medical data analysis [13]. Furthermore, four DL models are utilized in study [14] to diagnose coronary artery disease (CAD) using ECG data. DL approaches were preferred for automated diagnostic systems [15]. The paper's 16-layer LSTM model evaluated using 10-fold cross-validation, classified ECG signals achieved 98.5% accuracy.

Generative Adversarial Network (GAN) models were used to generate more data to balance skewed data [16]. For MIT-BIH and PTB-ECG datasets, the GAN-LSTM ensemble model performed best, with an accuracy of 0.992 and 0.994, respectively. Other ensemble methods and datasets might improve detection performance in future studies. A CAD method for ECG-based apnea diagnosis was proposed in the study [17] to simplify automated ECG arrhythmia identification. After removing noise with a Notch filter, the system retrieved features and used ML and DL models to diagnose. The suggested model detected obstructive sleep apnea with 86.25% accuracy. In study [18], an uneven number of ECG samples was utilized to identify normal and abnormal individuals. SVM, LR, and AdaBoost were used. The ensemble model with AdaBoost and LR performed best, with PTB-ECG accuracy of 0.946 and MIT-BIH accuracy of 0.921. This approach might be used for various illnesses with different signal inputs.

The CIGRU-ELM model [19] required preprocessing, data sampling, feature extraction, and classification. The class imbalance was resolved via ADASYN, GRU feature extraction, and ELM classification on the PTB-XL dataset. It excelled in accuracy, sensitivity, specificity, and other parameters, demonstrating its flexibility. A study in [20] examined Spark–Scala tools for massive dataset processing. GDB Tree and Random Forest methods gave the model 97.98% binary classification accuracy and 98.03% multiclass classification accuracy utilizing MIT-BIH datasets. This proved Spark–Scala's large-data handling ability.

Both 1D-CNN and SVM algorithms performed well in research [21] utilizing combined ECG datasets. The 1D-CNN method was 93.07% accurate, whereas the SVM classifier was somewhat lower. Combining datasets from diverse sources helped evaluate ML models. The Multi-Level Wavelet Convolutional Neural Network (ML-WCNN) in the study [22] recognized cardiac arrhythmias. The ML-WCNN used 1D-CNN and SWT for feature extraction and achieved improved performance with 10-fold cross-validation accuracy.

The Enhanced Deep Learning aided Convolutional Neural Network (EDCNN) was suggested to enhance heart disease prognostics [23]. The Internet of Medical Things (IoMT) platform enabled the EDCNN to reach 99.1% accuracy, indicating its clinical promise. ECG-based DenseNet and CNN models classified myocardial infarction (MI) [24]. Highly performing DenseNet beat CNN in computational complexity and classification accuracy. This study also revealed certain ECG leads that influence prediction choices. The study in [25] investigated the optimal signal information for categorization. Adding entropy-based features and extracted QRS complexes to raw ECG signals enhanced performance, demonstrating the benefits of using them in ECG analysis.

A hybrid deep learning-based 2D-CNN-LSTM technique was presented for cardiac arrhythmia detection and classification [26]. The model was useful due to its excellent accuracy, sensitivity, and specificity. Future studies might use Bi-LSTM instead of LSTM on real ECG data. CNN and GaborCNN models classified CAD, MI, and CHF in the study [27]. GaborCNN was picked for its excellent classification accuracy and low computing complexity. This technique might be clinically validated with larger databases. CNN, hybrid CNN-LSTM, and ensemble methods were used to create an automated MI detection system [28]. The models have great classification accuracy using SMOTE-Tomek Link for data balancing, suited for hospital use. ECG classification transfer learning techniques were compared [29]. CAA-TL employing ResNet50, AlexNet, and SqueezeNet exhibited outstanding accuracy, suggesting transfer learning improves heart disease diagnosis. A modified VGG16-based CNN-based ECG beat classifier was proposed in the study [30] and achieved good accuracy. SHAP values improved ECG interpretability, making this model suitable for automated cardiovascular diagnosis. A study [31] predicted vascular events from ECGs using LSTM-DBN. The algorithm outperformed deep learning and standard classification approaches, suggesting early cardiovascular event diagnosis and prevention.

Despite advances as described in Table I, many problems remain. High computing expenses of deep learning models and hybrid architectural complexity are major obstacles. Despite their great accuracy, these models need refinement to handle different and unexplored data. Further study should optimize computing efficiency via model compression and more efficient methods. These models might be strengthened by adding medical imaging and patient history data. These systems must have real-time processing and continual learning to be relevant in changing healthcare situations. Continuous security improvements will secure patient data, encouraging confidence and privacy compliance. Finally, numerous models and procedures using deep learning to identify cardiac disease have increased diagnostic accuracy and efficiency. Research and development on computational optimization, feature integration, and security will improve heart disease diagnostic technologies and make them more reliable and accessible.

III. Research Methodology

Fig. 1 shows this study's multi-layered strategy to establishing a safe and efficient cloud-based deep learning system for heart disease diagnostics. In the Preprocessing and Augmentation Layer, innovative signal processing and data augmentation procedures improve medical voice record input data quality. A unique adaptive wavelet-transform, and feature scaling method extracts and transforms spectral and temporal properties in the Feature Extraction and Transformation Layer. The Hyperband-GA hybrid approach optimizes a Transformer-Recurrent Neural Network (RNN) hybrid model in the Deep Learning and Hyperparameter Optimization Layer for accuracy and efficiency. Pre-trained models are fine-tuned using domain-specific data via transfer learning. Finally, the Evaluation and Security Layer thoroughly evaluates and verifies performance metrics while protecting sensitive medical data with strong encryption and access control. Encrypting data in transit and at rest and using authentication techniques, this layer secures data processing and cloud server deployment.

Algorithm 1: Overall secure and efficient heart disease detection system

```
[Input] ECG data
[Output] heart disease diagnosis with high accuracy and secure data handling
Compute
    Load necessary libraries and dependencies.
    Initialize cloud server for scalable deployment.
    Generate encryption keys for data security.
While () do
    For (every ECG class) do
    Update
        Preprocessing and Augmentation:
        Input raw ECG recordings.
        Apply noise reduction techniques.
        Normalize the recordings.
        Segment the recordings into smaller parts.
        Feature Extraction and Transformation:
        Input preprocessed and augmented recordings.
        Extract spectral features using adaptive wavelet transforms.
        Extract temporal features.
        Apply feature scaling techniques
        Deep Learning and Hyperparameter Optimization:
        Input extracted and transformed features.
        Initialize hybrid deep learning model (Transformer + LSTM).
        Apply transfer learning to fine-tune pre-trained models.
        Optimize the model using Hyperband-GA technique.

        Update and analyze
            If (condition) then
                Train the optimized model on the dataset.
            End
    End
    Deploy trained model and security mechanisms on cloud server.
    Set up real-time processing capabilities.
    Ensure automatic scaling for increased loads
End
```

A. Data Acquisitions

The Heart-SecureCloud system was trained and tested using three datasets. PhysioNet's MIT-BIH Arrhythmia database [32] contains annotated ECG recordings utilized in cardiovascular disease detection studies. Second, the PhysioNet's MIMIC-III Waveform database [33] contains ICU patients' ECGs and other physiological waveforms, which may be used to design and evaluate cardiovascular disease detection algorithms. Third, PhysioNet's PTB Diagnostic ECG [34] collection includes 549 ECG recordings from healthy volunteers and cardiac disease patients, including myocardial infarction. Table II describes the detailed parameters of each dataset.

In cardiovascular research, the MIT-BIH Arrhythmia Database from PhysioNet is commonly used for arrhythmia identification. Annotated electrocardiogram (ECG) recordings from varied patients are included. Each recording is properly annotated with arrhythmia annotations, helping build and validate cardiac rhythm problem detection algorithms. Also, the MIMIC-III data is used to study many cardiovascular diseases. The PTB Diagnostic ECG Database, available through PhysioNet, has 549 ECG recordings from healthy people and patients with cardiac problems, including myocardial infarction. From the above three datasets, Table II describes the details about the ECG dataset.

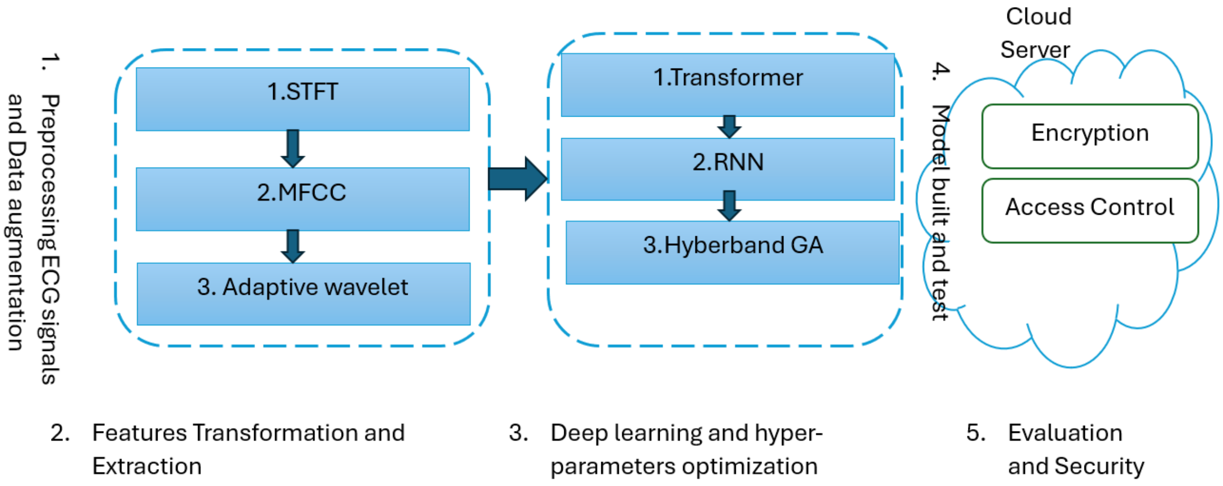

1. Preprocessing ECG signals and Data augmentation
1.STFT
2.MFCC
3. Adaptive wavelet
1.Transformer
2.RNN
3.Hyberband GA
4. Model built and test
Cloud Server
Encryption
Access Control
2. Features Transformation and Extraction
3. Deep learning and hyper-parameters optimization
5. Evaluation and Security

Fig. 1. A systematic flow diagram of proposed system for cardiovascular disease detection.

TABLE II. DATA FROM THE MIT-BIH ARRHYTHMIA, MIMIC-III, AND PTB ECG DATASETS FOR PREDICTING HEART DISEASES

Dataset	Sample ID	Source	Patient ID	ECG Lead Type	Sampling Rate (Hz)	Duration (s)	Diagnosis/Label	Annotation Format
MIT-BIH Arrhythmi [32]	MITBIH-Sample-1	MIT-BIH Arrhythmia	100	Lead II	360	30	Arrhythmia Type	AAMI ECG Codes
	MITBIH-Sample-2	MIT-BIH Arrhythmia	101	Lead V1	360	30	Normal, Atrial Fibrillation	AAMI ECG Codes
MIMIC-III Waveform [33]	MIMIC-Sample-1	MIMIC-III Waveform	201	Lead II	125	60	Various Cardiac Events	Custom Annotations
	MIMIC-Sample-2	MIMIC-III Waveform	202	Lead V5	125	60	Heart Failure, Myocardial Infarction	Custom Annotations
PTB Diagnostic ECG [34]	PTB-Sample-1	PTB Diagnostic ECG	301	Lead II	1000	10	Myocardial Infarction, Healthy	SCP-ECG Codes
	PTB-Sample-2	PTB Diagnostic ECG	302	Lead III	1000	10	Myocardial Ischemia, Healthy	SCP-ECG Codes

Three ECG datasets—MIT-BIH Arrhythmia, MIMIC-III Waveform, and PTB Diagnostic ECG—cover Normal/Healthy, Atrial Fibrillation, Ventricular Tachycardia, Myocardial Infarction, Premature Ventricular Contraction, Heart Failure, and Left Bundle Branch Block. These common heart diseases required early treatment and diagnosis. Fig. 2 and Fig. 3 show distribution plots of sampling rates by datasets. Whereas Fig. 4 shows the length of ECG recordings for each sample, grouped by diagnosis. Fig. 5 provides ECG samples for each cardiac sequence.

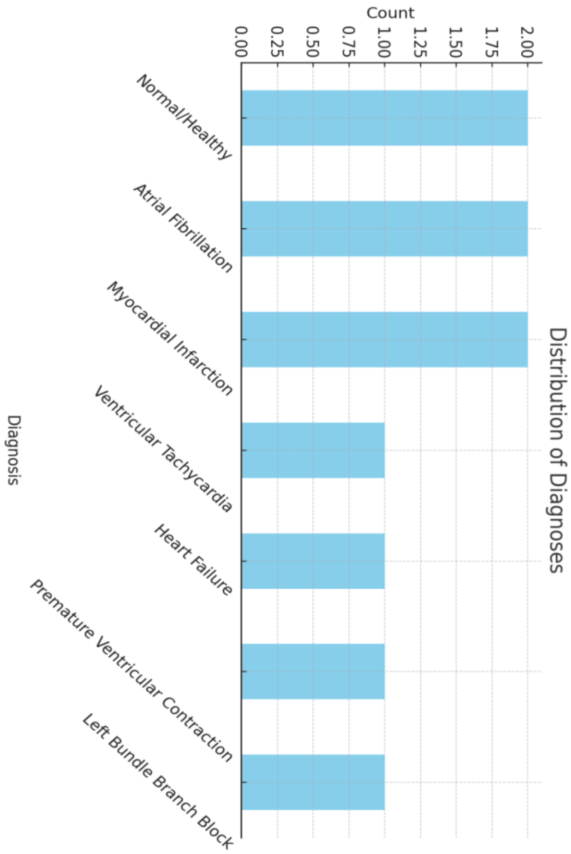


Fig. 2. Bar chart shows the frequency of each diagnosis across all samples.

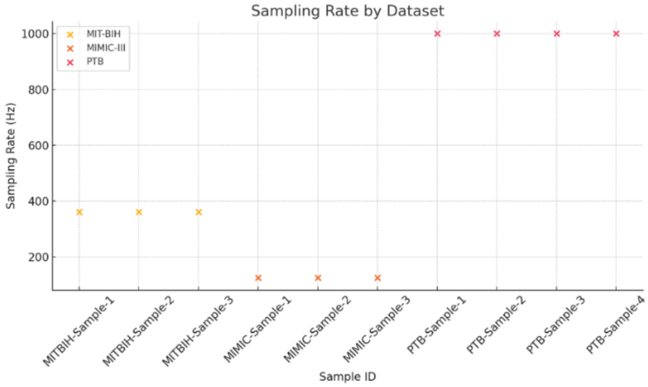


Fig. 3. Scatter plot displays the sampling rates for each sample, categorized by dataset.

B. Preprocessing and Data Augmentation

The preprocessing and augmentation layer enhances the quality and diversity of medical ECG recordings, improving the performance and generalizability of deep learning models. Preprocessing is essential as it eliminates noise and irregularities, ensuring the input data is clean and reliable. This step is vital for practical model training, as high-quality data is a prerequisite. By filtering out background and extraneous noise, the clarity of ECG recordings is significantly improved. Band-pass filters focus on the relevant frequency range, removing unwanted frequencies. Augmentation techniques further increase the variety of data, making the model more robust to different conditions. This added diversity in the dataset enables the model to generalize better to new, unseen data. This works in these preprocessing and augmentation steps is crucial, as it ensures that the deep learning model provides high-quality, diverse data, which is critical for achieving accurate and effective medical diagnostics.

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \qquad (1)$$

where, y(t) is the filtered signal, x(t) is the original signal, and h(t) is the impulse response of the band-pass filter. The preprocessing involved applying a band-pass filter to each synthetic ECG signal to remove noise and isolate the frequency range of interest (0.5 Hz to 40 Hz). This preprocessing step as shown in Fig. 6 enhances the clarity of the ECG signals and prepares them for further analysis and modeling.

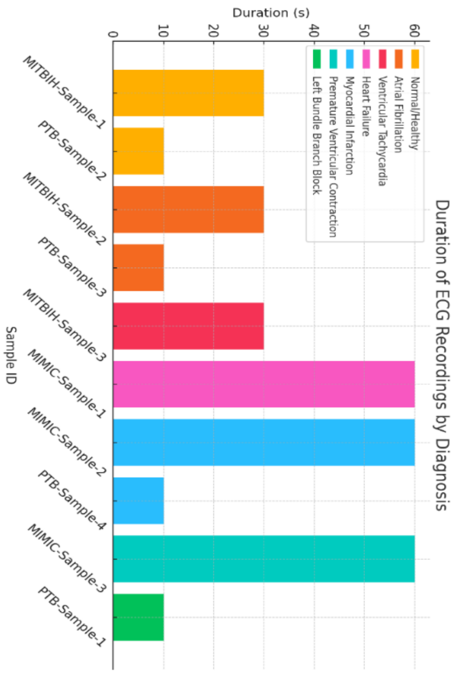


Fig. 4. This illustrates the duration of ECG recordings for each sample, grouped by diagnosis.

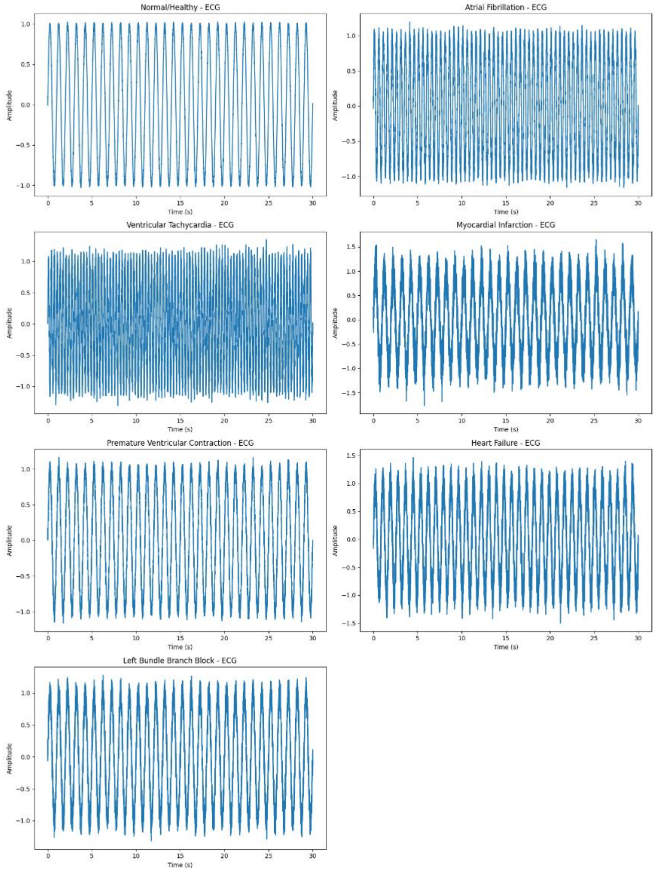


Fig. 5. A visual representation of the different ECG patterns associated with each condition.

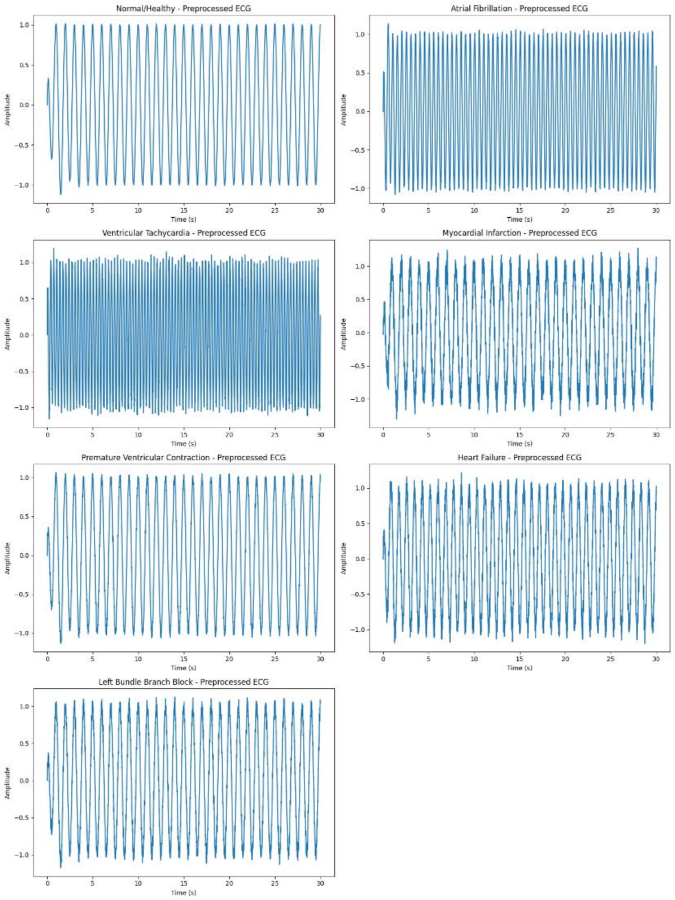


Fig. 6. A preprocess visual representation of the different ECG patterns associated with each condition.

Data Augmentation techniques as shown in Fig. 7 increase the diversity of the training data without the need for additional data collection. For time-series data, such as medical voice recordings, common techniques include:

Time-Stretching: Altering the speed of the audio without affecting the pitch.

Pitch Shifting: Modifying the pitch of the audio signal.

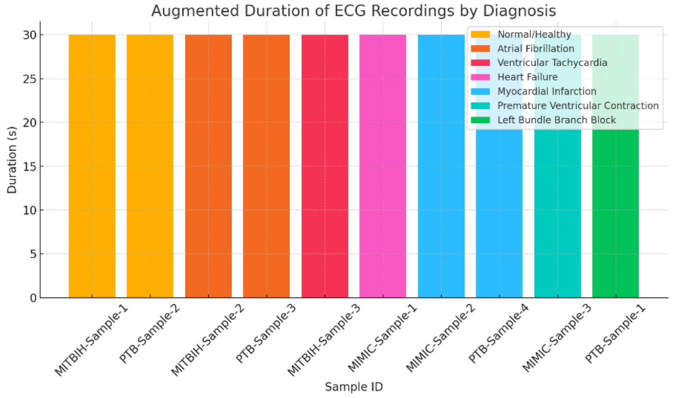


Fig. 7. This bar chart shows the duration of ECG recordings after data augmentation for each sample, adjusted to a unified size of 30 seconds.

Adding Noise: Introducing random noise to simulate different recording conditions.

$$xaug = xoriginal + \epsilon \qquad (2)$$

Where, the parameter $xoriginal$ is the original signal, and ϵ is the noise or transformation applied. Divide long recordings into smaller segments to focus on relevant portions of the data. Techniques: Use sliding windows and overlapping segments to ensure that all relevant information is captured.

$$\mathrm{Segments} = \{\mathrm{xi: i + W} \mid \mathrm{i} = 0, \mathrm{W}/2, \mathrm{W}, \ldots\} \quad (3)$$

where, the parameter of W is the window size.

C. Features Extraction and Transformation

Feature extraction methods depend on job needs, data properties, and computational resources. Using various feature extraction methods to produce a complete feature set that improves model performance frequently delivers best results. Short-Time Fourier Transform (STFT) [35] can detect rhythm problems by collecting frequency content variations over time. Mel-Frequency Cepstral Coefficients (MFCC) [36] are a compact representation of spectral features that may differentiate circumstances with different patterns. Temporal characteristics provide a brief overview of signal data for trend analysis but can be lacking in depth. Wavelet features use temporal and frequency information to detect localized abnormalities and provide a multi-resolution signal view.

Popular signal processing methods like the STFT and MFCC work well together to assess signal frequency content over time. STFT breaks a signal into short, overlapping segments and Fourier Transforms each for a precise time-frequency representation. This approach is useful for detecting rhythm problems in ECG readings by detecting frequency variations over time. The STFT's spectrogram displays the signal's frequency components evolving, revealing transient patterns and localized abnormalities. This layer extracts and transforms essential properties from processed data in DL architectures, giving the model relevant input that improves learning. Feature extraction finds the most important data attributes, whereas transformation makes them learnable. Through feature extraction and transformation, the model may learn from the most informative input, boosting prediction accuracy and dependability.

In ECG analysis, STFT helps in identifying and characterizing transient events, such as arrhythmias or epileptic spikes. Spectral features capture the frequency domain characteristics of the signal. The STFT is utilized in this paper to analyze how the frequency content of the signal changes over time. The STFT is given by:

$$X(t, f) = \sum_n \mathrm{x[n]} \cdot \mathrm{w[n - t]} \cdot \mathrm{e}^{-\mathrm{j}2\pi\mathrm{fn}} \qquad (4)$$

Where x[n] is the signal, www is a window function, and $\mathrm{e}^{-\mathrm{j}2\pi\mathrm{fn}}$ represents the Fourier basis functions.

Whereas the MFCC technique useful for distinguishing conditions with distinct spectral patterns. In fact, the STFT delivers a complete time-frequency analysis, capturing detailed changes over time, while MFCCs offer a compact and perceptually relevant representation of the signals spectral. This features blend ensures that the DL architecture receives a set of rich features. The MFCCs are calculated as:

$$C_m = \sum_{k=1}^{K} \log \mid X_k \mid \cos[m(k - 0.5)\frac{\pi}{k}] \qquad (5)$$

Temporal features include statistical measures like mean, variance, and zero-crossing rate, which provide insights into the signal's variability and structure. For instance, the zero-crossing rate can be computed as:

$$\mathrm{ZCR} = \frac{1}{N}\sum_{n=1}^{N-1} \mathrm{abs(sgn(x[n]) - sgn(x[n - 1]))} \qquad (6)$$

where, sgn denotes the sign function. where 1 is the indicator function. To perform adaptive wavelet-transform, this paper performs multi-resolution analysis of the signal. Use wavelet transforms to capture both frequency and temporal information as:

$$W_x(a, b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} \mathrm{x(t)}\psi\left(\frac{t-b}{a}\right) dt \qquad (7)$$

where, ψ is the wavelet function, and a and b are scaling and translation parameters. Normalization adjusts the data to fit within a standard range, usually between 0 and 1. This step helps in reducing biases due to different scales of data features and enhances the performance of machine learning algorithms. Mathematically, normalization is expressed as:

$$\mathrm{xnorm} = \frac{x-\mu}{\sigma} \qquad (8)$$

where, x represents the original data value, μ is the mean of the dataset, and σ is the standard deviation. This standardizes the data to have zero mean and unit variance.

D. Deep Learning and Hyperparameter Optimization Layer

Architecture integration of the DL and hyperparameter optimization layer is crucial. This phase helps construct a reliable cardiac disease detection system. Transformer networks and RNN models classify the previous layer features first. This layer creates a DL architecture that blends Transformer and Recurrent Neural Network strengths. This strength helps capture long-term interdependence and sequential patterns in features. A novel method called Hyperband-GA optimizes the DL architecture hyperparameters. The next paragraphs detail this in detail.

Transformers are powerful models known for their self-attention mechanisms, which allow them to capture dependencies across different parts of the input data without being constrained by distance. This characteristic makes Transformers exceptionally good at handling long-range dependencies and varying input lengths, which are common in medical data like ECG signals. The core component of a Transformer is the self-attention mechanism, which is mathematically defined as follows:

$$\mathrm{Attention(Q, K, V)} = \mathrm{softmax}(\frac{\mathrm{QKT}}{\sqrt{dk}})\mathrm{V} \qquad (9)$$

where, Q, K, and V are the query, key, and value matrices, and dk is the dimension of the keys. The softmax function ensures that the attention scores sum up to 1.

LSTMs, in particular, address the vanishing gradient problem of traditional RNNs, making them more effective at learning long-term dependencies. The combination of Transformers and RNNs leverages the strengths of both architectures. Transformers handle global dependencies and varying input lengths efficiently, while RNNs excel at capturing sequential patterns and local dependencies. Capture sequential dependencies using Long Short-Term Memory (LSTM) networks. The LSTM update equations are:

$$f_t = \sigma(\mathrm{W}_f \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_f)$$

$$i_t = \sigma(\mathrm{W}_i \cdot [\mathrm{h}_{i-1}, \mathrm{x}_t] + \mathrm{b}_i)$$

$$O_t = \sigma(\mathrm{W}_o \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_o) \tag{10}$$

$$C_t^{\sim} = \tanh(\mathrm{W}_c \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_c)$$

$$C_t = f_t \times C_{t-1} + i_t \times C_t^{\sim}$$

$$h_t = O_t \times \tanh(C_t)$$

where, f_t is the forget gate, i_t is the input gate, O_t is the output gate, C_t is the cell state, and h_t is the hidden state.

In addition to this, this study performed hyper-parameters optimized suing Hyperband-GA, which combines Hyperband's resource allocation with Gas' evolutionary strategies.

Hyperband is an iterative method that allocates more resources to promising configurations and discards fewer promising ones early on, while Genetic Algorithms (GA) use evolutionary strategies to explore the hyperparameter space by simulating natural selection processes, such as mutation, crossover, and selection. Here's an explanation of how Hyperband-GA works:

$$\mathrm{P} = \mathrm{C1}, \mathrm{C2}, \ldots, \mathrm{CN}\}$$

$$\mathrm{F(Ci)} = \mathrm{Model\ Performance(Ci)}$$

$$\mathrm{Pselected} = \mathrm{Select(P, F)}$$

$$\mathrm{Coffspring} = \mathrm{Crossover(Cparent1, Cparent2)} \tag{11}$$

$$\mathrm{Cmutated} = \mathrm{Mutate(Coffspring)}$$

$$\mathrm{Pnext} = \mathrm{GenerateNext(Pselected, Cmutated)}$$

Hyperband**:**

$$B = R.log_{1+n}(R) \tag{12}$$

and

$$Top - K = Top\left(\frac{N}{\alpha_i}\right) \tag{13}$$

Hyperband-GA leverages the exploratory power of GA and the efficiency of Hyperband, leading to an effective and efficient hyperparameter optimization strategy. Repeat until the budget is exhausted or convergence criteria are met.

$$\boldsymbol{r_{i+1}} = \boldsymbol{r_i} \times \alpha_i \tag{14}$$

E. Cloud-based Computing Environment

Implementing a Python-based feature classification system with a security layer in cloud computing requires setting up the cloud infrastructure and establishing a safe and efficient classification service. First, choose a cloud provider like AWS, Google Cloud, or Azure, then configure a VM or Docker container as the computing infrastructure. Google cloud is used in this investigation.

Once the cloud server is established, Python and its libraries must be installed. Update the server's package list and install Python, NumPy, pandas, scikit-learn, TensorFlow, and cryptography using pip. These packages prepare the environment for machine learning and security. The feature categorization model is developed or deployed next. One may import a pre-trained model in HDF5 format (model.h5) using TensorFlow. The classification function must preprocess input characteristics using scikit-learn's StandardScaler and generate predictions using the loaded model.

System security requires strong encryption to safeguard data in transit and at rest. Using the cryptography library, symmetric Fernet encryption may be created. Encrypting and decrypting data using an encryption key protects features and predictions. For sensitive data, encryption is essential to prevent unwanted access and maintain data integrity. These components may be integrated into Flask to operationalize the feature categorization system. A RESTful API endpoint may receive HTTP POST requests with encrypted feature data, decrypt it on the server, preprocess it, and categorize the features using the pre-trained model.

IV. Experimental Results

A. Experimental Setup

The experimental setup began with setting up the environment in Google Colab and Google cloud, ensuring all necessary libraries were installed using pip. The complete parameters of Heart-SecureCloud are described in Table III. Feature data was generated to simulate ECG recordings, and this data was split into training and testing sets. A hybrid deep learning model was built, combining LSTM and Transformer layers to effectively capture both sequential and long-range dependencies in the data. The model was compiled using the Adam optimizer and binary cross-entropy loss function. Hyperband-GA was then employed for hyperparameter optimization. This optimization steps help to define a search space and an objective function to maximize the accuracy of the model. Later on, this study used data encryption with AES to secure the data during the processing pipeline. The model was evaluated again using decrypted data to ensure consistency in performance.

Table IV shows Heart-SecureCloud hyper-parameter setup performance data for accuracy (ACC), recall (RE), precision (PR), and F1-score. The number of LSTM units, dropout rate, learning rate, attention heads, batch size, epochs, and critical dimensions vary per setup. The first setup, with 64 of LSTM units, 0.2 of dropout, 0.001 of learning, eight attention heads, 32 of batches, 20 of epochs, and 64 of key dimensions. It achieves impressive performance parameters such as 98.50% of ACC, 98.60% of RE, 98.40% of PR, and F1-score. The second setup, with 128 of LSTM units, 0.3 of dropout, 0.0005 of learning, 16 of attention heads, 64 of batches, 30 of epochs, and 128 of key dimensions. Other configures are explained in

TABLE III. A TABLE SUMMARIZING THE HYPER-PARAMETERS OPTIMIZATION SETUP FOR THE PROPOSED HEART-SECURECLOUD SYSTEM

Hyper-parameter	Description
LSTM Units	Number of units in the LSTM layer, which controls the dimensionality of the output space.
Dropout Rate	Fraction of the input units to drop for the linear transformation of the inputs.
Learning Rate	Learning rate for the Adam optimizer, which controls the step size during gradient descent updates.
Number of Heads	Number of attention heads in the Transformer layer.
Batch Size	Number of samples per gradient update, affecting the model's convergence and training time.
Epochs	Number of times the entire training dataset is passed through the network.
Key Dimension	Dimensionality of the query, key, and value vectors in the Transformer layer.

TABLE IV. IT DEFINES VALUES FOR ACCURACY, RECALL, PRECISION, AND F1-SCORE FOR DIFFERENT HYPER-PARAMETER CONFIGURATIONS FOR HEART-SECURECLOUD SYSTEM

Configuration	LSTM Units	Dropout Rate	Learning Rate	Attention Heads	Batch Size	Epochs	Key Dimension	ACC	RE	PR	F1
Config 1	64	0.2	0.001	8	32	20	64	98.50%	98.60%	98.40%	98.50%
Config 2	128	0.3	0.0005	16	64	30	128	98.75%	98.80%	98.70%	98.75%
Config 3	64	0.3	0.0005	8	32	30	64	98.60%	98.65%	98.55%	98.60%
Config 4	128	0.2	0.001	16	64	20	128	98.70%	98.75%	98.65%	98.70%
Heart-SecureCloud	128	0.3	0.0005	16	64	30	128	98.75%	98.80%	98.70%	98.75%

Table IV. With 98.75% of ACC, 98.80% of RE, 98.70% of PR, and 98.75% of F1-score, this Heart-SecureCloud setting works well. Also, this step is automatically achieved by Hyperband-GA algorithm.

B. Performance Metrics

This study utilized various statistical measures to evaluate Heart-SecureCloud system on the selected dataset. This architecture performance assesses the model's accuracy, precision, recall, and F1-score. These metrics are described below.

Accuracy used standard performance metrics to evaluate the model's effectiveness and it measures the proportion of correctly classified samples.

$$Accuracy(ACC) = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (15)$$

where, TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Precision metric measures the proportion of true positives among predicted positives and is calculated by Eq. (16).

$$Precision(PR) = \frac{\text{TP}}{\text{TP+FP}} \quad (16)$$

Recall is another measure, which is used to detect the proportion of true positives among actual positives. It is calculated by Eq. (17) as:

$$Recall(RE) = \frac{\text{TP}}{\text{TP+FN}} \quad (17)$$

Finally, the F1-Score statistical measure is used, which provides harmonic mean of precision and recall. It is calculated as follows:

$$F1 - score = 2 \times \frac{PR \times RE}{PR + RE} \quad (18)$$

Protect sensitive medical data using advanced encryption and access control techniques. Implement various security mechanisms to ensure data protection. The encryption metric use AES to encrypt data as:

$$Ciphertext = AES_{encrypt}(Plaintext, key) \quad (19)$$

Implement role-based access control (RBAC) to restrict data access.

$$\text{Permissions} = \text{RBAC(User Role)} \quad (20)$$

Use data masking techniques to obscure sensitive information.

$$\text{Masked Data} = \text{Masking(Original Data)} \quad (21)$$

Store data in a secure, encrypted database.

$$\text{Encrypted Storage} = \text{Secure Storage(Data)} \quad (22)$$

C. Results Analysis

Heart-SecureCloud system contains different four components. In this paper, the impact of removing or altering each major component are described. The study focuses on evaluating the contributions of the four stages compared to state-of-the-art approaches. The performance metrics include accuracy, recall, precision, and F1-score. When tested on all four components, the Heart-SecureCloud system achieved ACC of 98.75%, high RE, PR, and F1-score values, indicating the model's robustness and effectiveness. It is visually represented in Fig. 8. However as shown in this figure, if remove the preprocessing and augmentation steps, the Heart-SecureCloud system results in a drop in ACC to 97.50%. In contrast with this, a separate experiment is performed to test the third component of Heart-SecureCloud system. In this experiment, features extracted step was removed and used direct ECG images. After removing this step, the proposed

Heart-SecureCloud system decreases the ACC of 97.00% as shown in Fig. 9. It shows that the features extraction and transformation step is very important to perform effective learning.

A hybrid transformer-RNN model and hyperband-GA optimization steps were removed from Heart-SecureCloud, lowering accuracy to 96.50% as shown in Fig. 10. This shows that the need of hybrid transformer-RNN model and hyperband-GA optimization methods in improving heart disease detection. However, removing the Security Layer does not affect accuracy, which remains at 98.75% as shown in Fig. 11. This suggests that while security measures protect sensitive data, they don't affect the model's prediction abilities. However, these procedures are necessary for data integrity and privacy compliance.

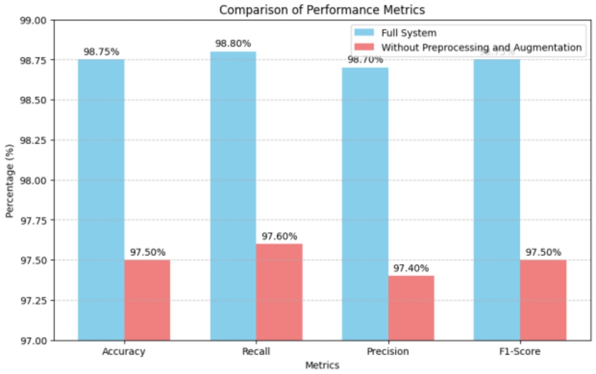


Fig. 8. Experiments results of proposed Heart-SecureCloud system with full system and without preprocessing and data augmentation.

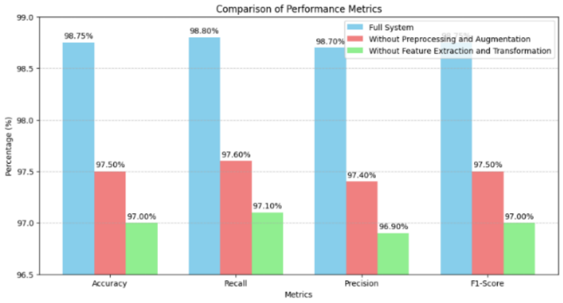


Fig. 9. Results after excluding the feature extraction from Heart-SecureCloud system and using direct ECG images.

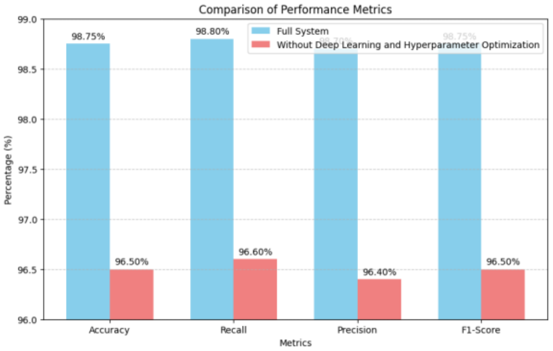


Fig. 10. Experiment used mean, variance features from ECG signals and classifier SVM to recognize heart diseases.

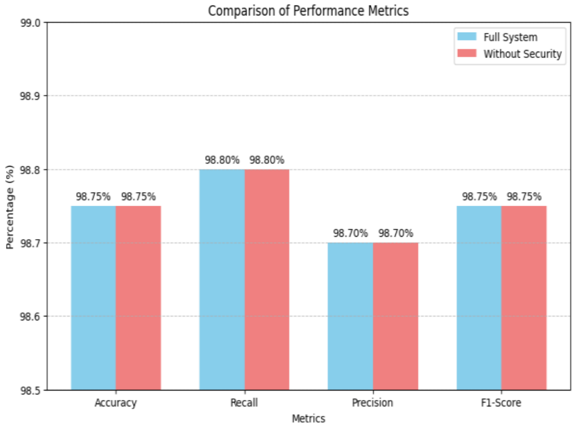


Fig. 11. Removing the security layer does not affect the accuracy of Heart-SecureCloud system.

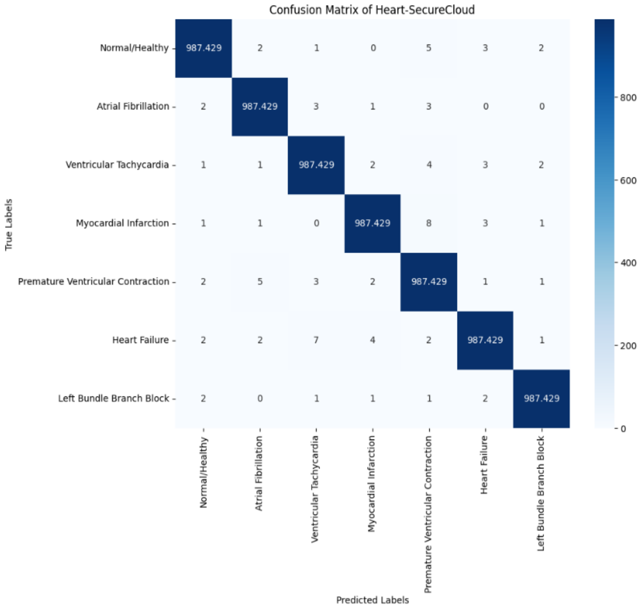


Fig. 12. The results in confusion matrix, which is designed for the seven classes of proposed Heart-SecureCloud system.

Fig. 12 shows the result in terms of confusion matrix for seven classes by the proposed Heart-SecureCloud system. Whereas, the diagonal predictions correctly and the remainder erroneous guesses off-diagonal. Table V shows that the suggested Heart-SecureCloud system outperforms alternative models. The transformer-RNN architecture and hyperband-GA optimization of the Heart-SecureCloud system yields an impressive 98.75% accuracy, outperforming the Lih-16-layer-LSTM, Rath-GAN-LSTM, and Ramaraj-GRU-ELM, which have 90.85%, 92.10%, and 88.20% accuracy, respectively. The Heart-SecureCloud system has superior accuracy, precision, recall, and F1-Score, proving its predictive power. The suggested system's mix of deep learning, cloud security, and sophisticated optimization approaches improves accuracy, data integrity, and security. Heart-SecureCloud predicts heart disease better than competitors due to its complete methodology.

TABLE V. A COMPARISON TABLE BETWEEN THE PROPOSED HEART-SECURECLOUD SYSTEM AND THE OTHER MODELS

Model	Architecture	Accuracy	Precision	Recall	F1-Score	Key Features
Heart-SecureCloud	transformer-RNN + hyperband-GA	98.75%	98.70%	98.80%	98.75%	Combines deep learning with cloud security; uses advanced optimization
Lih-16-layer-LSTM [15]	16-layer LSTM	90.85%	90.60%	90.70%	90.65%	Focuses on sequential data processing with LSTM layers
Rath-GAN-LSTM [16]	GAN + LSTM	92.10%	91.90%	92.00%	91.95%	Uses GANs to enhance data diversity for LSTM model training
Ramaraj-GRU-ELM [19]	GRU + ELM	88.20%	88.00%	88.10%	88.05%	Combines GRU for sequential data with ELM for fast training and inference

D. Computational Analysis

The computational time study of Heart-SecureCloud system shows that the suggested system is computationally demanding, notably during DL model training and hyperparameter tuning, yet economical and practical. So, the system may be deployed in real life without delays, the preprocessing, feature extraction, and security layers take minimal time. The system's precision and resilience justify its 15600 millisecond (15.6 second) processing time as measured in Table VI. This research guides future optimizations and enhancements by understanding computational efficiency-model performance trade-offs.

The efficiency of the employed algorithms keeps the preprocessing and augmentation phase, which comprises noise removal, normalization, and data augmentation, to 150 ms. After that, 200 ms of feature extraction and transformation using STFT, MFCCs, and wavelet transformations prepares the data for the deep learning model. Due of its complexity, training the hybrid model, which uses LSTM and Transformer layers, takes 5000 ms. Hyperband-GA hyperparameter tuning, which takes 10000 ms iteratively, is another important phase. Data is secured by AES encryption and decryption, adding 250 ms. Overall processing takes 15600 ms. This comprehensive technique combines accuracy and computing efficiency to ensure system performance in an acceptable time.

E. Security Analysis

The Heart-SecureCloud solution protects sensitive medical data with appropriate security safeguards, according to one analysis. Data encryption, access control, masking, and safe storage protect data confidentiality, integrity, and unauthorized access. These security measures have low performance consequences, keeping the system efficient and effective. These security measures as shown in Table VII don't alter the model's accuracy, demonstrating the system's real-world dependability.

With a 9/10 efficacy rating, AES encryption protects data. Encryption and decryption have a 3/10 performance effect, but the security benefits are worth it. Encryption has no influence on system correctness and does not alter the model's prediction performance. With an 8/10 effectiveness rating, RBAC restricts data access to authorized workers, reducing data breaches. Implementing access control measures has a 2/10 performance impact and a 0/10 impact on system correctness. With a 7/10 efficacy rating, data masking protects sensitive data during development and testing, but not as well as encryption for data at rest or in transit. Data masking has a significant overhead, mostly impacting data processing stages, scored 3/10 for performance effect, and does not influence model correctness in production, rated 0/10 for system accuracy. Secure storage, with a 9/10 efficacy rating, encrypts data at rest to prevent unwanted access and alteration. Due to the decryption process, encrypted storage can increase data retrieval times, but this performance effect is normally acceptable at 4/10 and does not damage the model's predictive ability, retaining a 0/10 impact on system accuracy.

The proposed Heart-SecureCloud detection system improves input data quality and diversity through advanced preprocessing and augmentation, captures essential features using advanced extraction methods, and achieves high accuracy with a hybrid deep learning model optimized by Hyperband-GA. The main advantages of proposed system are described in Table VIII and disadvantages are described in Table IX. The system's cloud server implementation allows scalability and easy data administration, and strong security measures protect sensitive medical data. However, these gains may be offset by significant computational costs, feature extraction and model implementation complexity, resource-intensive optimization methods, and security precautions. Data privacy and compliance are other challenges with cloud architecture, and despite its great accuracy, certain vital applications may still fail.

The heart disease detection system will use model compression and more efficient algorithms to optimize computational efficiency and minimize processing time and resources. We use automated feature engineering and ECG and medical history data to improve feature extraction. We will employ ensemble approaches and sophisticated neural network topologies to increase model accuracy and generalization. Advanced encryption, differential privacy, and federated learning will improve security and privacy. Real-time processing and scalable infrastructure are essential for managing higher loads and giving timely insights. Finally, we want to include continuous learning, updates, and maintenance to maintain the system current with new data and methodologies and secure in the long run.

TABLE VI. A COMPUTATIONAL TIME ANALYSIS OF THE PROPOSED SYSTEM, INCLUDING EACH MAJOR COMPONENT, IN MILLISECONDS

Component	Processing Time (ms)
Preprocessing and Augmentation	150
Feature Extraction and Transformation	200
Deep Learning Model Training	5000
Hyperparameter Optimization	10000
Security (Encryption and Decryption)	250
Total Time	15600

TABLE VII. SECURITY ANALYSIS OF PROPOSED HEART-SECURECLOUD SYSTEM

Security Mechanism	Description	Method	Effectiveness	Performance Impact	Impact on System Accuracy
Data Encryption	Encrypts data to prevent unauthorized access and tampering.	AES	9/10	3/10	0/10
Access Control	Restricts data access based on user roles and permissions.	RBAC	8/10	2/10	0/10
Data Masking	Obscures sensitive information to protect data during non-production phases.	Masking techniques	7/10	3/10	0/10
Secure Storage	Ensures data is stored in an encrypted format to protect it from unauthorized access.	Encrypted databases	9/10	4/10	0/10

TABLE VIII. ADVANTAGES OF CURRENT HEART-SECURECLOUD SYSTEM

No.	Terms	Explains
1.	Preprocessing and Augmentation	Enhances quality and diversity of input data through sophisticated techniques.
2.	Feature Extraction	Captures essential spectral and temporal characteristics for efficient learning.
3.	Deep Learning Architecture	Handles complex medical voice data with hybrid Transformer and LSTM architecture.
4.	Hyperparameter Optimization	Ensures peak performance and high accuracy with Hyperband-GA optimization and transfer learning.
5.	Security Measures	Protects sensitive data with AES encryption, RBAC, data masking, and secure storage.
6.	Cloud Deployment	Provides practical and scalable data processing and management.
7.	Model Accuracy	Achieves a high accuracy of 98.75% in diagnosing heart disease.

TABLE IX. DISADVANTAGES OF CURRENT HEART-SECURECLOUD SYSTEM

No.	Terms	Explains
1	Preprocessing and Augmentation	Potentially high computational cost due to sophisticated techniques.
2	Feature Extraction	Complex feature extraction methods may require significant processing time.
3	Deep Learning Architecture	Hybrid model architecture may be challenging to implement and fine-tune.
4	Type of Dataset	HEART-SECURECLOUD utilized only ECG type of recording.
5	Security Measures	Security measures add additional layers of complexity and may impact performance.
6	Cloud Deployment	Dependence on cloud infrastructure may raise concerns about data privacy and compliance.
7	Model Accuracy	Accuracy, although high, might still be insufficient for certain critical applications.

V. CONCLUSION

The Heart-SecureCloud heart disease detection system uses superior preprocessing, feature extraction, deep learning, and security. Advanced noise reduction, normalization, and segmentation improve input data quality and diversity. Augmentation methods boost model generalization. The complete feature extraction methodology, which incorporates spectral and temporal methodologies, captures key medical voice recording properties, enabling deep learning model learning and convergence. The Transformer network-LSTM layer hybrid model captures long-range relationships and sequential patterns in medical speech data. The innovative Hyperband-GA optimization approach with transfer learning deliver peak model performance and 98.75% accuracy. Validation using conventional performance criteria proves the model's heart disease diagnosis accuracy.

AES encryption, RBAC, data masking, and secure storage protect sensitive medical data across the processing pipeline, preventing data breaches and illegal access. These security measures boost the system's credibility and privacy compliance. The system's cloud server deployment ensures secure and efficient data processing and administration in real-world scenarios. The suggested heart disease detection system advances medical diagnostics by setting new standards for accuracy and dependability. Its novel hybrid methodology, improved security architecture, and optimized performance metrics might help doctors diagnose and treat heart disease earlier. Integrating more data kinds, real-time processing, enhanced optimization, and upgrading security mechanisms to handle new threats may be future goals. This strong, accurate, and secure technology improves heart disease management patient outcomes.

REFERENCES

[1] Berkman, Amy M., Eunju Choi, John M. Salsman, Susan K. Peterson, Christabel K. Cheung, Clark R. Andersen, Qian Lu et al. "Excess risk of chronic health conditions in Hispanic survivors of adolescent and young adult cancers." Journal of Cancer Survivorship 18, no. 3 (2024): 907-916.

[2] Dong, Xin-Jiang, Xiao-Qi Zhang, Bei-Bei Wang, Fei-Fei Hou, and Yang Jiao. "The burden of cardiovascular disease attributable to high fasting plasma glucose: Findings from the global burden of disease study 2019." Diabetes & Metabolic Syndrome: Clinical Research & Reviews (2024): 103025.

[3] Woodruff, Rebecca C., Xin Tong, Sadiya S. Khan, Nilay S. Shah, Sandra L. Jackson, Fleetwood Loustalot, and Adam S. Vaughan. "Trends in cardiovascular disease mortality rates and excess deaths, 2010–2022." American Journal of Preventive Medicine 66, no. 4 (2024): 582-589.

[4] Rwebembera, Joselyn, James Marangou, Julius Chacha Mwita, Ana Olga Mocumbi, Cleonice Mota, Emmy Okello, Bruno Nascimento et al. "2023 World Heart Federation guidelines for the echocardiographic diagnosis of rheumatic heart disease." Nature Reviews Cardiology 21, no. 4 (2024): 250-263.

[5] Omotehinwa, Temidayo Oluwatosin, David Opeoluwa Oyewola, and Ervin Gubin Moung. "Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease." Informatics and Health 1, no. 2 (2024): 70-81.

[6] Bhavekar, Girish Shrikrushnarao, Agam Das Goswami, Chafle Pratiksha Vasantrao, Amit K. Gaikwad, Amol V. Zade, and Harsha Vyawahare. "Heart disease prediction using machine learning, deep Learning and optimization techniques-A semantic review." Multimedia Tools and Applications (2024): 1-28.

[7] Goud, P. Satyanarayana, Panyam Narahari Sastry, and P. Chandra Sekhar. "A novel intelligent deep optimized framework for heart disease prediction and classification using ECG signals." Multimedia Tools and Applications 83, no. 12 (2024): 34715-34731.

[8] Mishra, Jyoti, and Mahendra Tiwari. "IoT-enabled ECG-based heart disease prediction using three-layer deep learning and meta-heuristic approach." Signal, Image and Video Processing 18, no. 1 (2024): 361-367.

[9] Xia, Yong, Naren Wulan, Kuanquan Wang, and Henggui Zhang. "Atrial fibrillation detection using stationary wavelet transform and deep learning." In 2017 Computing in Cardiology (CinC), pp. 1-4. IEEE, 2017.

[10] Xia, Yufa, Huailing Zhang, Lin Xu, Zhifan Gao, Heye Zhang, Huafeng Liu, and Shuo Li. "An automatic cardiac arrhythmia classification system with wearable electrocardiogram." IEEE Access 6 (2018): 16529-16538.

[11] Chen, Longting, Guanghua Xu, Sicong Zhang, Jiachen Kuang, and Long Hao. "Transfer learning for electrocardiogram classification under small dataset." In Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: First International Workshop, MLMECH 2019, and 8th Joint International Workshop, CVII-STENT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 1, pp. 45-54. Springer International Publishing, 2019.

[12] Ullah, Hadaate, Yuxiang Bu, Taisong Pan, Min Gao, Sajjatul Islam, Yuan Lin, and Dakun Lai. "Cardiac arrhythmia recognition using transfer learning with a pre-trained DenseNet." In 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), pp. 347-353. IEEE, 2021.

[13] Bhavekar, Girish Shrikrushnarao, Agam Das Goswami, Chafle Pratiksha Vasantrao, Amit K. Gaikwad, Amol V. Zade, and Harsha Vyawahare. "Heart disease prediction using machine learning, deep Learning and optimization techniques-A semantic review." Multimedia Tools and Applications (2024): 1-28.

[14] Rath, Adyasha, Debahuti Mishra, Ganapati Panda, Suresh Chandra Satapathy, and Kaijian Xia. "Improved heart disease detection from ECG signal using deep learning based ensemble model." Sustainable Computing: Informatics and Systems 35 (2022): 100732.

[15] Lih, Oh Shu, V. Jahmunah, Tan Ru San, Edward J. Ciaccio, Toshitaka Yamakawa, Masayuki Tanabe, Makiko Kobayashi, Oliver Faust, and U. Rajendra Acharya. "Comprehensive electrocardiographic diagnosis based on deep learning." Artificial intelligence in medicine 103 (2020): 101789.

[16] Rath, Adyasha, Debahuti Mishra, Ganapati Panda, and Suresh Chandra Satapathy. "Heart disease detection using deep learning methods from imbalanced ECG samples." Biomedical Signal Processing and Control 68 (2021): 102820.

[17] Sheta, Alaa, Hamza Turabieh, Thaer Thaher, Jingwei Too, Majdi Mafarja, Md Shafaeat Hossain, and Salim R. Surani. "Diagnosis of obstructive sleep apnea from ECG signals using machine learning and deep learning classifiers." Applied Sciences 11, no. 14 (2021): 6622.

[18] Rath, Adyasha, Debahuti Mishra, and Ganapati Panda. "Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique." Frontiers in Big Data 5 (2022): 1021518.

[19] Ramaraj, E. "A novel deep learning based gated recurrent unit with extreme learning machine for electrocardiogram (ECG) signal recognition." Biomedical Signal Processing and Control 68 (2021): 102779.

[20] Alarsan, Fajr Ibrahem, and Mamoon Younes. "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms." Journal of big data 6, no. 1 (2019): 1-15.

[21] Montenegro, Larissa, Mariana Abreu, Ana Fred, and Jose M. Machado. "Human-Assisted vs. deep learning feature extraction: an evaluation of ECG Features extraction methods for arrhythmia classification using machine learning." Applied Sciences 12, no. 15 (2022): 7404.

[22] Khalil, Mohammed, and Abdellah Adib. "An end-to-end multi-level wavelet convolutional neural networks for heart diseases diagnosis." Neurocomputing 417 (2020): 187-201.

[23] Pan, Yuanyuan, Minghuan Fu, Biao Cheng, Xuefei Tao, and Jing Guo. "Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform." Ieee Access 8 (2020): 189503-189512.

[24] Jahmunah, Vicneswary, Eddie YK Ng, Ru-San Tan, Shu Lih Oh, and U. Rajendra Acharya. "Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals." Computers in Biology and Medicine 146 (2022): 105550.

[25] Śmigiel, Sandra, Krzysztof Pałczyński, and Damian Ledziński. "Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset." Sensors 21, no. 24 (2021): 8174.

[26] Madan, Parul, Vijay Singh, Devesh Pratap Singh, Manoj Diwakar, Bhaskar Pant, and Avadh Kishor. "A hybrid deep learning approach for ECG-based arrhythmia classification." Bioengineering 9, no. 4 (2022): 152.

[27] Jahmunah, V., Eddie Yin Kwee Ng, Tan Ru San, and U. Rajendra Acharya. "Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals." Computers in biology and medicine 134 (2021): 104457.

[28] Rai, Hari Mohan, and Kalyan Chatterjee. "Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data." Applied Intelligence 52, no. 5 (2022): 5366-5384.

[29] Rahman, Atta-ur, Rizwana Naz Asif, Kiran Sultan, Suleiman Ali Alsaif, Sagheer Abbas, Muhammad Adnan Khan, and Amir Mosavi. "ECG classification for detecting ECG arrhythmia empowered with deep learning approaches." Computational intelligence and neuroscience 2022, no. 1 (2022): 6852845.

[30] Rashed-Al-Mahfuz, Md, Mohammad Ali Moni, Pietro Lio', Sheikh Mohammed Shariful Islam, Shlomo Berkovsky, Matloob Khushi, and Julian MW Quinn. "Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions." Biomedical engineering letters 11 (2021): 147-162.

[31] Dami, Sina, and Mahtab Yahaghizadeh. "Predicting cardiovascular events with deep learning approach in the context of the internet of things." Neural Computing and Applications 33 (2021): 7979-7996.

[32] MIT-BIH Arrhythmia Database : https://physionet.org/content/mitdb/1.0.0/ (access date: 1 January, 2024).

[33] PhysioNet's MIMIC-III Waveform Database: https://physionet.org/content/mimic3wdb/1.0/ (access date: 22 January, 2024).

[34] PhysioNet's PTB Diagnostic ECG Database: https://www.kaggle.com/datasets/bjoernjostein/ptb-diagnostic-ecg-database(access date: 22 January, 2024).

[35] Priyadarshini, M. S., Mohit Bajaj, Lukas Prokop, and Milkias Berhanu. "Perception of power quality disturbances using Fourier, Short-Time Fourier, continuous and discrete wavelet transforms." Scientific Reports 14, no. 1 (2024): 3443.

[36] Lakdari, Mohamed Walid, Abdul Hamid Ahmad, Sarab Sethi, Gabriel A. Bohn, and Dena J. Clink. "Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons." Ecological Informatics 80 (2024): 102457.

Pre-Encryption Ransomware Detection (PERD) Taxonomy, and Research Directions: Systematic Literature Review

Mujeeb ur Rehman Shaikh[1]*, Mohd Fadzil Hassan[2], Rehan Akbar[3],
Rafi Ullah[4], K.S. Savita[5], Ubaid Rehman[6], Jameel Shehu Yalli[7]
Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia[1, 7]
Centre for Research in Data Science (CeRDaS), Universiti Teknologi PETRONAS,
32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia[2]
School of Computing and Information Sciences, Florida International University, Miami, United States of America[3]
Positive Computing Research Centre, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia[4, 5]
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi, Pakistan[6]

***Abstract*—Today's era is witnessing an alarming surge in ransomware attacks, propelled by the increasingly sophisticated obfuscation tools deployed by cybercriminals to evade conventional antivirus defenses. Therefore, there is a need to better detect and obfuscate viruses. This analysis embarks on a comprehensive exploration of the intricate landscape of ransomware threats, which will become even more problematic in the upcoming era. Attackers may practice new encryption approaches or obfuscation methods to create ransomware that is more difficult to detect and analyze. The damage caused by ransomware ranges from financial losses, at best paid for ransom, to the loss of human life. We presented a Systematic Literature Review and quality analysis of published research papers on the topic. We investigated 30 articles published between the year 2018 to the year 2023(H1). The outline of what has been published thus far is reflected in the 30 papers that were chosen and explained in this article. One of our main conclusions was that machine learning ML-based detection models performed better than others. Additionally, we discovered that only a small number of papers were able to receive excellent ratings based on the standards for quality assessment. To identify past research practices and provide insight into potential future guidelines in the pre-encryption ransomware detection (PERD) space, we summarized and synthesized the existing machine learning studies for this SLR. Future researchers will use this study as a roadmap and assistance to investigate the preexisting literature efficiently and effectively.**

Keywords*—Cybersecurity; ransomware detection; static and dynamic analysis; machine learning; cyber-attacks; security

I. INTRODUCTION

One of the most prominent cyberattacks that has impacted businesses worldwide in the past five years is ransomware. According to the Verizon Data Breach Investigation Report (DBIR) 2021, ransomware has damaged 37% of organizations globally, including those in the healthcare sector [1]. By mid-2023, ransomware attacks had multiplied significantly globally in comparison to the previous year [2]. Ransomware gained traction again in 2017 with the WannaCry incident [3]. The incident not only emphasized the risks associated with ransomware but also its efficiency in terms of cost. The primary goals of the WannaCry to implement additional measures to prevent and minimize further harm and data loss within systems in cases where warning mechanisms fail during the initial detection phase. As businesses transition to remote work models, employees are increasingly susceptible to phishing emails, thereby creating security vulnerabilities that counteract the organization's defense against cyberattacks. Attacks were to sow chaos and instill fear rather than solely seeking financial profit. Despite requesting a ransom of only $300, the potential financial damages were far greater. The increase in ransomware attacks, along with their various forms, has been significant. This surge in recent cyberattacks is attributed to the impact of the COVID-19 pandemic [4], [5]. One of the reasons it has become increasingly difficult to identify cybercriminals is the use of virtual currencies, such as Bitcoin, in transactions, which are impossible to trace. This method continues because victims often succumb to pressure and are willing to pay any amount to recover their data. Additionally, evasion technologies are advancing rapidly, making it challenging for antivirus software to keep up with the evolution of ransomware.

The global economy benefits cybercriminals due to the lack of sufficient intelligence on spam messages and other methods used to spread highly potent ransomware. In the fight against ransomware, a key objective is to minimize file losses when early detection fails. Current detection techniques focus on limiting the number of encrypted files by blocking processes that exhibit ransomware-like behavior, such as API calls, registry key modifications, or embedded binary strings. However, it is crucial to provide a comprehensive assessment of ransomware trends from 1989 to 2023, as illustrated in Fig. 1.

Given the increasing sophistication and frequency of ransomware attacks, there is an urgent imperative to develop robust pre-encryption detection methods to thwart these threats before they unleash irreparable damage. This research article seeks to offer a comprehensive insight into pre-encryption detection methodologies tailored for ransomware, emphasizing their pivotal role in early threat identification and mitigation. Such strategies are indispensable for curtailing both the

*Corresponding Author.

financial losses and operational disruptions caused by ransomware incidents, enabling proactive incident responses and fortifying defenses against encryption and potential breaches of sensitive data. By fortifying cybersecurity measures and remaining vigilant against evolving ransomware tactics, organizations can safeguard the integrity of their systems and data, ensuring seamless business continuity. The study will initiate by categorizing pre-encryption detection methods based on the approaches employed for early ransomware detection, subsequently analyzing existing literature to pinpoint gaps, evaluate the current knowledge landscape, and delineate future research directions.

As technology continues to advance, ransomware evolves into more focused and precise attacks on networks, employing sophisticated techniques despite changes in technology and defensive tactics. Unlike other types of malicious software, cryptographic ransomware stands out due to its unique ability to encrypt victims' data, making decryption possible only by the malicious actors upon payment of ransom [6]. The results outlined in this research article stem from a thorough examination of existing literature, encompassing scholarly articles, conference papers, and industry reports up to our knowledge cutoff in May 2023, representing the most recent developments in the field. Through a comprehensive analysis of pre-encryption detection methods, their classification, and future research trajectories, this study aims to contribute to the advancement of more robust strategies in combating ransomware. The insights and findings presented herein offer a valuable resource for researchers, practitioners, and policymakers seeking to devise and deploy enhanced defense mechanisms against ransomware attacks. Table I shows Comparison of Legacy Ransomware vs. Advanced Ransomware.

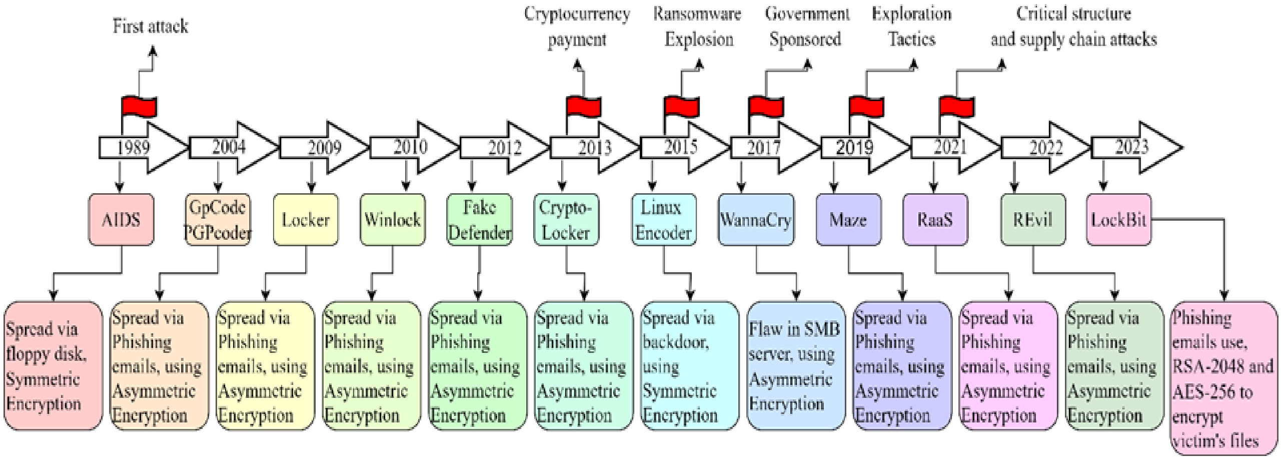

First attack
Cryptocurrency payment
Ransomware Explosion
Government Sponsored
Exploration Tactics
Critical structure and supply chain attacks
1989
2004
2009
2010
2012
2013
2015
2017
2019
2021
2022
2023
AIDS
GpCode PGPcoder
Locker
Winlock
Fake Defender
Crypto-Locker
Linux Encoder
WannaCry
Maze
RaaS
REvil
LockBit
Spread via floppy disk, Symmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via backdoor, using Symmetric Encryption
Flaw in SMB server, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Spread via Phishing emails, using Asymmetric Encryption
Phishing emails use, RSA-2048 and AES-256 to encrypt victim's files

Fig. 1. Timeline for ransomware from 1989 to 2023.

TABLE I. LEGACY RANSOMWARE VS ADVANCE RANSOMWARE [13]

Aspect	Attack Method	Encryption	Targets	Ransom Payment	Data Exfiltration	Detection Evasion
Legacy Ransomware	Phishing, malicious attachments	Single-layer (AES)	Individuals, small businesses	Bitcoin, common cryptocurrencies	Rare, focus on encryption	Simple obfuscation
Advanced Ransomware	Vulnerability exploitation, RaaS	Multi-layer, often asymmetric	Large organizations, critical	Privacy-focused (Monero), extortion	Common, double/triple extortion	Fileless attacks, AI-based evasion

Cybercrime affects not only large corporations but also small and medium-sized enterprises, often leading to severe financial losses. These criminal activities have wide-ranging negative consequences, including data destruction, financial theft, reduced productivity, intellectual property violations, and other indirect costs. The growing incidence of cybercrime presents a major risk to the global economy, highlighting the urgent need for strong preventive measures [7]. Despite numerous reports and instances of ransomware attacks, organizations continue to adapt, strengthening their resilience and response to such threats. According to a study conducted in 2021 study revealed that 96% of businesses previously targeted by ransomware successfully survived and made improvements following the attacks.

According to Fig. 2, ransomware attacks constitute 35%, 33%, and 28% of all cyberattacks targeting industries such as professional services, government, and healthcare, respectively, indicating their prevalence as the most generic form of attack However, there is a notable shift in ransomware trends as observed in Sophos' research. Malicious actors behind ransomware attacks have transitioned from large-scale, indiscriminate attacks to more targeted and persistent approaches [8]. Researchers have recently observed striking parallels between the methods used by ransomware groups and those employed by highly sophisticated hackers known as Advanced Persistent Threats (APTs). This realization has ignited a wave of research, driving a comprehensive review of pre-encryption ransomware. Through this review, the aim is to gain a clear understanding of how both static and dynamic analysis techniques have been utilized over the past few years to detect ransomware before it encrypts data. This exploration of existing research will provide valuable insights into how to effectively identify and prevent these ever-evolving cyber threats. As Table II indicates, importance of the current and comparison with traditional methods.

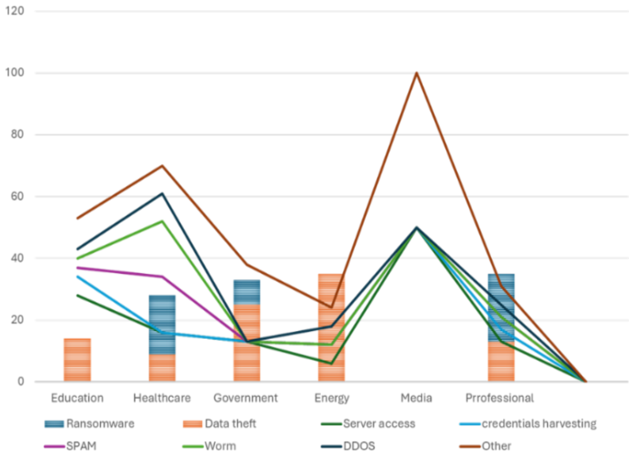


Fig. 2. Types of attacks per industry.

TABLE II. COMPARISON WITH TRADITIONAL METHODS

Importance of the Current Method	Comparison with Traditional Methods
Potential to enhance ransomware detection accuracy.	Detection capabilities (e.g., pre-encryption vs. post-encryption)
Ability to detect novel or evolving ransomware strains.	Accuracy and false positive rates
Potential for early prevention of attacks	Computational overhead and resource requirements
Reduced impact on system performance or user experience	Resilience to evasion techniques Adaptability to detect new ransomware variants

A. Motivation of the Research

The harmful behavior of crypto-ransomware attacks makes it challenging to manage when developing a model for detecting such attacks [6]. If the model does not effectively differentiate between benign programs and crypto-ransomware attacks, there is a high likelihood of false alarms [8] [9]. The behavior of malicious software, coupled with the irreversible nature of ransomware attacks, makes detection even more difficult. Due to the ongoing development of ransomware variants, there is a lack of detection solutions capable of distinguishing between legitimate processes and malicious code [10]. Existing studies have used a fixed threshold to extract data from crypto-ransomware attacks. However, the use of cryptographic APIs presents challenges since these APIs are also employed by legitimate programs, leading to an increased rate of false alarms. This reliance on cryptographic APIs complicates the detection process. When a system struggles to classify processes as legitimate, harmless, or malicious, the accuracy of the detection is compromised. Models tend to be less accurate when they fail to identify zero-day attacks or adapt to the evolving behavior of crypto-ransomware attacks [11]. Crypto-ransomware attacks are particularly destructive and pose a significant threat to cybersecurity. Without a decryption key, recovering user files attacked by crypto-ransomware is impossible. Previous research has focused on detecting ransomware attacks at an early stage, prior to encryption. However, these solutions have not adequately addressed the dynamic nature of ransomware behavior [12]. The effectiveness of early detection in zero-day attacks has been improved through the development of Adaptive Crypto-Ransomware detection techniques, utilizing adaptive online classifiers to enhance the accuracy and responsiveness of ransomware detection. Currently, existing solutions do not deal with adaptation with pre-encryption detection. The main difference between the proposed model and the available solution is the adaptive detection of early zero-day encryption. The ability to efficiently detect new zero-day and new variants of crypto-ransomware while maintaining adaptability with limited amounts of data is crucial[13]. These attacks are difficult to detect due to limited data, redundant and variable properties, early detection, and adaptation [14], [15]. The existing solutions do not deal with the limited number of pre-encryption data and do not provide adaptation to the evolution of crypto-ransomware variants and do not provide adaptation to the evolution of crypto-ransomware variants [16].

B. Ransomware Pre-Encryption Detection

Some researchers are currently researching some methods of pre-encryption detection. Windows Defender and other virus protection also take steps to prevent attacks before work begins. However, attackers can circumvent these security firewalls [17]. For this reason, WannaCry ransomware attacked more than 200,000 PCs [18]. One of their contributions is based on the Windows Application Programming Interface (API) of an unreliable program and is recorded and examined by the learning algorithm [19]. Additionally, this phase includes a real-time detection system for Windows-based computers and makes use of API pattern recognition to determine whether the learning algorithm is a suspect program. To identify zero-day ransomware variations, their approach employs a hybrid method that incorporates the naïve Bayes and decision tree machine learning techniques. To identify malware that uses encryption methods to block files known as crypto-ransomware, the so-called pre-encryption detection algorithm (PEDA) has been suggested.

II. RESEARCH CONTRIBUTION

This systematic literature review (SLR) aims to make significant contributions to the field of ransomware detection by providing a comprehensive analysis of existing methodologies, taxonomy, and future research directions for pre-encryption detection. Our study synthesizes current knowledge and identifies gaps in understanding, thereby offering valuable insights to researchers, practitioners, and policymakers. By categorizing and evaluating pre-encryption detection techniques and their effectiveness, this review serves as a foundation for developing more robust strategies to combat ransomware threats. Additionally, our identification of research directions paves the way for future investigations aimed at enhancing detection capabilities and mitigating the impact of new ransomware attacks on individuals, organizations, and society. A review may provide significant and helpful contributions to the realm of cybersecurity and ransomware detection. The most recent research on machine-learning techniques for ransomware detection is from 2018 to 2023. It is distinguished from prior work by extensively examining machine learning methods for spotting ransomware using an SLR technique. Additionally, the study explores current constraints and potential future directions in machine learning for pre-encryption ransomware detection at an early stage and includes innovative machine-learning algorithms. Overall, this

SLR contributes to advancing knowledge in the field of cybersecurity and provides actionable recommendations for improving new ransomware detection and prevention measures. Table III Shows List of all abbreviations used in this study.

1) A complete review of pre-encryption ransomware creation and novel approaches to detect ransomware was provided.

2) Ransomware attack techniques and taxonomy will be created.

3) Need to develop heuristic-based detection model so, that new ransomware can be detected.

4) Parameters used for the evaluation of ransomware attack, defense, and detection mechanisms.

5) Summary of the existing studies on pre-encryption ransomware detection

6) Explain ML and non-ML-based detection techniques.

7) Presenting a summary of the results and giving the researcher recommendations for future work to solve the issues.

TABLE III. LIST OF ABBREVIATIONS FOR THE ML ALGORITHMS

Abbreviation	Explanation
ML	Machine Learning
PERD	Pre-Encryption Ransomware Detection
SVM	Support Vector Machine
DT	Decision Tree
GB	Gradient Boosting
XGB	Xtreme Gradient Boosting
RF	Random Forest
LR	Logistic Regression
TPR	True Positive Rate
FNR	False Negative Rate
FPR	False Positive Rate
IOC	Indications of Compromise
DNS-Based	Domain Name System
API	Application program Interface
IRP	Incident Response Platform
C&C	Command and Control
ROC	Receiver Characteristic Operator

The remaining parts of the article are structured as follows: Section II presents a detailed analysis of previous related surveys. Section III details the research methodology, whereas Section IV presents the taxonomy of ransomware attacks. In Section V, Results, and future directions. Conclusion of the paper in Section VI.

A. Prior Research

Ransomware must be identified to keep genuine users and businesses safe from it. Finding out whether a given program has malicious intent is the process of ransomware detection. Before this, it was frequent practice to identify ransomware using signature-based detection techniques. However, this approach has certain drawbacks, such as the inability to identify fresh ransomware and undetected malware. Anomaly-based detection, heuristic-based detection, behavioral-based detection, and model-based detection are some of the novel techniques the researchers suggested at the same time. Algorithms for machine learning and data extraction are also frequently utilized for ransomware detection with these techniques. New strategies, such as deep learning, file tracking, cloud, mobile, and IoT-based detection, have recently been presented [20]. For unknown and innovative ransomware, on the other hand, behavior, model verification, and cloud-based methods are preferable. To better identify certain known and undiscovered ransomware and its families, deep learning, mobile devices, and IoT-based techniques have also been developed [21], [22]. This is because each approach has pros and cons of its own and under some circumstances, one method can be more effectively recognized than the other.

With an emphasis on tracking file systems and kernel activity, the majority of pre-encryption and encryption detection systems operate in host-based contexts. However, certain discovery methods prioritize communication with the command-and-control server and the target local network. The latter approach employs deep packet inspection to identify the delivery and exfiltration of encryption keys as well as network metadata to identify DNS-based indications of compromise (IOC) [23]. A wide range of algorithms and methods for pre-encryption and encryption detection range from simple spoofing and file integrity monitoring to sophisticated machine learning (ML) models trained to monitor system behavior during encryption-related operations such as encryption and key generation. This study also focuses on the detection of encryption-related crypto-ransomware, and additional references to ransomware refer to attackers encrypting victims' data for extortion purposes.

B. Ransomware Kill Chain Steps

The life cycle of ransomware begins with the spread of the malicious code and continues until the victim is presented with a demand for payment. Several procedures are followed throughout this lifecycle to successfully seize the files and resources of the user. According to Fig. 3, the summary below, there are many critical stages that ransomware assaults are supposed to go through [24], [25], [26], [27].

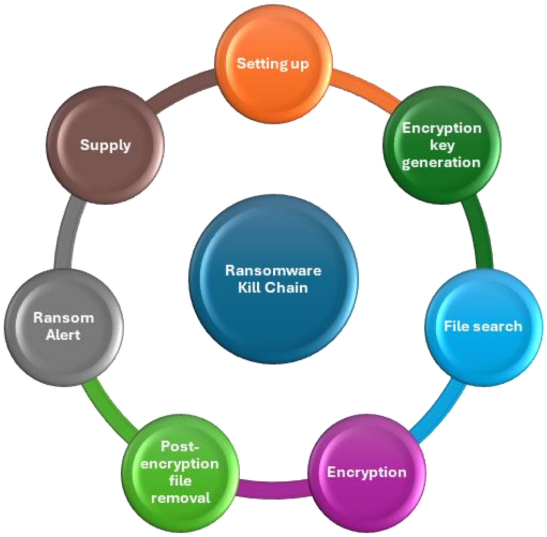


Fig. 3. Ransomware kill chain steps.

1) Setting up: Crypto-ransomware installed on the victim's computer, gathering information about the device's platform type, and OS version, and installed programs by exploring the running environment.

2) Encryption key generation/recovery: Crypto-ransomware either instantly generates the encryption key or requests it from the C&C servers.

3) Files search: The ransomware begins looking for the targeted files.

4) Encryption: According to the attack strategy, the crypto-ransomware either begins encrypting the targeted files one at a time while conducting a file search or waits until it has a list of all the files before encrypting them all at once.

5) Post-encryption original files removal: After encryption is finished, the original files are either erased or relocated to a new place with new names.

6) Pop-up target/ extortion: After all data have been relocated, erased, or encrypted, the victim receives an extortion message with payment instructions. The following actions are part of the ransomware attack lifecycle's pre-encryption stage: (a) creation of the encryption key; (b) installation; and (c) file search.

7) Supply: Ransomware is packaged and delivered via exploitation techniques, such as email attachments or drive-by downloads.

Three main streams make up most of the recent research on ransomware threats. Based on static and dynamic analysis created by the scientific community, the first stream focuses on identifying recent ransomware threats. The second stream focuses on categorizing ransomware threats rather than necessarily concentrating on detection algorithms [28], [29], [30].

- Prepare: “Identifying all active assets”
- Prevent: “Blocking common ransomware spread methods”
- Detect: “Alert an unauthorized access attempt”
- Remediate: “Initiate quarantine upon attack detection”
- Recover: “Visualization for phased recovery strategies”

The third stream engages with comprehensive strategies for countering ransomware techniques and tactics. Despite the title and the general subject of the paper, this survey addresses both crypto and locking ransomware types and includes some Android ransomware incidents. Since data from different papers cannot be compared because of different metrics and approaches to ransomware, the existing surveys strictly focus on crypto ransomware while noting the challenges of surveying this novel topic [31].

Despite efforts to identify ransomware early in the pre-encryption stage, existing solutions do not consider the dynamic nature of ransomware attacks. The evolution of zero-day attacks makes detection work more difficult [32], [33] . An adaptive pre-encryption detection system is therefore needed to identify crypto-ransomware attacks before they cause extortion [34]. Studying all the APIs before any encryption function was called, also known as pre-encryption APIs, was the data of interest in this research.

General ransomware detection and analysis system. First, a ransomware dataset sample is provided for pre-encryption, and then a feature extraction module that generates a feature representation vector. A feature reduction/selection process is conducted on the feature representation vector to obtain fixed dimensionality despite the length of the input sample for increased performance. Classification/clustering approach is trained using ransomware and benign samples that are currently available. Unseen samples are reported as ransomware or not during detection and analysis by the classification/clustering approaches to warn the user. Sometimes further analysis is conducted, such as outlining any suspicious (or advantageous) traits found in the sample. Ransomware detection analysis system is a cybersecurity tool designed to detect and prevent ransomware attacks in advance.

It uses signature- and behavior-based detection techniques to identify and stop known ransomware versions. Behavior-based detection observes program and process behavior to detect ransomware-like behaviors, such as widespread file-encrypting or shady network activity. Based on ransomware-specific patterns and qualities, machine learning algorithms may also be utilized to recognize fresh and developing ransomware outbreaks, as illustrated in Fig. 4.

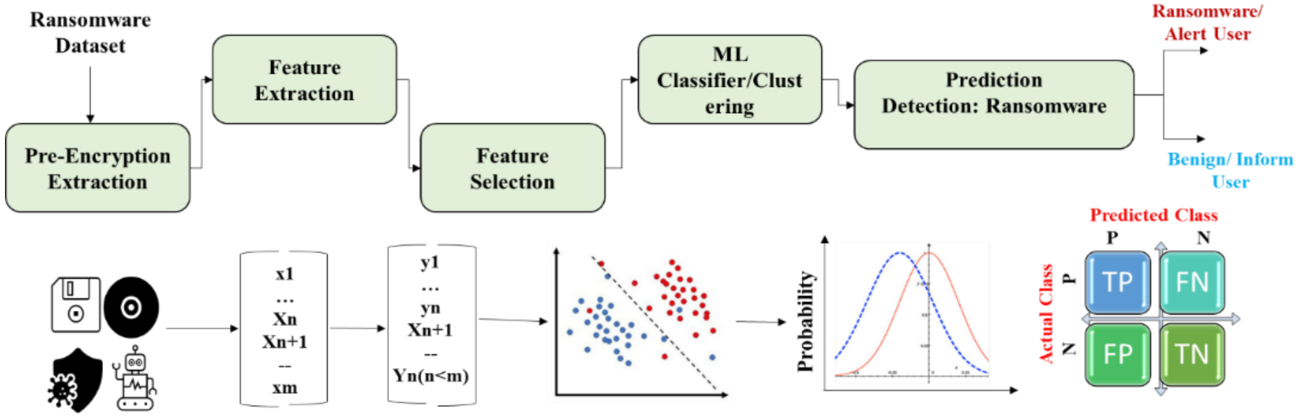


Fig. 4. Ransomware detection analysis system.

III. Research Methodology

The measures taken to examine earlier studies about ransomware attacks and detection systems are described in the methodology section. Also, inclusion and exclusion criteria were utilized to select the available research. The details of each phase of this investigation are provided in the sections that follows.

A. Systematic Literature Review

The PRISMA standards were used for the selection procedure, and the SLR guidelines were taken directly from [35]. The creation of review questions is the primary step. The next stage is to develop and evaluate a review technique, and then we will use the review protocol's criteria to look for primary screen studies as shown in Fig. 5. As Table IV shows different ransomware analysis tools.

TABLE IV. Ransomware Analysis Tools

Tools	Functions	Platform	Key Features	Pricing
IDA Pro	Disassembling and analysis	Win OS, Linux, macOS	Advanced disassembly and debugging capabilities	Contact for pricing
Cuckoo Sandbox	Automated malware analysis	Win OS, Linux	Dynamic behavioral analysis, threat intelligence integration	Open source
YARA	Pattern matching and detection	Win OS, Linux, macOS	Rule-based detection, custom signature creation	Open source
Wireshark	Network traffic analysis	Win OS, Linux, macOS	Packet-level analysis, protocol dissectors	Open source
Volatility	Memory forensics	Win OS, Linux	Memory image analysis, process, and DLL extraction	Open source
PEStudio	Static analysis of PE files	Win OS	Analysis of portable executable files	Free, Contact for advanced pricing
Ghidra	Reverse engineering and analysis	Win, Linux, macOS	Decomplication, the scriptable analysis environment	Open source
Procmon	Process monitoring and analysis	Win OS	Real-time process monitoring, event logging	Open source

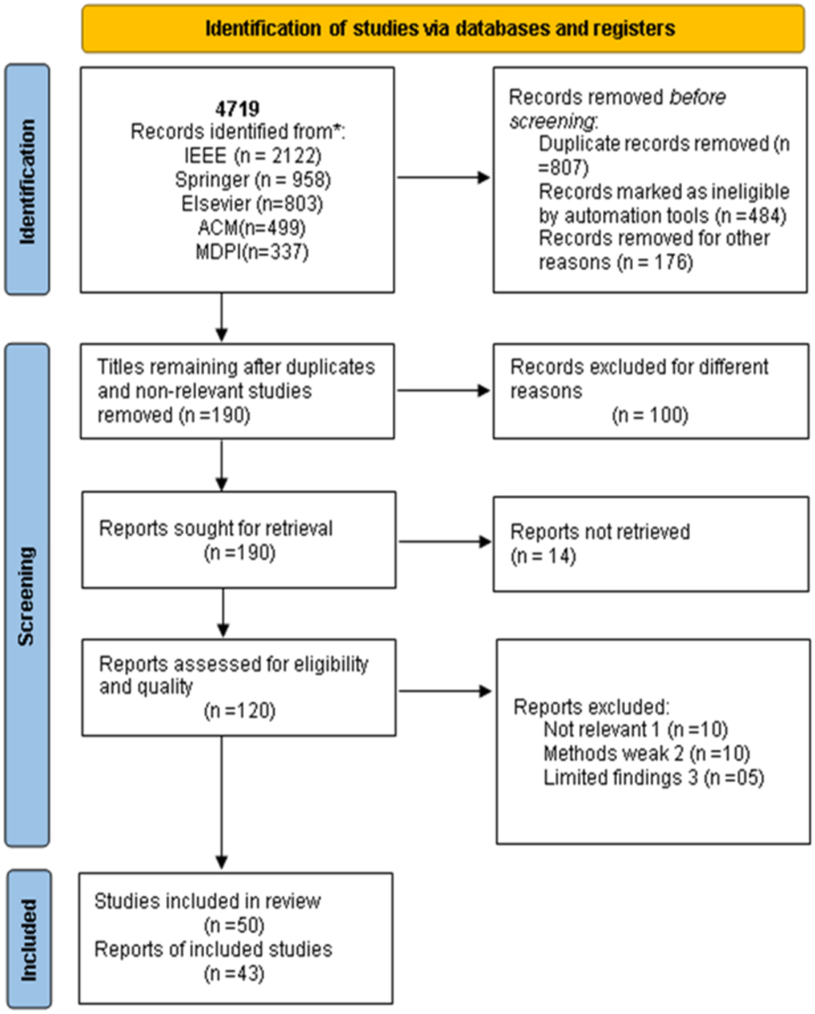

Identification of studies via databases and registers
Identification
4719
Records identified from*:
IEEE (n = 2122)
Springer (n = 958)
Elsevier (n=803)
ACM(n=499)
MDPI(n=337)
Records removed *before screening*:
Duplicate records removed (n =807)
Records marked as ineligible by automation tools (n =484)
Records removed for other reasons (n = 176)
Screening
Titles remaining after duplicates and non-relevant studies removed (n =190)
Records excluded for different reasons (n = 100)
Reports sought for retrieval (n =190)
Reports not retrieved (n = 14)
Reports assessed for eligibility and quality (n =120)
Reports excluded:
Not relevant 1 (n =10)
Methods weak 2 (n =10)
Limited findings 3 (n =05)
Included
Studies included in review (n =50)
Reports of included studies (n =43)

Fig. 5. Scoping literature review PRISMA.

B. Data Sources Information

The data sources utilized for the implementation of this article include IEEE Explore, ACM's digital library, Springer, Elsevier, MDPI, and online libraries. A search string is used to browse the code by the recommendations [36], [37]. The information sources that are mentioned in this review have been picked because they have high-quality, high-impact articles. Data sources were looked up in May 2023 utilizing sophisticated search tools. Table V illustrates searched databases sources.

TABLE V. SEARCH DATABASE SOURCES

Electronic Database	URLs
IEEE Xplore	(http://ieeexplore.ieee.org)
ACM Digital Library	(http://dl.acm.org/)
ScienceDirect	(http://www.sciencedirect.com)
Springer	(http://www.springer.com)
MDPI	https://www.mdpi.com/journal/futureinternet/special issues/SEO

IV. RESEARCH QUESTIONS

The goal of this study is to examine and assess several machine-learning algorithms for new ransomware detection. Research questions and research objectives (RQs and ROs) have been made to be emphasized in this SLR in Table VI.

TABLE VI. FORMULATED RQS AND ROS

No	Research Questions	Objectives
RQ1	State the current limitations in existing ransomware detection techniques that affect during the early phases.	RO1 To analyze and identify the current limitations and challenges in existing ransomware detection techniques, with a specific focus on their impact on the early phases of new ransomware attacks.
RQ2	What factors contribute to the improvement of pre-encryption for new-ransomware detection?	RO2 To Explore machine learning algorithms to detect unusual pre-encryption ransomware activities.
RQ3	How can the pre-encryption of ransomware be improved using machine learning and non-machine learning?	RO3 To identify the recent advances and techniques to overcome and improve the issue in new-ransomware pre-encryption prevention, and detection.

A. Review Protocol

The SLR metrics produced by [35] search strategy, inclusion and exclusion criteria, quality assessment, data extraction, and data analysis serve as the foundation for the review procedures.

B. Search Strategy

The most pertinent keywords and their variants were used to create the search string by the main goals of this study. Boolean operators and the keywords that were specified were used to create the search query. The search parameters and query strings are shown below in Fig. 6.

- TITLE-ABS-KEY "pre-encryption" OR "ransomware" OR "ransom" OR "malware" AND "security" ((‘‘cybersecurity’’ OR ‘‘security’’ OR ‘‘pre-encryption’’ OR ‘‘ransomware’’ OR ‘‘ransom’’) AND (‘‘ransomware detection’’ OR ‘‘ransom-ware’’ OR ‘‘malware’’ OR ‘‘encryption’’) AND (‘‘machine learning’’ OR ‘‘deep learning’’ OR ‘‘information security’’))
- The timespan to collect the studies is from 2018 to 2023 H1.
- The survey is in the English language medium.
- After finalizing the search terms, the appropriate digital repositories were chosen. We conducted searches across five electronic databases, which are listed below.
  - IEEE Xplore
  - ACM Digital Library
  - MDPI
  - Springer
  - ScienceDirect

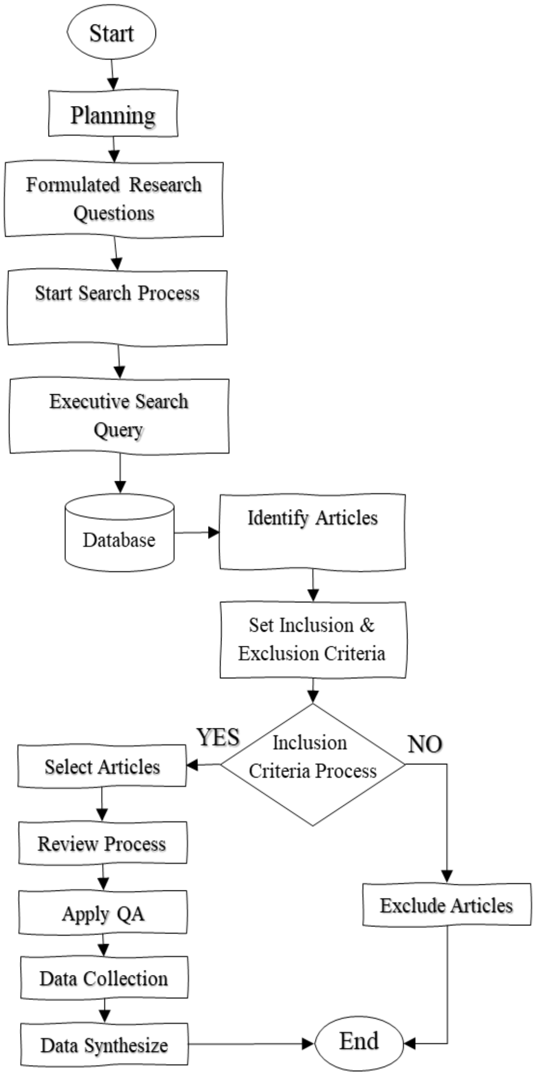


Fig. 6. Scoping review process.

The previously listed five electronic databases, which include the major publications and conferences, are searched. Second, while studies only make up just a small portion of the major research, we also compile the studies that are connected to pre-encryption ransomware detection using static analysis in the reference section. The search period covers from January 2018 to April 2023, and all research related to search phrases has been considered.

C. Inclusion and Exclusion Criteria

For the inclusion and exclusion criteria for this research, to focus on the most important criteria, scholarly works for this SLR. The shortlist is shown in Table VII. According to their capacities, studies are used to determine whether inclusion and exclusion satisfy the requirements.

TABLE VII. INCLUSION AND EXCLUSION CRITERIA

No	Inclusion Criteria	Exclusion Criteria
1	A research article published in English that focuses on the activities, assaults, defenses, and detection methods of ransomware in the Windows operating system, as well as data from peer-reviewed, reputable publications or conference papers included in the above-mentioned databases.	Analyze information from news and magazine publications, non-English articles, and information on the latest ransomware variants and vulnerabilities in other operating systems and mobile devices.
2	The article offers insights and practical advice to protect against ransomware attacks and other cyber threats and the early stage.	The article should discuss papers that investigate the impact of ransomware attacks on businesses or the legal system.
3	The document should provide an in-depth examination of ransomware or any other relevant technological advancement in your writing.	Governmental documents and blogs should not be included in the article.

Three steps made up the selection process. The first step was to look for any possible primary studies. The next step involved examining and reading the titles. Abstracts of all the papers that were returned by the search. So, we discovered every piece of research that met the requirements for inclusion and exclusion. After that every study that had been found was read already being selected for final selection.

Diagram illustrating the preferred reporting items for systematic reviews and Meta-Analysis (PRISMA) flow process showing the ultimate number of studies included in the systematic review and meta-analysis as well as the inclusion and exclusion of studies. Fig. 7 shows selection criteria for this study.

D. Quality Assessment Criteria

We screen the chosen studies following the quality assessment criteria and Score listed in Table VIII, Table IX to evaluate their quality. We determine whether the chosen studies meet these requirements using the cross-checking method to guarantee the reliability of the results. The final studies, which include, are obtained following the stage of quality assessment criteria, concerning the detection of ransomware, there are 103 studies and 2 SLRs related to this.

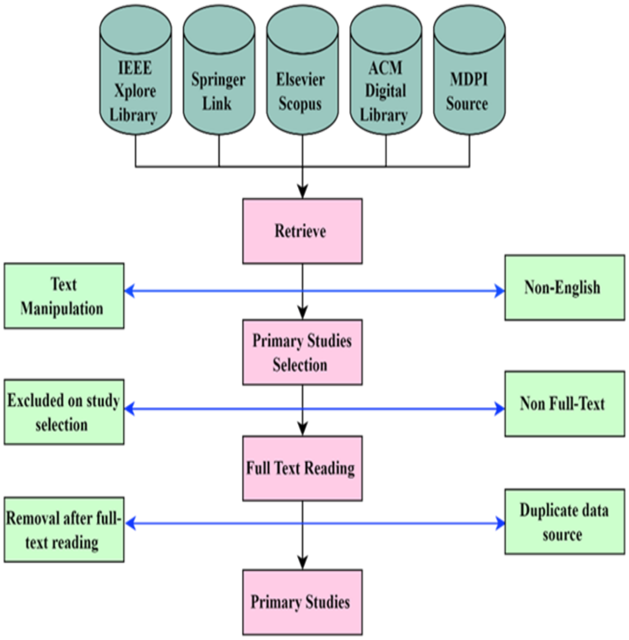


Fig. 7. Study selection criteria.

TABLE VIII. QUALITY ASSESSMENT CRITERIA

No	Quality Assessment Criteria
1	Is the study's direction clear?
2	Is the approach for static and dynamic well stated?
3	Do the experimental datasets provide clear descriptions?
4	Exactly what features are being used?
5	Is the model mentioned clearly?
6	Do the empirical experiments provide a clear description?
7	Are performance metrics given in a transparent manner?
8	Does the research study contribute to this SLR?

TABLE IX. QUALITY ASSESSMENT SCORE

No	Category	Result
1	Systematically Adopted	3
2	Reviewed effectively	2
3	Minor declared	1
4	Does not mention	0

E. Data Extraction

To support the study questions, the required data was acquired, and a detailed analysis was conducted. The following details were extracted from the selected primary studies and entered an extraction form that had been pre-made. Fig. 8 shows extraction of complete information.

- Class, Info, and reference ID
- Publication name
- Country of organization
- Authority of research

- The type of machine learning methods used to mitigate ransomware.
- Algorithms, models, and ideas are essential.
- Categorization of machine learning algorithm with a certain approach and analysis type.
- For ransomware detection process tools were used.

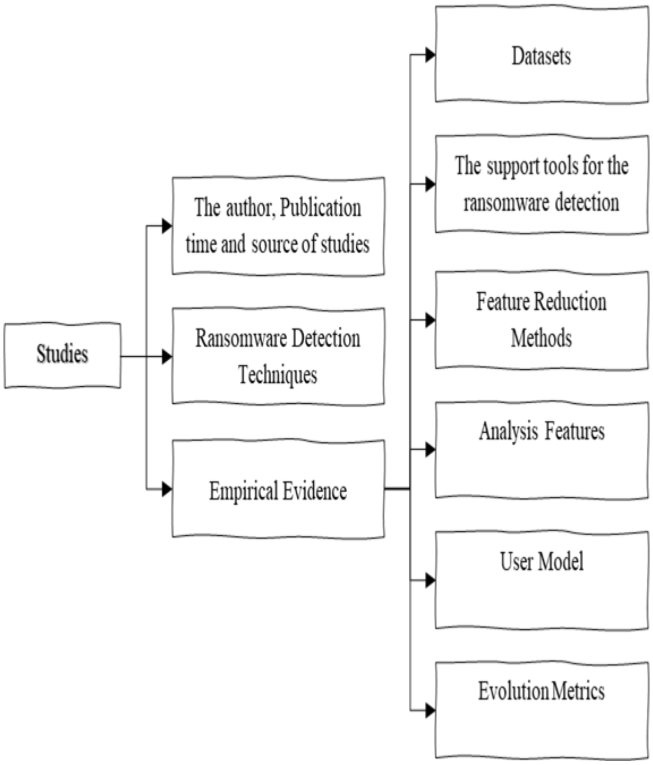

Studies
The author, Publication time and source of studies
Ransomware Detection Techniques
Empirical Evidence
Datasets
The support tools for the ransomware detection
Feature Reduction Methods
Analysis Features
User Model
Evolution Metrics

Fig. 8. Information extraction.

F. Data Analysis

Each main study's data was extracted, and then each research question was addressed with thorough data analysis. Machine learning algorithms that had been implemented were identified to respond to RQ1, and their effectiveness was assessed to respond to RQ2. Related theories or models were found for each category and key features of successful pre-encryption detection as RQ2.1. The outputs of the algorithm were evaluated in terms of their ability to respond to RQ3.

1) Difficulty of problem in practice: It is highly recommended to study the idea of ransomware camouflage to learn more about and develop the field of malware analysis and new ransomware detection. Malware camouflage involves using techniques to hide harmful code, extending its undetected presence by eluding conventional malware detection tools. Malware authors use a variety of strategies, from straightforward ones like encryption to more complex ones like metamorphism. For academics and security experts to create efficient defenses against developing malware threats and improve overall detection methods, they must be aware of various disguise strategies [20].

2) Sophisticated techniques: In the context of ransomware, sophisticated tactics relate to sophisticated and complicated ways employed by attackers to conduct effective and evasive ransomware operations. These strategies aim to circumvent established cybersecurity defenses and increase the difficulty of discovery, prevention, and recovery.

3) Encryption and data exfiltration: Data security and privacy are seriously threatened by ransomware attacks, which are crucially based on encryption and data exfiltration. Effective cybersecurity measures need a thorough understanding of data exfiltration risks, ransomware use of encryption, and related issues.

4) Impact on critical systems: New ransomware attacks can have severe consequences, especially when targeting critical infrastructure systems, healthcare institutions, or government agencies. Disruptions caused by new ransomware can lead to financial losses, endanger lives, or compromise sensitive national security information.

V. Ransomware Detection Taxonomy

The methodologies utilized to identify ransomware, the operation of the machine learning algorithm, the performance outcome, the classification strategy, and the chosen analysis type used to respond to RQ1 through RQ3 are all covered in this part.

The motivating methodology, findings, restrictions, and future directions of the investigated approaches were all covered in the authors' assessment of the ransomware detection methods put out in the literature. They also examined several ransomware detection methods about factors including the operating system for mobile devices and PCs, the cloud, data sources, various machine learning algorithms in use, and result and assessment standards. Fig. 10 demonstrates the ransomware detection environments, along with the many standards and related metrics. The comparison charts of the detection environment, data analysis, machine learning, results, and assessment criteria charts are shown in Fig. 9 to 13.

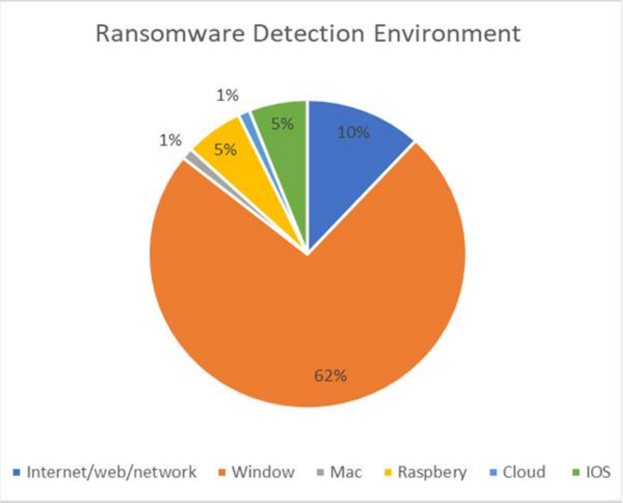


Fig. 9. Ransomware detection environments.

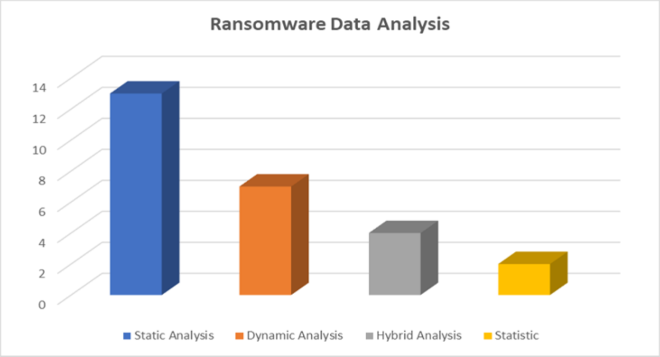


Fig. 10. Ransomware data analysis understanding.

A. Early Detection

Early detection of ransomware is crucial for preventing data encryption and minimizing attacks. The Pre-Encryption Detection Algorithm (PEDA) is a machine learning-based algorithm that detects ransomware behavior patterns before the encryption process begins. This helps identify patterns and characteristics indicative of ransomware before it can cause severe damage. Runtime data analysis captures runtime data during the initial phases of ransomware attacks, allowing for the identification of patterns and indicators of ransomware. However, challenges like accurately defining pre-encryption phases and limited data availability require further research to develop more robust techniques. Combining machine learning algorithms like PEDA with runtime data analysis can contribute to the early detection of ransomware and improve the effectiveness of preventive measures [16].

Ransomware detection environments play a crucial role in safeguarding organizations and individuals from the ever-increasing threat of ransomware attacks. These detection environments are designed to identify and mitigate ransomware activities during the initial stages, before encryption occurs, and severe damage is inflicted.

“Ransomware data analysis understanding" refers to the process of examining, interpreting, and making sense of data related to ransomware attacks. This involves delving into various aspects of ransomware incidents, and analysis techniques as shown in Fig. 10.

A thorough taxonomy of ransomware detection is comparable to an orderly road map of the various kinds, techniques, and strategies applied to the identification and mitigation of ransomware threats in Fig. 11 and classifies the various methods, tools, and approaches used in cybersecurity.

Comparison of the effectiveness of different machine learning classifiers used especially for ransomware detection. It suggests assessing various models or algorithms for ransomware classification, stressing their relative performance and efficacy as shown in Fig. 12.Machine learning classifiers.

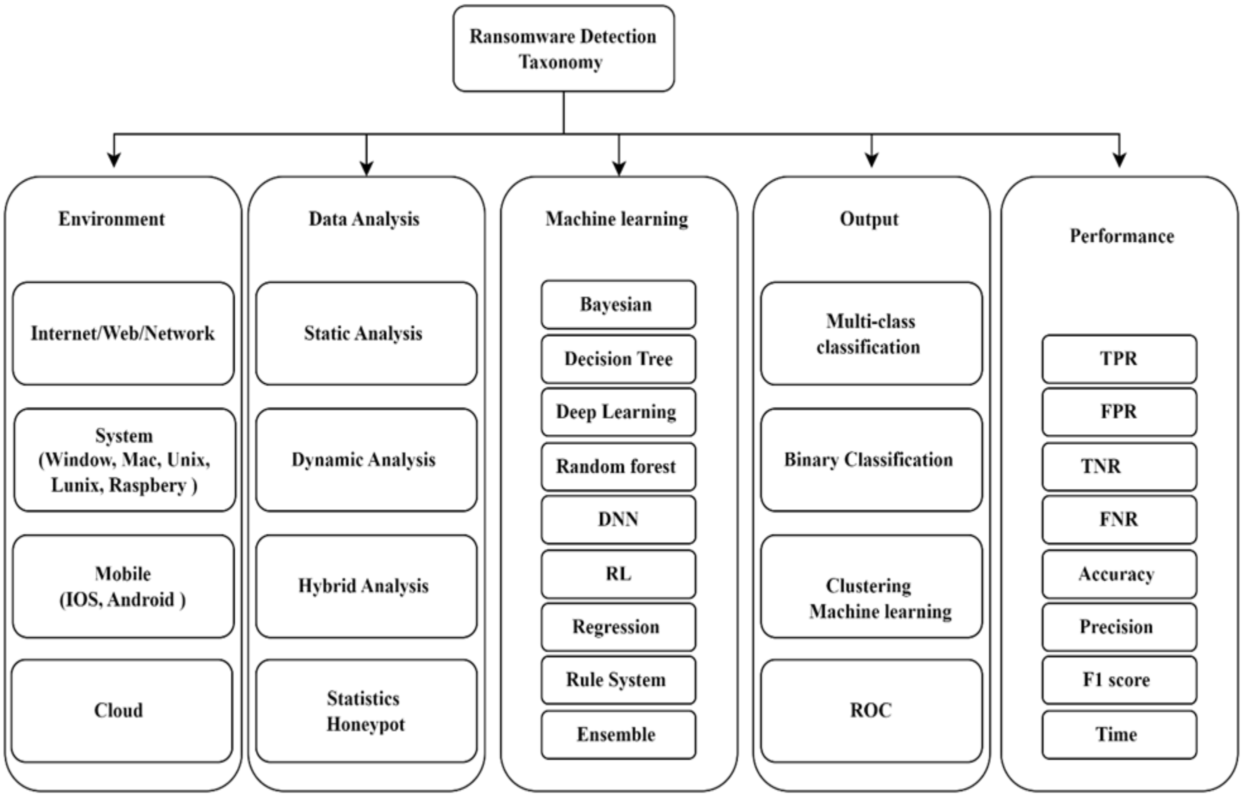


Fig. 11. A comprehensive ransomware detection taxonomy assists cybersecurity.

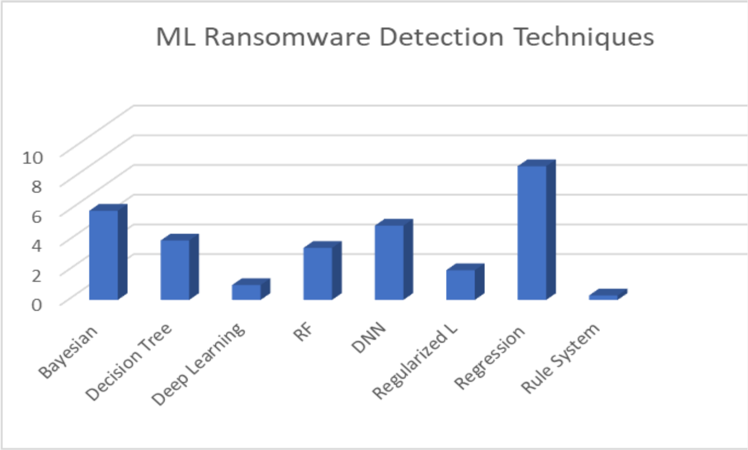


Fig. 12. Machine learning classifiers.

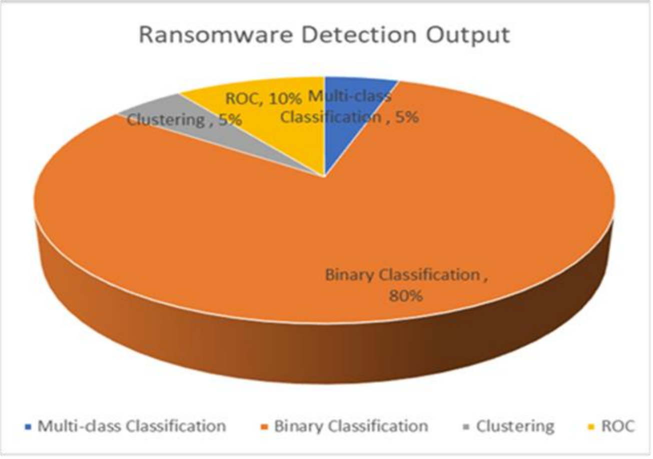


Fig. 13. Ransomware output responding to threats promptly.

Outcomes produced by a system for detecting ransomware. It highlights how quickly the system can detect any ransomware threats. The picture illustrates the steps involved in detecting these threats and taking appropriate action in response, emphasizing how crucial prompt and efficient action is in lessening the impact of ransomware attacks.

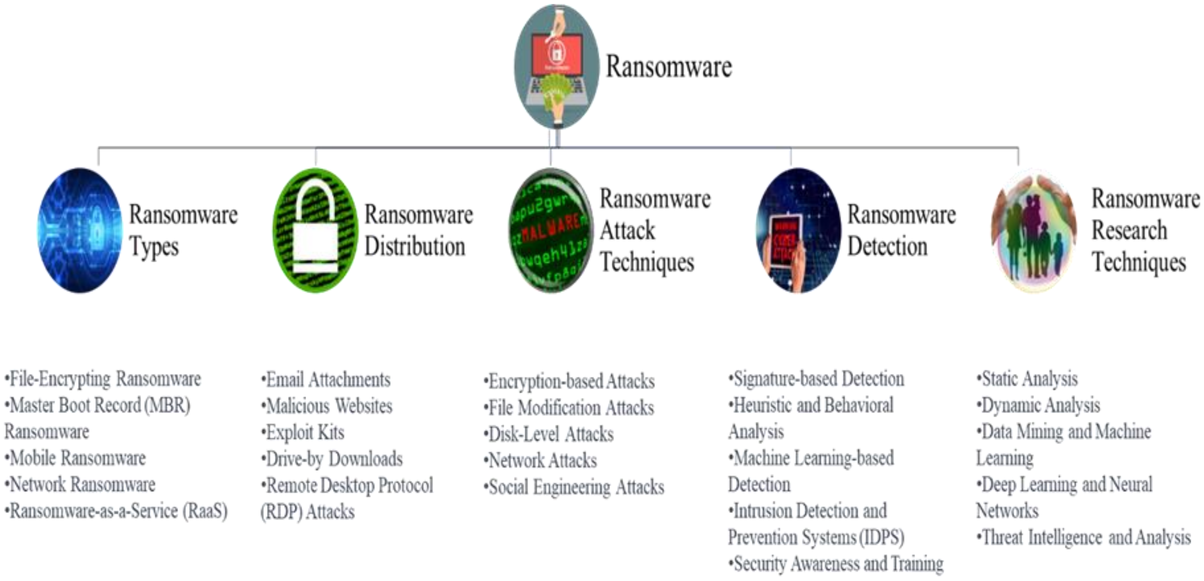

Ransomware
Ransomware Types
Ransomware Distribution
Ransomware Attack Techniques
Ransomware Detection
Ransomware Research Techniques
•File-Encrypting Ransomware
•Master Boot Record (MBR) Ransomware
•Mobile Ransomware
•Network Ransomware
•Ransomware-as-a-Service (RaaS)
•Email Attachments
•Malicious Websites
•Exploit Kits
•Drive-by Downloads
•Remote Desktop Protocol (RDP) Attacks
•Encryption-based Attacks
•File Modification Attacks
•Disk-Level Attacks
•Network Attacks
•Social Engineering Attacks
•Signature-based Detection
•Heuristic and Behavioral Analysis
•Machine Learning-based Detection
•Intrusion Detection and Prevention Systems (IDPS)
•Security Awareness and Training
•Static Analysis
•Dynamic Analysis
•Data Mining and Machine Learning
•Deep Learning and Neural Networks
•Threat Intelligence and Analysis

Fig. 14. Ransomware techniques overview.

Fig. 14 presents a summary of the ransomware approaches discussed in this section, categorized by approach type, analysis features, and availability. However, upon thorough examination of the literature, it became evident that previous studies had certain limitations. Specifically, there was a lack of research focusing on ransomware, with most works primarily utilizing static analysis for detection. Moreover, since ransomware can evade static analysis through code obfuscation techniques, it is crucial to incorporate dynamic analysis. Unfortunately, existing dynamic analysis tools target malicious programs rather than ransomware, and some tools are either immature, outdated or only accessible commercially. Consequently, a need to propose a hybrid system that investigates the effectiveness of integrating techniques, and static and dynamic analysis to detect ransomware more efficiently and accurately, thereby safeguarding system users from falling victim to such attacks. This hybrid system should incorporate various established static analysis approaches and evaluate their ability to differentiate between ransomware apps and benign apps. Based on the results of the static analysis, a decision can be made regarding the need for additional dynamic analysis on these apps. Furthermore, careful consideration should be given to selecting appropriate tools for conducting dynamic analysis. The primary objective is to achieve accurate pre-encryption ransomware detection while minimizing costs. Table X contains the different ransomware detection approaches.

TABLE X. PRE-ENCRYPTION RANSOMWARE DETECTION METHODS

Detection Method	Description
File Signature	Finds ransomware based on known file signatures
Behavior Analysis	Monitors unusual file access or encryption behavior
Heuristic Analysis	Finds potential ransomware based on patterns
Machine Learning	Uses algorithms to detect ransomware-like behavior
Sandbox Analysis	Executes files in a controlled environment for detection

Metrics used to evaluate the efficacy of ransomware detection and mitigation techniques are included, such as detection accuracy, false positive rates, mitigation speed, and recovery efficiency. An examination of these metrics across time or several approaches may be shown in Fig. 15 along with trends.

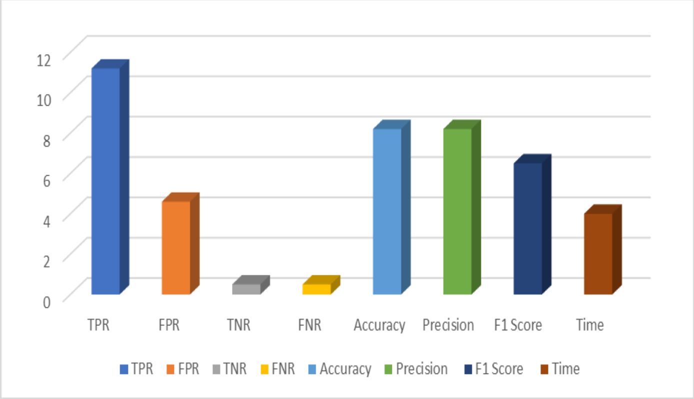


Fig. 15. Performance evaluation metrics of ransomware detection and mitigation.

B. Ransomware Detection Based on Machine Learning

Preventing ransomware is difficult for several reasons. Ransomware often mimics the behavior of legitimate software, operating in a covert manner. As a result, detecting ransomware in zero-day attacks has become a critical priority. The main goals are to prevent system damage caused by ransomware, identify previously unknown malware (zero-day attacks), and reduce detection time. Various tools and techniques are available for detecting ransomware. Static analysis methods, for example, examine source code without executing it. However, these methods tend to produce many false positives and struggle to detect ransomware that has been obfuscated. As Table XI represent existing detection techniques. Attackers frequently develop new variants and modify their code using different packing techniques. To address these challenges, researchers have turned to dynamic behavior analysis, which observes how executed code interacts with a virtual environment. While effective, these methods can be resource-heavy and slow. Machine learning, by contrast, excels at analyzing the behavior of applications or processes. Several machine learning-based detection systems follow well-established methodologies: Table XII summarizes previous studies on Machine learning techniques (behavioral techniques) for ransomware detection.

TABLE XI. EXISTING RANSOMWARE DETECTION TECHNIQUES

Study	Year	Features used	Static	Dynamic	Available
[38]	2018	Employs Droid Bot, test response creator, and API Packages	×	✓	×
[39]	2018	UI widgets, users' finger activities	×	✓	×
[40]	2019	Text, sysadmin, win pro, sys Opp, Priority, Consent	×	✓	×
[15]	2020	27 API-level, permissions	✓	×	✓
[41]	2020	General features in Static	✓	×	✓
[42]	2020	API call level 27	✓	×	✓
[43]	2021	API call level 30s	✓	×	✓
[48]	2022	ML-based API Calls	×	✓	✓
[44]	2021	API call level 30	✓	×	✓

TABLE XII. COLLECTED STUDIES ON ML-BASED RANSOMWARE DETECTION

Study	Features	AI Techniques												Accuracy
		ML Classifier								DL Techniques				
		DT	RF	GB	NB	SVM	LR	KNN	XGB	LSTM	ANN	RNN	MLP	
[45]	Network traffic	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	99.8%
[46]	API calls	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	98.63%
[47]	Access privileges, read/write/implement/copy	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	96.28%
[48]	API calls	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗	94.9%
[15]	API calls	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	✓	97.08%
[19]	API	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	--
[49]	IRPs	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	96.6%
[50]	Opcode sequence	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	99.3%
[39]	API calls packages	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	97%
[11]	APIs, IRPs	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	--
[51]	C&C, no of bytes read/written	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	99.9%
[52]	Power/energy consumption patterns	✗	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	83.7%
[53]	API calls	-	-	-	-	-	-	-	-	✓	✗	✓	✗	93%
[56]	System logs, network logs	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	98.5%
[54]	DLL, function calls assembly levels	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	99.7%
[55]	Raw byte	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	97.7%

1) *File behavior analysis:* Machine learning algorithms can analyze the behavior of files on a system to detect ransomware. By creating baselines of legitimate code executions, the algorithms can detect any behavior that deviates from those baselines.

2) *Network traffic analysis:* Algorithms based on machine learning can analyze network traffic to detect ransomware. By monitoring the traffic patterns and identifying anomalies, the algorithms can detect ransomware attacks.

3) *Dynamic feature dataset:* A dynamic feature dataset can be used to detect ransomware using machine learning algorithms. The dataset contains features that are extracted from the binary file of the ransomware. By analyzing these features, the algorithms can detect ransomware attacks.

4) *Multi-classifier network-based system:* A multi-classifier network-based system can be used to detect ransomware. The system uses machine learning algorithms to analyze the behavior of files and network traffic. By combining the results of multiple classifiers, the system can improve the accuracy of ransomware detection.

C. Limitations of Machine Learning Based Ransomware Detection

The limitations of machine learning-based ransomware detection can be summarized as follows:

1) *False negatives:* Machine learning algorithms can sometimes fail to detect ransomware attacks, resulting in false negatives. This can be due to the lack of training data or the inability of the algorithm to detect new variants of ransomware.

2) *Limited dataset:* The accuracy of machine learning algorithms depends on the quality and quantity of the dataset used for training. A limited dataset may not capture all the variations of ransomware, leading to inaccurate results.

3) *Overfitting:* Machine learning algorithms can sometimes overfit the training data, resulting in deficient performance on new data. This can be due to the algorithm's complexity or the lack of regularization.

4) *Encryption:* Ransomware attacks often involve the encryption of files, which can make it difficult for machine learning algorithms to detect them. This is because the encrypted files may not contain the same features as the original files.

5) *Adversarial attacks:* Adversarial attacks can be used to evade ransomware detection based on machine learning. Attackers can modify the ransomware code to bypass the detection algorithm.

Study has explored the use of machine learning algorithms such as the J48 decision tree and random forest to detect and classify different ransomware families based on TCP malware network traffic [45]. Another study introduced a new approach called WmRmR to detect early ransomware, effectively evaluating the fundamental characteristics of large-scale datasets at low complexity and false positive rates [46]. A related study proposes a detection method focused on analyzing access privileges in process memory and enabling accurate and efficient identification of key functions of ransomware [47]. In the field of ransomware classification, researchers developed an advanced technique to exploit the suspicious behaviors displayed by ransomware, in particular several API requests to find an optimal execution environment. By generating fingerprints of these behaviors from more than 3,000 recently known ransomware samples, the authors achieved an impressive classification accuracy of 94.92% [48]. The redundancy coefficient gradual upweighting (RCGU) approach improves the selection of crypto-ransomware detection features by dynamically adjusting the weight of redundancy terms. The combination of RCGU and other mutual information methods further improved accuracy compared to previous studies [15]. Several studies focused on the detection of Android ransomware. Researchers have been using decoy techniques to detect ransomware in real-time, monitor file systems and running processes, and identify and prevent benign file changes from triggering alerts based on learned encryption behavior [49]. Developed a classification model. Using N-gram sequences from ransomware sample opcode sequences, it is possible to classify families more accurately [50]. The API-based ransomware detection system (API-RDS) was used to study the static and dynamic analytical approach of ransomware detection in mobile devices. Although this approach has not been put into practice or proven through simulation, the author presented it as a framework for the early detection of ransomware, considering the temporal correlations between IRPs and APIs. In the context of network traffic analysis [11]. This study presented a detection approach based on the analysis of file-sharing traffic, effective detection, and prevention of crypto-ransomware activity [51]. The author introduced a unique method for detecting Android ransomware with energy consumption levels [52]. The study incorporated an attention mechanism in learning malware sequences to detect ransomware based on repetitive patterns associated with repeated encryption [53]. The author proposed artificial intelligence-powered hybrid models that would overcome the challenges of detecting ransomware using functions such as assembly, dynamic link libraries, and function calls [54]. Based on the ease of static malware analysis, an approach based on data mining techniques in particular, frequent pattern mining was developed to identify ransomware. Similarly, a pre-distributed model was created using convolutional neural networks to classify binary items and improve performance using transfer learning [55].

D. Non-Machine Learning Based Ransomware Detection

The term "non-machine learning-based ransomware detection" refers to techniques that do not rely on machine learning algorithms but rather are conventional and rule-based. This method is useful in some circumstances since it makes use of predetermined rules, patterns, and heuristics to find ransomware activity. Known signatures or patterns of previously recognized ransomware variants are used to identify and stop ransomware in one popular technique called signature-based detection. Another method is behavior-based detection, which keeps an eye on system activity for ransomware-specific suspicious behaviors such as quick file encryption. Furthermore, even if the precise ransomware strain is unknown, heuristics may be used to spot ransomware-like behavior. Non-machine learning-based techniques may be able to provide quick detection and reaction capabilities shown in Table XIII, but they could have trouble spotting new or polymorphic ransomware versions. The authors also discussed the integration of this contextual detection technique into digital forensics for ransomware mitigation and prevention [57].

A software-defined network (SDN)-based detection technique for Windows computers was also demonstrated in [58]. The technique extracts pertinent HTTP message sequences as key features from network traffic between the crypto-ransomware variants Crypto Wall and Locky. The reading and writing activities of backup files and ransomware samples with significant read/write operations are tracked. The context-aware detection model uses entropy data to spot unusual file activity [24]. The basis for the detection is manipulation files (such as desktop files and/or user files). The system creates a fake user environment and can identify file modifications caused by ransomware. The system keeps track of system modifications as well as their behavior. The detection can spot previously unreported, unknown (zero-day), and evasive ransomware. Passively observes traffic produced by 19 ransomware families using a network prober. Less than 10 files are lost prior to the ransomware activity being detected by the model, which focuses on early detection [59] and examines the characteristics of cryptographic ransomware. To stop ransomware, they suggested deceptive file protection methods. Incorporated a dynamic analysis-based automated malware detection technology. The latter extracts a call to the Application Programming Interface (API) from logs to find ransomware [62]. The researchers demonstrated how their techniques could enhance the automatic analysis of numerous malware samples [63], [64]. To categorize tweets to fulfill the requirement for file protection on rootless devices, they developed and deployed KRProtector, which can recognize ransomware and protect files using deception [65], [66]. The author used static and dynamic analysis of the executable malware to extract both static and running-time behavior. Reverse engineering is used to extract binary signatures using the CRSTATIC model. No matter what kind of malware is being assessed, the authors show that using YARA rules with fuzzy hashing can enhance the evaluation's outcomes [67], [68], [69], [70].

TABLE XIII. COLLECTED STUDIES ON NON ML-BASED RANSOMWARE DETECTION

Study	Methods	Features	Evaluation metrics	Correctness	Platform	Environment
[57]	Rule-based	API calls DLL libraries windows registry	Trigger threshold CAT	-	OS	Cuckoo sandbox
[58]	Software-defined network	HTTP	ROC curve	-	OS-7	Cuckoo Sandbox, VMware
[24]	Hardware-based	I/O requests	Accuracy	96.3%	OS	Cuckoo sandbox
[59]	Rule-based	IP traffic	Overhead detection rate, file lost	100%	OS-7	Virtual
[60]	Decoy-based	File access read/write/remove	Precision Accuracy	96.2%	IoT Android	Real in Android 7.1
[61], [64]	Forensic based	Network sign Function calls	-	-	OS	Real in testbed

VI. Results

The results of primary research are intended to be presented in this section. We begin by outlining the main studies. In terms of fact, we next provide the SLR's findings considering the study's questions.

A. *Study Description*

103 Studies are addressed in this section in terms of publication time.

B. *Publication Time*

Fig. 16 demonstrates that there were 26, 25, 25, 23, 10, and 26, respectively, studies from 2018 to 2023-H1. This data shows that the number of studies in 2018 accounts for the biggest share. The number of research connected to ransomware detection using static analysis and dynamic is increasing from 2018 to 2023, except for certain papers in 2023 (some publications in 2023 are not released, thus the time of these papers in 2023 is from January to May) [71], [72], [73]. This data indicates that pre-encryption ransomware detection has consistently been a popular issue in recent years.

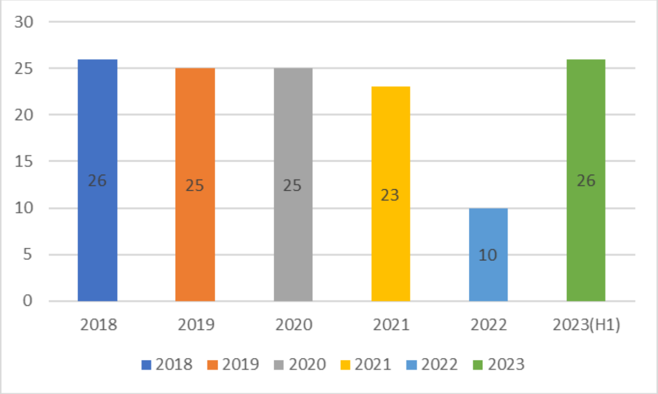


Fig. 16. Year-wise distribution of studies.

RQ1: What are the current limitations/challenges in existing ransomware detection techniques that affect during the early phases?

There are several limitations and challenges in existing ransomware detection techniques that affect early detection. These include:

- Signature-based detection is easily bypassed by malware authors. Signature-based detection relies on identifying known malicious files or patterns. However, malware authors can easily obfuscate their code to evade detection by signature-based tools [74], [75], [76], [77].

- Behavior-based detection can be triggered by legitimate applications. Behavior-based detection looks for suspicious or malicious behavior, such as file encryption or network traffic patterns. However, legitimate applications can also exhibit these same behaviors, which can lead to false positives [78], [79], [80], [81].

- Ransomware attacks are often targeted and stealthy. Ransomware authors often target specific organizations or individuals, and they may take steps to conceal their attack. This can make it difficult for detection tools to identify the attack in its initial stages [82], [83], [84], [85].

- Ransomware is constantly evolving. Ransomware authors are constantly developing new techniques to evade detection. This makes it difficult for detection tools to keep up with the latest threats.

As a result of these limitations and challenges, it can be difficult to detect ransomware in its initial stages. This is why it is important to have a layered security approach that includes a variety of detection techniques. By combining signature-based, behavior-based, and other detection techniques, organizations can improve their chances of detecting ransomware early and preventing a successful attack [86], [87], [88], [89].

- Insufficient data and attack patterns

- Evolving tactics and techniques

- Lack of awareness

- Visibility into systems

The current limitations/challenges in existing ransomware detection techniques during the early phases include insufficient data and attack patterns, evolving tactics and techniques, limited detection capabilities, lack of awareness, and limited visibility into systems. These challenges require innovative solutions and collaborative efforts to combat the rise of ransomware attacks.

RQ2: What factors contribute to the improvement of pre-encryption ransomware detection?

Improving pre-encryption ransomware detection requires the development of advanced detection techniques, such as behavior matching, machine learning, detection of symmetrical and asymmetrical encryption, early detection, high detection rate, and continuous updates. These factors can help improve the detection and prevention of ransomware attacks.

1) Pre-encryption detection algorithms: Pre-encryption detection algorithms, such as the Pre-Encryption Detection Algorithm (PEDA), can detect ransomware before it starts its encryption function. These algorithms use machine learning or behavior matching to identify patterns in the ransomware code and create a signature repository to detect future attacks [19].

2) Adaptive models: Adaptive models that combine machine learning and non-machine learning techniques can improve pre-encryption ransomware detection. For example, an adaptive pre-encryption crypto-ransomware early detection model uses both machine learning and non-machine learning techniques to detect ransomware before it can be executed [90], [91].

3) Behavior matching: Behavior matching can be used to detect small variants of unknown crypto-ransomware. This approach involves comparing the behavior of a file to a known set of behaviors associated with ransomware [92], [93], [94].

4) Improved visibility: Improved visibility into network activities can help detect ransomware during its early phases. This involves monitoring, aggregation, correlation, and analysis of network activities to identify suspicious behavior [95], [96], [97].

Continuous research and collaboration: Continuous research and collaboration are needed to stay ahead of evolving ransomware tactics and techniques. This includes sharing threat intelligence and developing new detection techniques to address emerging threats and evaluation metrics for crypto-ransomware.

Q3 How can the pre-encryption of ransomware be improved using machine learning and non-machine learning?

Improving the pre-encryption detection of ransomware can be achieved through the integration of both machine learning and non-machine learning techniques. Machine learning approaches enhance pre-encryption ransomware detection by leveraging advanced algorithms to analyze data and identify patterns indicative of ransomware behavior. Through feature engineering, anomaly detection, and deep learning models, machine learning can detect ransomware with higher accuracy and adapt to new variants. Ensemble methods and continuous learning mechanisms further enhance the detection capabilities. On the other hand, non-machine learning approaches such as signature-based detection, heuristics, behavioral analysis, and network traffic analysis provide additional layers of defense. By combining these approaches, organizations can leverage the strengths of both methods. As shown in Fig. 17, detection system analyzes ransomware behaviors, utilizing classifiers.

Non-machine learning techniques offer rule-based detection and proactive measures such as authorization, block-listing, and user education. Integrating machine learning with non-machine learning techniques creates a comprehensive defense strategy that improves pre-encryption ransomware detection, providing more effective protection against emerging threats and reducing the risk of successful attacks [98], [99], [101]. Machine learning techniques have progressively been widely used for ransomware detection in recent years due to the rapid growth of these techniques in natural language processing, image recognition, and other areas. Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Ensemble Learning (EL), and neural networks are some of the machine learning models that are frequently utilized in primary investigations [100], [102], [103]. Table XIV shows ML and Non-ML based methods.

By utilizing a combination of machine learning and non-machine learning approaches, organizations can improve pre-encryption ransomware detection, providing more robust and proactive defenses against ransomware threats. The integration of these methods complements each other, resulting in a comprehensive approach that enhances detection accuracy and responsiveness, thus minimizing the potential impact of ransomware attacks (Fig. 18) [93].

To extract configuration, "MalConfScan with Cuckoo" launches malware on the host computer. MalConfScan can extract the configuration of recognized malware from a memory image that is dumped when malware is registered on Cuckoo and executed on the host computer. A report will then display the extracted configuration that can be seen in Fig. 18 [93].

JPCERTCC/MalConfScan-GitHub
https://github.com/JPCERTCC/MalConfScan-with-Cuckoo

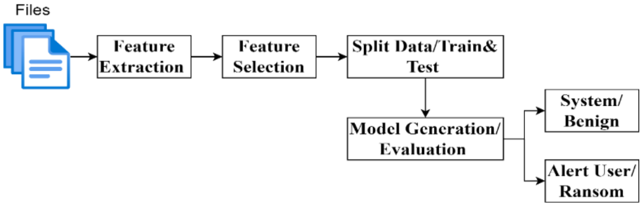


Fig. 17. Machine learning detection for ransomware.

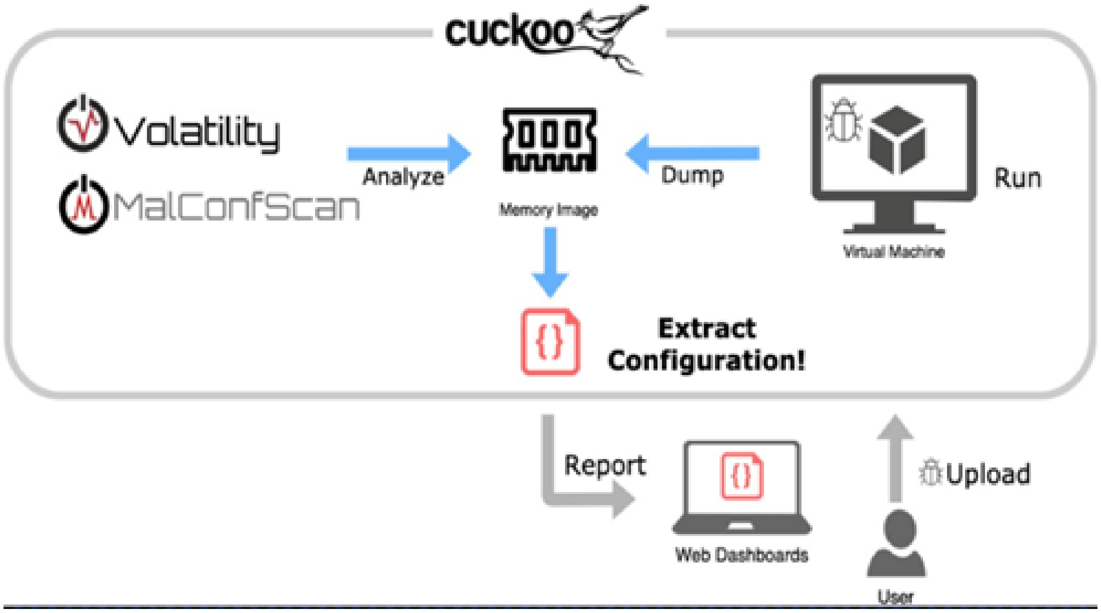


Fig. 18. Non-machine learning detection for ransomware [93].

TABLE XIV. ML AND NON-ML-BASED METHODS

Method	Machine Learning Approach	Non-Machine Learning Approach
Feature Engineering	- Extract relevant and discriminative features from data	- Define heuristics based on known ransomware characteristics
	- Find patterns using deep learning models	- Create rules to detect ransomware behaviors
Anomaly Detection	- Detect deviations from normal system behavior	- Monitor network traffic for unusual patterns
	- Find abnormal file access patterns	- See unusual file encryption behavior
Ensemble Methods	- Combine multiple models to enhance detection performance	- Integrate various detection techniques for comprehensive analysis
	- Reduce false positives through ensemble approaches	- Combine signature-based and heuristic-based detection
Continuous Learning	- Adapt models in real-time with new ransomware samples	- Update signature databases regularly
	- Stay up to date with evolving ransomware variants	- Check for new ransomware families
Dynamic Analysis	- Analyze ransomware behavior in sandboxed environments	- See ransomware actions in isolated systems
	- Find malicious code execution within the sandbox	

VII. RESEARCH DIRECTION

This paper provides a brief overview of machine learning, and deep learning techniques applied to the detection of ransomware. To increase the effectiveness of ransomware detection systems, additional research is required on several open issues.

1) High computational complexity and time: Develop efficient detection systems for new ransomware attacks, considering computational overhead for low-resource devices like embedded systems and IoT.

2) Hardware complexity: Modern systems rely on RAM-intensive hard drives, requiring careful consideration of hardware limitations for sophisticated detection systems and solutions.

3) Evasion and obfuscation: Ransomware detection is dynamic, requiring evasive and secretive methods for accuracy, less false alarms, and dependable handling of escape and confusion.

4) Rich Dataset: Dataset for ransomware attack patterns training machine learning and deep learning models; regular updates needed for effective ransomware detection systems.

VIII. CONCLUSION

This article presents an overview of ransomware detection using heuristic-based machine learning, and non-machine learning technologies. It investigates various ransomware platform detection tools and uses datasets containing different methods. The study provides taxonomy and related concepts for research on new ransomware detection methods, categorizing studies into classical, conventional, and early detection before encryption. It examines the frequency of attack patterns across different platforms and analyzes attacks targeting these platforms. Using heuristic-based machine learning approaches can produce a reliable and precise solution for new ransomware attack patterns. The study aims to encourage academics to use contemporary technologies in ransomware attack detection, evaluating potential solutions and creating more effective models. The main findings and contributions of the reviews shed light on new ransomware detection and pre-encryption strategies. Heuristic-based machine learning should be the focus of future research to identify new ransomware patterns, adjust to changing strategies, and combat evasion methods. In future noise data can be reduced during feature extraction process. Sustained innovation is essential to keep up with the evolution of ransomware.

ACKNOWLEDGMENT

This research is funded by the Fundamental Research Grant Scheme (FRGS) under cost Centre (015MA0-151) from the Ministry of Higher Education (MOHE) Malaysia.

REFERENCES

[1] T. B. Slayton, “Ransomware: The virus attacking the healthcare industry,” J. Leg. Med., vol. 38, no. 2, pp. 287–311, 2018, doi: 10.1080/01947648.2018.1473186.

[2] P. Kuper, “The state of security,” IEEE Secur. Priv., vol. 3, no. 5, pp. 51–53, 2005, doi: 10.1109/MSP.2005.134.

[3] M. Akbanov, V. G. Vassilakis, and M. D. Logothetis, “Ransomware detection and mitigation using software-defined networking: The case of WannaCry,” Comput. Electr. Eng., vol. 76, no. March 2019, pp. 111–121, 2019, doi: 10.1016/j.compeleceng.2019.03.012.

[4] S. Razaulla et al., “The Age of Ransomware: A Survey on the Evolution, Taxonomy, and Research Directions,” IEEE Access, vol. 11, no. April, pp. 40698–40723, 2023, doi: 10.1109/ACCESS.2023.3268535.

[5] ENISA, Threat Landscape Report 2018 - 15 Top Cyberthreats and Trends, no. January. 2018. doi: 10.2824/622757.

[6] N. Scaife, H. Carter, P. Traynor and K. R. B. Butler, "CryptoLock (and Drop It): Stopping Ransomware Attacks on User Data," 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Nara, Japan, 2016, pp. 303-312, doi: 10.1109/ICDCS.2016.46.

[7] M. Hamad and D. Eleyan, “Survey On Ransomware Evolution, Prevention, And Mitigation,” Lume, vol. 10, no. October, p. 2, 2021, [Online]. Available: www.ijstr.org

[8] Bo Li, Kevin Roundy, Chris Gates, and Yevgeniy Vorobeychik. 2017. Large-Scale Identification of Malicious Singleton Files. In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy (CODASPY '17). Association for Computing Machinery, New York, NY, USA, 227–238. https://doi.org/10.1145/3029806.3029815

[9] Heena, “Advances In Malware Detection- An Overview,” 2021, [Online]. Available: http://arxiv.org/abs/2104.01835

[10] S. I. Popoola, U. B. Iyekekpolo, S. O. Ojewande, F. O. Sweetwilliams, S. N. John, and A. A. Atayero, “Ransomware: Current trend, challenges, and research directions,” Lect. Notes Eng. Comput. Sci., vol. 1, pp. 169–174, 2017.

[11] U. Urooj, M. A. B. Maarof and B. A. S. Al-rimy, "A proposed Adaptive Pre-Encryption Crypto-Ransomware Early Detection Model," 2021 3rd International Cyber Resilience Conference (CRC), Langkawi Island, Malaysia, 2021, pp. 1-6, doi: 10.1109/CRC50527.2021.9392548.

[12] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, “Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions,” Comput. Secur., vol. 74, pp. 144–166, May 2018, doi: 10.1016/j.cose.2018.01.001.

[13] Hassan, M. F., Akbar, R., Savita, K. S., Ullah, R., & Mandala, S. (2024). Ransomware Classification with Deep Neural Network and Bi-LSTM. Journal of Advanced Research in Applied Sciences and Engineering Technology, 47(2), 266-280.

[14] A. Moser et al., “Cyber security threats and mitigation techniques for multifunctional devices,” Comput. Secur., vol. 10, no. 1, pp. 1–6, Dec. 2022, doi: 10.1109/ICTAS.2018.8368745.

[15] B. A. S. Al-rimy et al., “Redundancy Coefficient Gradual Up-weighting-based Mutual Information Feature Selection technique for Crypto-ransomware early detection,” Futur. Gener. Comput. Syst., vol. 115, pp. 641–658, 2021, doi: 10.1016/j.future.2020.10.002.

[16] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, “Crypto-ransomware early detection model using novel incremental bagging with enhanced semi-random subspace selection,” Futur. Gener. Comput. Syst., vol. 101, pp. 476–491, Dec. 2019, doi: 10.1016/j.future.2019.06.005.

[17] C. Moore, "Detecting Ransomware with Honeypot Techniques," 2016 Cybersecurity and Cyberforensics Conference (CCC), Amman, Jordan, 2016, pp. 77-81, doi: 10.1109/CCC.2016.14.

[18] S. R. Davies, R. Macfarlane, and W. J. Buchanan, “Differential area analysis for ransomware attack detection within mixed file datasets,” Comput. Secur., vol. 108, p. 102377, 2021, doi: 10.1016/j.cose.2021.102377.

[19] S. H. Kok, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, “Prevention of crypto-ransomware using a pre-encryption detection algorithm,” Computers, vol. 8, no. 4, pp. 1–15, 2019, doi: 10.3390/computers8040079.

[20] O. Aslan and R. Samet, “A Comprehensive Review on Malware Detection Approaches,” IEEE Access, vol. 8, no. March 2021, pp. 6249–6271, 2020, doi: 10.1109/ACCESS.2019.2963724.

[21] M. E. Ahmed, H. Kim, S. Camtepe, and S. Nepal, “Peeler: Profiling Kernel-Level Events to Detect Ransomware,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12972 LNCS, pp. 240–260, 2021, doi: 10.1007/978-3-030-88418-5_12.

[22] V. Madhushalini and L. Raja, “A Novel Ransomware Virus Detection Technique using Machine and Deep Learning Methods,” pp. 8–14, 2023, doi: 10.1109/ICICCS56967.2023.10142938.

[23] K. Begovic, A. Al-ali, and Q. Malluhi, “Cryptographic Ransomware Encryption Detection : Survey,” Comput. Secur., vol. 132, no. February 2022, p. 103349, 2023, doi: 10.1016/j.cose.2023.103349.

[24] V. A. Popescu, G. N. Popescu, and C. R. Popescu, “The relation productivity – Environment in the context of sustainable development–Case study on the Romanian industry,” Metalurgija, vol. 54, no. 1, pp. 286–288, 2015.

[25] G. Cusack, O. Michel, and E. Keller, “Machine learning-based detection of ransomware using SDN,” SDN-NFVSec 2018 - Proc. 2018 ACM Int. Work. Secur. Softw. Defin. Networks Netw. Funct. Virtualization, Co-located with CODASPY 2018, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1145/3180465.3180467.

[26] B. A. S. Al-Rimy et al., “A Pseudo Feedback-Based Annotated TF-IDF Technique for Dynamic Crypto-Ransomware Pre-Encryption Boundary Delineation and Features Extraction,” IEEE Access, vol. 8, pp. 140586–140598, 2020, doi: 10.1109/ACCESS.2020.3012674.

[27] L. J. García Villalba, A. L. Sandoval Orozco, A. López Vivar, E. A. Armas Vega, and T. H. Kim, “Ransomware Automatic Data Acquisition Tool,” IEEE Access, vol. 6, pp. 55043–55051, 2018, doi: 10.1109/ACCESS.2018.2868885.

[28] R. Moussaileb, N. Cuppens, J. L. Lanet, and H. Le Bouder, “A Survey on Windows-based Ransomware Taxonomy and Detection Mechanisms: Case Closed?,” ACM Comput. Surv., vol. 54, no. 6, 2021, doi: 10.1145/3453153.

[29] E. Berrueta, D. Morato, E. Magana, and M. Izal, “A Survey on Detection Techniques for Cryptographic Ransomware,” IEEE Access, vol. 7, pp. 144925–144944, 2019, doi: 10.1109/ACCESS.2019.2945839.

[30] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, “A 0-day aware crypto-ransomware early behavioral detection framework,” Lect. Notes Data Eng. Commun. Technol., vol. 5, pp. 758–766, 2018, doi: 10.1007/978-3-319-59427-9_78.

[31] Monika, P. Zavarsky, and D. Lindskog, “Experimental Analysis of Ransomware on Windows and Android Platforms: Evolution and Characterization,” Procedia Comput. Sci., vol. 94, pp. 465–472, 2016, doi: 10.1016/j.procs.2016.08.072.

[32] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, “Cybersecurity data science: an overview from machine learning perspective,” J. Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00318-5.

[33] B. A. S. Al-Rimy, M. A. Maarof, Y. A. Prasetyo, S. Z. M. Shaid, and A. F. M. Ariffin, “Zero-day aware decision fusion-based model for crypto-ransomware early detection,” Int. J. Integr. Eng., vol. 10, no. 6, pp. 82–88, 2018, doi: 10.30880/ijie.2018.10.06.011.

[34] A. I. M. Detection et al., “A proposed Adaptive Pre-Encryption Crypto-Ransomware Early Detection Model,” IEEE Access, vol. 10, no. 1, pp. 3–8, Jan. 2023, doi: 10.1109/CRC50527.2021.9392548.

[35] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.,” Ann. Intern. Med., vol. 151, no. 4, pp. 264–9, W64, Aug. 2009, doi: 10.7326/0003-4819-151-4-200908180-00135.

[36] K. Adnan, R. Akbar, and K. S. Wang, “Development of Usability Enhancement Model for Unstructured Big Data Using SLR,” IEEE Access, vol. 9, pp. 87391–87409, 2021, doi: 10.1109/ACCESS.2021.3089100.

[37] N. Ariffin, A. Zainal, M. A. Maarof and M. Nizam Kassim, "A Conceptual Scheme for Ransomware Background Knowledge Construction," 2018 Cyber Resilience Conference (CRC), Putrajaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/CR.2018.8626868.

[38] B. Celiktas and E. Karacuha, “Ransomware , Detection and Prevention Techniques , Cyber Security , Malware Analysis Istanbul Technical University Informatics Institute Using Signature and Anomaly Based Detection Methods Barış Çeliktaş Department of Applied Informatics Applied Informa,” no. July, 2018.

[39] S. Alsoghyer and I. Almomani, “Ransomware detection system for android applications,” Electron., vol. 8, no. 8, pp. 1–36, 2019, doi: 10.3390/electronics8080868.

[40] E. G. Dada, J. Stephen Bassi, Y. J. Hurcha, and A. H. Alkali, “Performance Evaluation of Machine Learning Algorithms for Detection and Prevention of Malware Attacks Related papers Performance Evaluation of Machine Learning Algorithms for Detection and Prevention of Malware Attacks,” vol. 21, no. 3, pp. 18–27, 2019, doi: 10.9790/0661-2103011827.

[41] H. Aghakhani et al., “When Malware is Packin’ Heat; Limits of Machine Learning Classifiers Based on Static Analysis Features,” no. February, 2020, doi: 10.14722/ndss.2020.24310.

[42] D. W. Fernando, N. Komninos, and T. Chen, “A Study on the Evolution of Ransomware Detection Using Machine Learning and Deep Learning Techniques,” IoT, vol. 1, no. 2, pp. 551–604, Dec. 2020, doi: 10.3390/iot1020030.

[43] M. Almousa, S. Basavaraju, and M. Anwar, “API-Based Ransomware Detection Using Machine Learning-Based Threat Detection Models,” in 2021 18th International Conference on Privacy, Security and Trust, PST 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/PST52912.2021.9647816. 13-15 December 2021 at Auckland, New Zealand

[44] M. Rhode, P. Burnap, and A. Wedgbury, “Real-Time Malware Process Detection and Automated Process Killing,” Secur. Commun. Networks, vol. 2021, 2021, doi: 10.1155/2021/8933681.

[45] M. Almousa, J. Osawere and M. Anwar, "Identification of Ransomware families by Analyzing Network Traffic Using Machine Learning Techniques," 2021 Third International Conference on Transdisciplinary AI (TransAI), Laguna Hills, CA, USA, 2021, pp. 19-24, doi: 10.1109/TransAI51903.2021.00012.

[46] Y. A. Ahmed et al., “A Weighted Minimum Redundancy Maximum Relevance Technique for Ransomware Early Detection in Industrial IoT,” Sustain., vol. 14, no. 3, pp. 1–16, 2022, doi: 10.3390/su14031231.

[47] A. Singh, R. Ikuesan, and H. Venter, “Ransomware Detection using Process Memory,”. 17th International Conference on Cyber Warfare and Security (ICCWS 2022), hosted State University of New York at Albany, USA on 17-18 March 2022. doi: 10.34190/iccws.17.1.53

[48] R. M. A. Molina, S. Torabi, K. Sarieddine, E. Bou-Harb, N. Bouguila, and C. Assi, “On Ransomware Family Attribution Using Pre-Attack Paranoia Activities,” IEEE Trans. Netw. Serv. Manag., vol. 19, no. 1, pp. 19–36, 2022, doi: 10.1109/TNSM.2021.3112056.

[49] S. Mehnaz, A. Mudgerikar, and E. Bertino, RWGuard: A real-time detection system against cryptographic ransomware, vol. 11050 LNCS,

no. March 2019. Springer International Publishing, 2018. doi: 10.1007/978-3-030-00470-5_6.

[50] H. Zhang, X. Xiao, F. Mercaldo, S. Ni, F. Martinelli, and A. K. Sangaiah, “Classification of ransomware families with machine learning based on N-gram of opcodes,” Futur. Gener. Comput. Syst., vol. 90, pp. 211–221, 2019, doi: 10.1016/j.future.2018.07.052.

[51] E. Berrueta, D. Morato, E. Magaña, and M. Izal, “Crypto-ransomware detection using machine learning models in file-sharing network scenarios with encrypted traffic,” Expert Syst. Appl., vol. 209, Dec. 2022, doi: 10.1016/j.eswa.2022.118299.

[52] A. Azmoodeh, A. Dehghantanha, M. Conti, and K.-K. K. R. Choo, “Detecting crypto-ransomware in IoT networks based on energy consumption footprint,” J. Ambient Intell. Humaniz. Comput., vol. 9, no. 4, pp. 1141–1152, 2018, doi: 10.1007/s12652-017-0558-5.

[53] R. Agrawal, J. W. Stokes, K. Selvaraj, and M. Marinescu, “University of California , Santa Cruz , Santa Cruz , CA 95064 USA Microsoft Corp ., One Microsoft Way , Redmond , WA 98052 USA,” pp. 3222–3226, 2019.

[54] S. Poudyal and Di. Dasgupta, “Analysis of Crypto-Ransomware Using ML-Based Multi-Level Profiling,” IEEE Access, vol. 9, pp. 122532–122547, 2021, doi: 10.1109/ACCESS.2021.3109260.

[55] B. M. Khammas, “Ransomware Detection using Random Forest Technique,” ICT Express, vol. 6, no. 4, pp. 325–331, Dec. 2020, doi: 10.1016/j.icte.2020.11.001.

[56] R. Moussaileb, N. Cuppens, J. L. Lanet, and H. Le Bouder, “Ransomware Network Traffic Analysis for Pre-encryption Alert,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12056 LNCS, pp. 20–38, 2020, doi: 10.1007/978-3-030-45371-8_2.

[57] A. Singh, A. Ikuesan, and H. Venter, “A context-aware trigger mechanism for ransomware forensics,” 14th Int. Conf. Cyber Warf. Secur. ICCWS 2019, Feb 28- Mar 1, 2019 at Stellenbosch, South Africa pp. 629–638, 2019.

[58] K. Cabaj, M. Gregorczyk, and W. Mazurczyk, “Software-defined networking-based crypto ransomware detection using HTTP traffic characteristics,” Comput. Electr. Eng., vol. 66, no. November 2016, pp. 353–368, 2018, doi: 10.1016/j.compeleceng.2017.10.012.

[59] D. Morato, E. Berrueta, E. Magaña, and M. Izal, “Ransomware early detection by the analysis of file sharing traffic,” J. Netw. Comput. Appl., vol. 124, no. September, pp. 14–32, 2018, doi: 10.1016/j.jnca.2018.09.013.

[60] S. Wang et al., “KRProtector: Detection and Files Protection for IoT Devices on Android Without ROOT Against Ransomware Based on Decoys,” IEEE Internet Things J., vol. 9, no. 19, pp. 18251–18266, 2022, doi: 10.1109/JIOT.2022.3156571.

[61] K. P. Subedi, D. R. Budhathoki, and D. Dasgupta, “Forensic analysis of ransomware families using static and dynamic analysis,” Proc. - 2018 IEEE Symp. Secur. Priv. Work. SPW 2018, pp. 180–185, 2018, doi: 10.1109/SPW.2018.00033.

[62] S. Sharmeen, Y. A. Ahmed, S. Huda, B. S. Kocer, and M. M. Hassan, “Avoiding Future Digital Extortion through Robust Protection against Ransomware Threats Using Deep Learning Based Adaptive Approaches,” IEEE Access, vol. 8, pp. 24522–24534, 2020, doi: 10.1109/ACCESS.2020.2970466.

[63] S. Maniath, A. Ashok, P. Poornachandran, V. G. Sujadevi, A. U. P. Sankar, and S. Jan, “Deep learning LSTM based ransomware detection,” 2017 Recent Dev. Control. Autom. Power Eng. RDCAPE 2017, vol. 3, pp. 442–446, 2018, doi: 10.1109/RDCAPE.2017.8358312.

[64] V. R., M. Alazab, A. Jolfaei, S. K.P. and P. Poornachandran, "Ransomware Triage Using Deep Learning: Twitter as a Case Study," 2019 Cybersecurity and Cyberforensics Conference (CCC), Melbourne, VIC, Australia, 2019, pp. 67-73, doi: 10.1109/CCC.2019.000-7.,

[65] S. Homayoun et al., “DRTHIS: Deep ransomware threat hunting and intelligence system at the fog layer,” Futur. Gener. Comput. Syst., vol. 90, pp. 94–104, 2019, doi: 10.1016/j.future.2018.07.045.

[66] S. Usharani, P. M. Bala, and M. M. J. Mary, “Dynamic analysis on crypto-ransomware by using machine learning: Gandcrab ransomware,” J. Phys. Conf. Ser., vol. 1717, no. 1, 2021, doi: 10.1088/1742-6596/1717/1/012024.

[67] S. Song, B. Kim, and S. Lee, “The Effective Ransomware Prevention Technique Using Process Monitoring on Android Platform,” Mob. Inf. Syst., vol. 2016, 2016, doi: 10.1155/2016/2946735.

[68] J. K. Lee, S. Y. Moon, and J. H. Park, “CloudRPS: a cloud analysis based enhanced ransomware prevention system,” J. Supercomput., vol. 73, no. 7, pp. 3065–3084, 2017, doi: 10.1007/s11227-016-1825-5.

[69] A. Wani and S. Revathi, “Ransomware protection in loT using software defined networking,” Int. J. Electr. Comput. Eng., vol. 10, no. 3, pp. 3166–3174, 2020, doi: 10.11591/ijece.v10i3.pp3166-3175.

[70] S. K. Shaukat and V. J. Ribeiro, "RansomWall: A layered defense system against cryptographic ransomware attacks using machine learning," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 2018, pp. 356-363, doi: 10.1109/COMSNETS.2018.8328219.

[71] A. Kharraz and E. Kirda, “Redemption: Real-Time Protection Against Ransomware at End-Hosts BT - Recent Advances in Intrusion Detection,” Recent Adv. Intrusion Detect., vol. 10453, no. Chapter 5, pp. 98–119, 2017, [Online]. Available: http://link.springer.com/10.1007/978-3-319-66332-6_5%0Apapers3://publication/doi/10.1007/978-3-319-66332-6_5

[72] J. Zhang et al., "Scarecrow: Deactivating Evasive Malware via Its Own Evasive Logic," 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Valencia, Spain, 2020, pp. 76-87, doi: 10.1109/DSN48063.2020.00027.

[73] S. Lee, H. K. Kim, and K. Kim, “Ransomware protection using the moving target defense perspective,” Comput. Electr. Eng., vol. 78, pp. 288–299, 2019, doi: 10.1016/j.compeleceng.2019.07.014.

[74] K. Cabaj and W. Mazurczyk, “Using software-defined networking for ransomware mitigation: The case of cryptowall,” IEEE Netw., vol. 30, no. 6, pp. 14–20, 2016, doi: 10.1109/MNET.2016.1600110NM.

[75] Eugene Kolodenker, William Koch, Gianluca Stringhini, and Manuel Egele. 2017. PayBreak: Defense Against Cryptographic Ransomware. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17). Association for Computing Machinery, New York, NY, USA, 599–611. https://doi.org/10.1145/3052973.3053035

[76] J. S. Aidan, Zeenia and U. Garg, "Advanced Petya Ransomware and Mitigation Strategies," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 23-28, doi: 10.1109/ICSCCC.2018.8703323.

[77] E. Rouka, C. Birkinshaw and V. G. Vassilakis, "SDN-based Malware Detection and Mitigation: The Case of ExPetr Ransomware," 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2020, pp. 150-155, doi: 10.1109/ICIoT48696.2020.9089514.

[78] Marco Antonio Sotelo Monge, Jorge Maestre Vidal, and Luis Javier García Villalba. 2018. A novel Self-Organizing Network solution towards Crypto-ransomware Mitigation. In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES '18). Association for Computing Machinery, New York, NY, USA, Article 48, 1–10. https://doi.org/10.1145/3230833.3233249

[79] S. R. Davies, R. Macfarlane, and W. J. Buchanan, “Evaluation of live forensic techniques in ransomware attack mitigation,” Forensic Sci. Int. Digit. Investig., vol. 33, 2020, doi: 10.1016/j.fsidi.2020.300979.

[80] R. Umar, I. Riadi, and R. S. Kusuma, “Mitigating sodinokibi ransomware attack on cloud network using software-defined networking (SDN),” Int. J. Saf. Secur. Eng., vol. 11, no. 3, pp. 239–246, 2021, doi: 10.18280/ijsse.110304.

[81] V. Mathane and P. V. Lakshmi, “Predictive Analysis of Ransomware Attacks using Context-aware AI in IoT Systems,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 4, pp. 240–244, 2021, doi: 10.14569/IJACSA.2021.0120432.

[82] C. G. Akcora, Y. Li, Y. R. Gel, and M. Kantarcioglu, “Bitcoin heist: Topological data analysis for ransomware prediction on the bitcoin blockchain,” 17th Pacific Rim International Conference on Artificial Intelligence!IJCAI-PRICAI2020 7-15-2021 Yokohama, Japan,

[83] M. Rhode, P. Burnap, and K. Jones, “Early-stage malware prediction using recurrent neural networks,” Comput. Secur., vol. 77, no. December 2017, pp. 578–594, 2018, doi: 10.1016/j.cose.2018.05.010.

[84] Shengyun Xu. 2021. The Application of Machine Learning in Bitcoin Ransomware Family Prediction. In Proceedings of the 2021 5th International Conference on Information System and Data Mining (ICISDM '21). Association for Computing Machinery, New York, NY, USA, 21–27. https://doi.org/10.1145/3471287.3471300

[85] H. Y. Chang, T. L. Lin, T. F. Hsu, Y. S. Shen, and G. R. Li, "Implementation of ransomware prediction system based on weighted-KNN and real-time isolation architecture on SDN Networks," 2019 IEEE Int. Conf. Consum. Electron. - Taiwan, ICCE-TW 2019, pp. 4–5, 2019, doi: 10.1109/ICCE-TW46550.2019.8991771.

[86] U. Adamu and I. Awan, "Ransomware Prediction Using Supervised Learning Algorithms," 2019 7th International Conference on Future Internet of Things and Cloud (FiCloud), Istanbul, Turkey, 2019, pp. 57-63, doi: 10.1109/FiCloud.2019.00016.

[87] W. Song et al., "Crypto-ransomware Detection through Quantitative API-based Behavioral Profiling," 2023, [Online]. Available: http://arxiv.org/abs/2306.02270

[88] J. Modi and B. Eng, "Detecting Ransomware in Encrypted Network Traffic Using Machine Learning," 2019, [Online]. Available: https://dspace.library.uvic.ca/handle/1828/11076

[89] R. Brewer, "Ransomware attacks: detection, prevention and cure," Netw. Secur., vol. 2016, no. 9, pp. 5–9, 2016, doi: 10.1016/S1353-4858(16)30086-1.

[90] A. Djenna, A. Bouridane, S. Rubab, and I. M. Marou, "Artificial Intelligence-Based Malware Detection, Analysis, and Mitigation," Symmetry (Basel)., vol. 15, no. 3, pp. 1–24, 2023, doi: 10.3390/sym15030677.

[91] H. S. Talabani and H. M. T. Abdulhadi, "Bitcoin Ransomware Detection Employing Rule-Based Algorithms," Sci. J. Univ. Zakho, vol. 10, no. 1, pp. 5–10, 2022, doi: 10.25271/sjuoz.2022.10.1.865.

[92] S. H. Kok, A. Azween, and N. Z. Jhanjhi, "Evaluation metric for crypto-ransomware detection using machine learning," J. Inf. Secur. Appl., vol. 55, p. 102646, 2020, doi: 10.1016/j.jisa.2020.102646.

[93] R. S. Rahul, "Malware analysis detection," Iconic Res. Eng. Journals, vol. 1, no. 10, pp. 132–135, 2018, [Online]. Available: http://irejournals.com/formatedpaper/1700619.pd

[94] D. Sgandurra, L. Muñoz-González, R. Mohsen, and E. C. Lupu, "Automated Dynamic Analysis of Ransomware: Benefits, Limitations and use for Detection," no. September, 2016, [Online]. Available: http://arxiv.org/abs/1609.03020

[95] Rehman, M. U., Akbar, R., Omar, M., & Gilal, A. R. (2023, September). A Systematic Literature Review of Ransomware Detection Methods and Tools for Mitigating Potential Attacks. In International Conference on Computing and Informatics (pp. 80-95). Singapore: Springer Nature Singapore.

[96] Shaikh, M. R., Ullah, R., Akbar, R., Savita, K. S., & Mandala, S. (2024). Fortifying Against Ransomware: Navigating Cybersecurity Risk Management with a Focus on Ransomware Insurance Strategies. Int. J. Acad. Res. Bus. Soc. Sci, 14(1), 1415-1430.

[97] Sathio, A. A., Dootio, M. A., Lakhan, A., ur Rehman, M., Pnhwar, A. O., & Sahito, M. A. (2021, August). Pervasive futuristic healthcare and blockchain enabled digital identities-challenges and future intensions. In 2021 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 30-35). IEEE.

[98] Yalli, J. S., Hasan, M. H., Haron, N. S., Rehman Shaikh, M. U., Murad, N. Y., & Bako, A. L. (2023). Quality of Data (QoD) in Internet of Things (IOT): An Overview, State-of-the-Art, Taxonomy and Future Directions. International Journal of Advanced Computer Science & Applications, 14(12).

[99] Mujeeb-ur-Rehman, A. L., Hussain, Z., Khoso, F. H., & Arain, A. A. (2021). Cyber security intelligence and ethereum blockchain technology for e-commerce. International Journal, 9(7).

[100] Ur Rehman, M., Akbar, R., Mujeeb, S., & Janisar, A. A. Deep-learning enabled early detection of COVID-19 infection in IoMT fog-cloud healthcare in LPWAN. In Low-Power Wide Area Network for Large Scale Internet of Things (pp. 235-258). CRC Press

[101] Panhwar, A. O., Sathio, A. A., Lakhan, A., Umer, M., Mithiani, R. M., & Khan, S. (2022). Plant health detection enabled CNN scheme in IoT network. International Journal of Computing and Digital Systems, 11(1), 344-335.

[102] Sahito, M. A., & Kehar, A. (2021). Dynamic content enabled microservice for business applications in distributed cloudlet cloud network. International Journal, 9(7), 1035-1039

[103] Chandio, S. A., & Mahar, J. A. (2017, April). Gadget improved security alert monitoring, management and mitigation system to control the crowded occasions. In 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT) (pp. 1-3). IEEE.

Texture Feature and Mel-Spectrogram Analysis for Music Sound Classification

M. E. ElAlami[1], S. M. K. Tobar[2], S. M. Khater[3], Eman. A. Esmaeil[4]
Computer Science Department-Faculty of Specific Education, Mansoura University, Mansoura, Egypt[1, 3, 4]
Musical Education Department-Faculty of Specific Education, Mansoura University, Mansoura, Egypt[2]

***Abstract*—The categorization of music has received substantial interest in the management of large-scale databases. However, the sound of music classification (MC) is poorly interesting, making it a big challenge. For this reason, this paper has proposed a new robust combining method based on texture feature with Mel-spectrogram to classify Arabic music sound. A music audio dataset consisting of 404 sound recordings for different four classes of Arabic music sounds has been collected. The collected data became available for free on the Kaggle website. Firstly, music sound is transformed into a Mel spectrogram, and then several texture features are extracted from these Mel spectrogram images. A two-dimensional Haar wavelet is applied to each Mel-spectrogram image, and Local Binary Patterns (LBP), Gray Level Co-occurrence Matrix (GLCM), and Histogram of Oriented Gradient (HOG) are utilized for feature extraction. K-nearest neighbors (KNN), random forest (RF), decision tree (DT), logistic regression (LR), AdaBoost, extreme gradient boosting (XGB), and support vector machine (SVM) classifiers were utilized in a comparative analysis of Machine Learning (ML) algorithms. Two different datasets have been employed in order to evaluate the effectiveness of our approach: the collected dataset that the authors had gathered and the global GTZAN dataset. Our method demonstrates superior performance with a five-fold cross-validation. The experimental findings indicated that the XGB exhibited a high accuracy with an average performance of 97.80% for accuracy, 97.72% for F1-Score, 97.75% for recall, and 97.81% for precision.**

***Keywords*—Mel-spectrograms; ML; texture features; MC**

I. Introduction

Music is considered an inseparable part of our culture and tradition [1]. The advent of online social networks and cloud technology have led to a massive rise in the demand for online data storage and data sharing services [2, 3]. Music Information Retrieval (MIR) systems have gained huge popularity in recent years and are used in many fields, such as musical similarity and genre categorization, music emotion identification, music source separation, acoustic descriptions of music, and music transcription [4]. Music classification (MC) has emerged as an important area for digital music services such as Tidal, SoundCloud, and Apple Music and is used for classifying and overseeing extensive musical datasets [5, 6]. Regarding this, musical sound classification is an intriguing area challenge in the field [7]. Among the several methods for representing the contents of an audio clip, extracting distinguishing features is the most widely employed. However, due to the subjectivity associated with the concept of musical genre, as well as the enormous variety of music genres, strong feature extraction has proven difficult. In the Arab world, Arabic music is an essential component of global music, but Western music dominates the field. A machine learning approach is extensively used in music information retrieval applications [7, 8, 9, 10]. As well, texture features have a high capacity for extracting features of musical patterns [11, 12].

In related works, Western music using the GTZAN dataset dominates the field, and Arabic music is not yet equivalent to them. So, Arabic musical instruments must be moved out of the country and promoted for it in the works. Therefore, we hope that this work will contribute to solving this problem and overcoming the absence of a dataset based on Arabic music. Therefore, this paper presents a new robust approach based on ML techniques with texture features and Mel-spectrogram for Arabic music sound classification using a newly collected dataset in favour of this work and became available free for use.

Our contribution can be summarized as follows:

- A music audio dataset consisting of 440 sound recordings for different four classes of Arabic music sounds has been collected. The collected data became available for free at: (https://www.kaggle.com/datasets/emanatyaesmaeil/zekrayati-dataset).
- Gathered and annotated a large corpus of Arabic music clips to cover the lack of a dataset for Arabic music. Although the GTZAN dataset is a benchmark for MGC, it has limitations such as mislabeling, distortions, and replicas (Strum, [13]).
- A new robust feature extraction approach is presented for music signal classification using Mel-spectrogram images. A two-dimensional Haar wavelet is applied to each image and texture features (GLCM, HOG, and LBP) are extracted from all wavelet transform sub-bands.
- Comparative analysis to examine the efficiency of most various machine learning algorithms in Arabic music sound classification and then determine which algorithm is better for this type of data.
- The best accuracy had resulted compared to previous studies using global GTZAN dataset.

The general structure of this paper is as follows: Section II presents the related studies, Section III covers the materials and methodologies, and Section IV provides the results and

*Corresponding Author.

discussion. Discussion is given in Section V. Finally, Section VI covers conclusion and future work.

II. Related Work

Recently, there have been a lot of studies related to music classification. [14]. These works have been supported by recent advancements in machine learning (ML) and deep learning (DL) methodologies. This section provides a thorough overview of many ML and DL applications in the music industry and examination of the prospects for AI in this domain as show in Table I.

In summary, western music dominates the field. Most of the literature focuses on it by using the GTZAN dataset. Arabic music needs to move out of the country, so it is hoped that this paper will make a small contribution to this goal and benefit from the ML approach that has proven its effectiveness in music classification.

TABLE I. PREVIOUS STUDIES RELATED TO ML AND DL APPROACH

REF.	Year	Task	Algorithm	Dataset	Accuracy
[15]	**2022**	Dissecting the Nigerian music genre.	Timbral texture feature using SVM, XGB, RF, and K-NN	ORIN dataset	XGB=0.82, SVM= 0.74, RF=0.71 and K-NN=0.51
[16]	**2024**	Classification of Musical Genres	14 audio features in total when using XGB	GTZAN Dataset	Accuracy=81%
[17]	**2023**	Classification of musical genres	CNN-based mel-frequency cepstral coefficients (MFCC)	GTZAN	Accuracy=85%
[18]	**2024**	Classification of Music Categories	MFCC + STFT + CNN	GTZAN and Extended-Ballroom datasets	Accuracy of dataset1=95.71 Accuracy of dataset2=95.20
[19]	**2023**	categorization of music genres	hybrid model for wavelet + spectrogram analysis	Ballroom and GTZAN datasets	Accuracy of dataset1=81% Accuracy of dataset2=71%
[20]	**2021**	categorization of music genres	MEL-Spectrogram based on logs and Transfer Learning	GTZAN dataset	The best accuracy is Resnet34=97%
[21]	**2023**	Automated Genre Classification of Music	MFCC and CNN	GTZAN dataset	Accuracy=83%
[22]	**2022**	Identification of musical genres	CNN with Mel-spectrograms: The Best Feature	GTZAN dataset	Accuracy=91%
[23]	**2023**	Suggested Music Track	DCNN and Mel-spectrograms	JUNO, GTZAN, and FMA-Small datasets	Dataset1 Accuracy=63%, Dataset2 Accuracy=78% and Dataset3 Accuracy=89%
[24]	**2023**	Categorization of Music Genres	Ideal model with CRNN and Mel-spectrograms	FMA-Small dataset	Accuracy=90%
[25]	**2024**	Indian Category of Musical Instruments	K-NN, RF, RNN, XGB, LR, DT, and SVM in an MFCC	Gathered 1177 audio samples in total with six classes from different online sources.	RNN has best accuracy=0.9872
[26]	**2020**	Categorization of Music Genres	feature extraction from metadata using SVM, K-NN, and NB	Spotify music dataset	SVM=80%, K-NN=77.18% and NB=76.08%
[27]	**2020**	Identification of Music Genres	CNN and the Mel Spectrum	GTZAN dataset	Accuracy=84%
[28]	**2021**	categorization of music	Spectrograms with many DNN models; the best model is ResNet50.	The datasets FMA, GTZAN_4, and EMA	Dataset1 Accuracy=80.14%, Dataset2 Accuracy=81.09% and Dataset3 Accuracy =77.03%
[29]	**2022**	Identification of Music Genres	CNN, LSTM, and MLP in an MFCC	GTZAN dataset	CNN = 70.42%; LSTM = 61.50%; and MLP = 63.28%
[30]	**2020**	Identification of Music Genres	combined (FC1, FC2, FC3, FC4) with SVM	Spotify Music Dataset	Accuracy of FC1 and FC2=80%
[31]	**2022**	Bangla music's classification by genre	Feature Scaling Method plus PCA utilizing NN, RF, K-NN, and SVM	Bangla Music Dataset	SVM-RBF=68.77%, K-NN=61.32%, RF=69.05% and NN=77.68%

III. Materials and Methods

The flow diagram of the suggested model is illustrated in Fig. 1. The following three subsections will describe each of these stages in detail.

A. *Mel-Spectrogram Production Stage*

The Mel-spectrogram is resulted through the following steps:

1) Pre-emphasizing audio which improves clarity and reduces volume.

2) Blocking frames that render every audio frame end-to-end, maintaining audio continuity.

3) Introducing a window function to enhance the role of audio framing and prevent audio discontinuity caused by sampling and quantization.

4) Fast Fourier Transform which transforms the audio from the time domain to the frequency domain.

5) Map the FFT produce to the Mel scale; multiply it by the total number of triangular bandpass filters.

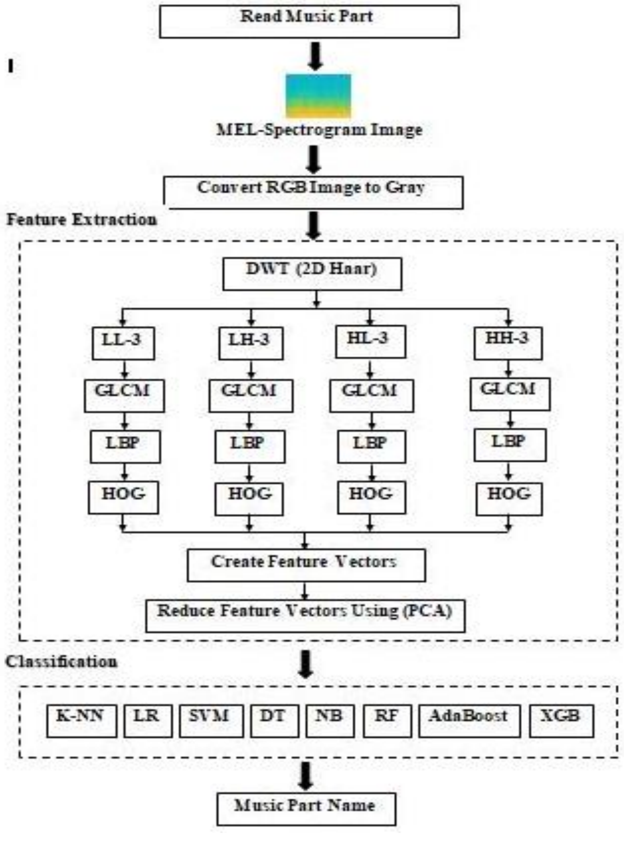


Fig. 1. The proposed system for lute signal classification.

Mel-spectrograms in this study are a two-dimensional representation of input signals. All audio signals are produced using the Short Time Fourier Transform (STFT) with Mel-frequency rather than normal frequency. The parameters used to generate the power Mel-spectrograms are stated in Table II, and Fig. 2 depicts various Mel-spectrograms for the dataset and shows the region of the image used for feature extraction as the black box, and the output is saved as a .png file with a size of 256*256. The output photos are transformed to grayscale before the textural features are extracted.

Parameter	Value
Audio Length (second)	12:91
Window Length (frames)	1024
Overlap Length (frames)	512
FFT Length (frames)	4096
Number bands (Filters)	64

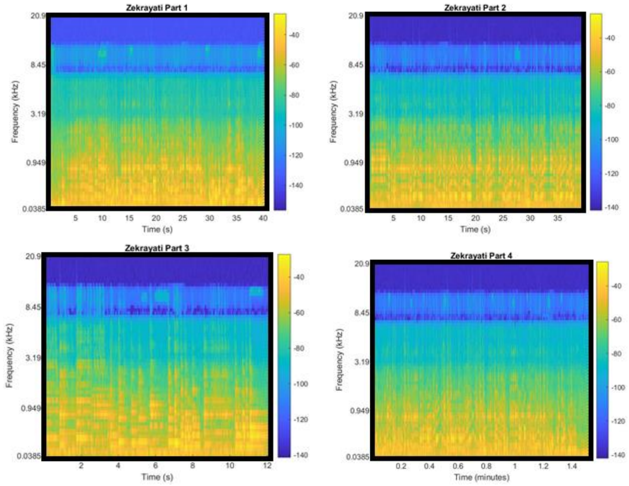


Fig. 2. Mel-spectrogram for some datasets.

B. *Feature Extraction Stage*

The primary phases for feature extraction are as follows:

Step 1: apply the discrete wavelet transforms to the Mel-spectrogram image to extract sub-bands.

Step 2: extracted all of (GLCM, HOG, and LBP) from all wavelet sub-bands and combine all features into one feature vector.

Step 3: reduce the final feature vector by using Principal component analysis (PCA).

1) Discrete Wavelet Transforms (DWT): It is a powerful image signal analysis tool. It has an effective analysis function and multi-resolution analysis capability, making it suited for the image signal analysis area [32]. The spatial domain (DWT or 2D_DWT) is resulted by first applying the output to the DWT along the vertical axis and then applying the horizontal axis to the (1D_DWT). Hence, (2D-DWT) contains four bands: (LL, LH, HL, and HH) bands [33, 34].

Eq. (1) symbolizes the transformation (DWT) of any signal, x(t).

$$x(t) = \sum a_{j.k} \Psi_{j.k}(t) \tag{1}$$

Where $a_{j.k}$ are called wavelet coefficients, $\Psi_{j.k}(t)$ is called the fundamental function. j is the scale and k is mother wavelet translated $\Psi(t)$.

The (2_D DWT) can be achieved by using Eq. (2) to apply DWT across rows and columns of a picture in both the (x and y) dimensions.

$$f(x.y) = \sum_{j.k} C_{j_0}(k.l)\varphi_{j.k.l}(x.y) + \sum_{s=H.V.D} \sum_{j=j_0}^{\infty} D_j^s[k.l]\Psi_{j.k.l}^s(x.y) \quad (2)$$

Where C_{j_0} is the approximation coefficient, $\varphi_{j.k.l}(x.y)$ is scaling function, D_j^s is set of detailed coefficients and $\Psi_{j.k.l}^s$ is wavelet function.

In this work, three level wavelet decomposition has been performed by using 'haar' mother wavelet function as shown in Fig. 3 and the sub bands of level three were used for extracting the features.

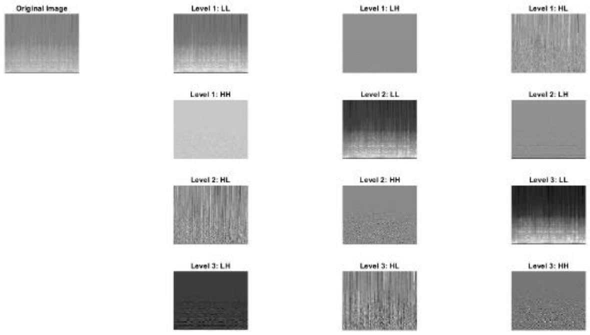


Fig. 3. Three-level wavelet decomposition.

2) GLCM Algorithm: The GLCM texture extraction approach has become more and more popular in recent years for picture classification and detection [35, 36]. It alludes to a widely used technique for characterizing texture through an examination of grayscale's spatial correlation features. It determines the frequency of occurrence for each piece of grayscale data it contains. The GLCM is a (L × L) counting matrix, where each element in the GLCM represents a potential combination of pixels, assuming the original image has (L) grayscale levels. Several studies have demonstrated that this approach is highly adaptable and stable in capturing detailed information such as direction, distance, and variation range between image pixel grayscales. The contrasts and patterns of texture features acquired using this method accurately characterize the properties of picture texture [37]. Some GLCM features [38] used in this work are briefly explained below.

- Contrast: This feature calculates the intensity of a pixel and its surrounding pixels over the entire image. The contrast function also calculates the color and brightness differences between each cellular object and other objects in the same field of view. It can be computed using the following equation.

$$Contrast = \sum_i \sum_j (i-j)^2 p(i.j) \quad (3)$$

For an image, $p(i.j)$ reflects the chance of a pair of pixels with gray level values(i and j)occurring in a specific space and direction.

- Correlation: It computes gray-level linear dependence among pixels at specified distances from one another. The(μ_i and μ_j)are the average of each row and column, and (σ_i and $\sigma_{j)}$ are, correspondingly, the standard deviations for each row and column.

$$Correlation = \sum_i \sum_j \frac{p(i.j)[(i-\mu_i)(j-\mu_j)]}{\sigma_i \sigma_j} \quad (4)$$

- Energy: It calculates regularity or pixel pair repetitions, as illustrated in the equation below. When a pixel pair is repeated multiple times, the energy characteristic returns a higher value.

$$Energy = \sum_i \sum_j p(i.j)^2 \quad (5)$$

- Homogeneity: it is refers to the consistency of element distribution along a GLCM's diagonal. When matrix elements are spread diagonally, homogeneity is high, as calculated by the equation below.

$$Homogeneity = \sum_i \sum_j \frac{p(i.j)}{1+|i-j|} \quad (6)$$

In this study, Mel-spectrogram image characteristics are extracted using the GLCM approach, first set the order of the grayscale co-generation matrix to 16 and selected 0°, 45°, 90° and 135° as the four directions of the grayscale co-generation matrix, The final eigenvalue co-generation matrix was calculated by averaging the four directional matrix eigenvalues. Finally, we retrieved four-dimensional GLCM features for each subband, yielding a final feature vector of 16 features.

3) LBP Algorithm: Local Binary Pattern is a well-known texture descriptor that has been successfully used in works made for several application domains, such as music genre detection [39, 40]. According to [41], it is uses the local neighborhood of a center pixel to find a local binary pattern. The feature vector, which characterizes the image's textural richness, corresponds to the histogram of local binary patterns present in all pixels. There are two basic parameters that can be adjusted to extract the LBP from an image. The first is the number of nearby pixels that will be considered for the central pixel, while the second is the distance between the central pixel and its neighbors. These values are referred to, in turn, as (P and R). Fig. 4 shows examples of Mel-spectrogram images, corresponding maps of LBP values, and the LBP histograms.

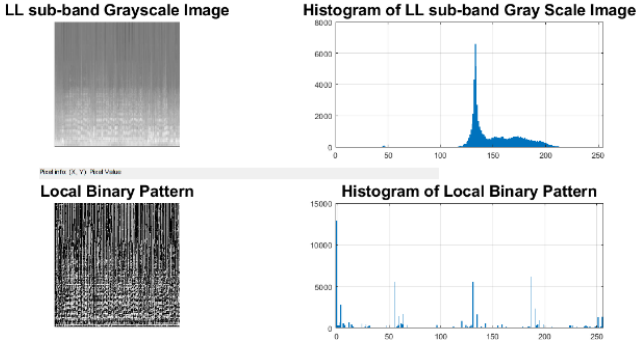


Fig. 4. Local binary patterns visualization.

In this study, 8 neighbors at a distance of 2 was used to extracted 59 features for each sub-bands and the final feature vector for this step was 236 feature value.

4) HOG Algorithm: HOG is a feature descriptor used to detect targets in image processing. It creates features by calculating the histogram of directional gradients in discrete parts of the image [42]. This method consists of two major processes [43]. The first is the histogram extraction, as the gradient of direction and magnitude is retrieved from each pixel in the input image. These steps are used to generate an angular histogram of gradients, which is then applied as an image texture feature vector. The vertical and horizontal components of the image I (i, j) are derivatives of pixel (i, j). They're computed as follows:

$$G_i(i.j) = I(i+1.j) - I(i-1.j) \qquad (7)$$

$$G_j(i.j) = I(i.j+1) - I(i.j-1) \qquad (8)$$

$$G(i.j) = \sqrt{G_i(i.j)^2 + G_j(i.j)^2} \qquad (9)$$

$$\alpha_0(i.j) = \tan^{-1}\left[\frac{G_j(i.j)}{G_i(i.j)}\right].\alpha_0\epsilon\left[-\frac{\pi}{2}.\frac{\pi}{2}\right] \qquad (10)$$

where $G_i(i.j)$, $G_j(i.j)$ are the derivatives in the horizontal and vertical directions at pixel (i ,j).

The second phase involves the generation of the HOG descriptor, which is built based on the gradient of the image. The whole image is split into blocks with sizes [2 2], [4 4], and [8 8]. The gradient direction range [-π/2, π/2] is calculated equally into nine direction intervals (bins). To provide a strong vector to brightness changes, the HOG feature values are normalized by segmenting each bin with the total of the histogram. Fig. 5 shows different block sizes of the sub-band HH of the Mel-spectrogram image.

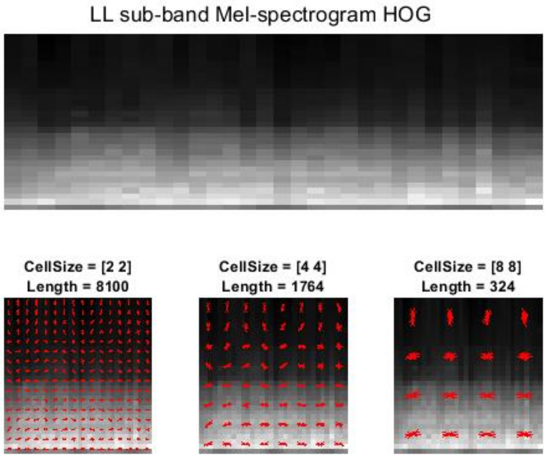


Fig. 5. HOG features of Mel-spectrogram image LL sub-band with different cell sizes.

In this paper the HOG Cell Size of [4 4] is used for extracted 1764 feature value for Each sub-bands of Mel-spectrogram image and the final feature vectored of this step was 7056 feature values.

5) Combining the features of the proposed system and PCA: This part describes a hybrid feature extraction technology. The characteristic of this method is the merger of characteristics derived from HOG, GLCM, WDT, and LBP. The suggested approach is distinguished by its expeditiousness in training the dataset and its demand for computer resources of moderate cost. At first, a Haar 2D wavelet is It is used for extraction 4 sub-bands from the Mel-spectrogram image and extract GLCM, HOG, and LBP for all sub-bands for training the Mel-spectrogram image to create a 440 x 7308 matrix of features. The PCA algorithm [44] is applied to the feature matrix in order to minimize the dimensions and select the most appropriate characteristics for each image.

C. Ml Algorithms

In this paper, many ML algorithms were used to classify the Mel-spectrogram image.

1) K-NN classifier: The K-NN technology is a basic data mining strategy where all samples are assigned to the same group in a feature space, and the algorithm has the same properties for both regression and classification [45]. This technique is considered effective for classification problems.

2) LR classifier: This Classifier [46] is a Strong statistic tool for developing resilient methods. It applies the linear regression principle to classification problems. It predicts dependent data by examining the connection between one or more pre-existing independent variables. The LR formula is represented by the following equation:

$$P = \frac{e^y}{1+e^y} \qquad (11)$$

3) SVM classifier: SVM is a collection of supervised learning algorithms. These are commonly used for classification and regression tasks on both linear and nonlinear data [47]. This method finds a decision boundary among two classes in order to forecast labels using one or more feature vectors. It lacks a natural growth to several courses and performs slowly throughout training.

4) DT classifier: One of the most powerful categorization methods is the DT, which simulates decisions using a tree framework. [48]. Computing the data allows it to classify the dataset and assign values to each of its attributes. The decision tree follows a top-down approach. Information gain is a typical strategy for choosing a decision node in DT. The equation for information gain is as follows:

$$IG(D.A) = Entropy(D) - \sum_{v\in Values(A)}\frac{|D_v|}{|D|}\cdot Entropy(D_v) \qquad (12)$$

5) NB classifier: NB is a class of supervised learning approaches which employ likely reasoning to anticipate the optimal result [49]. Using the Bayes theorem to it is easy to construct the classifier and the Gaussian normal distribution to forecast the class. The collection of probabilities for a certain set of data is determined by counting the value and frequency of the value. The Bayesian formula is:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \quad (13)$$

6) RF classifier: The RF technique, which is based on Ho's method and was later developed and introduced to the literature by [50], is a collective learning technique that determines the output class by training a large number of decision trees and taking the mode or average of their results. It's a popular algorithm because of its high prediction performance, capacity to deal with imbalance issues, and ability to produce consistent results in a variety of applications.

7) AdaBoost classifier: The adaboost boosting algorithm is a well-known ensemble technique for binary classification [51]. The group moves toward AdaBoost trains and installs trees in a sequential manner. AdaBoost combines a series of weak classifiers to perform boosting. The goal of each iteration of the weak classifier is to correct samples that were incorrectly classified by the preceding weak classifier. AdaBoost uses an iterative approach to help bad classifiers get better by using the mistakes they have made.

8) XGB classifier: XGB is an additional ensemble ML method that tackles regression and classification issues by utilizing many decision trees [52]. To lessen overfitting and boost performance, it uses more regularized prediction models.

Table II illustrates the hyperparameter method for determining the optimal parameters for ML algorithms.

TABLE II. THE FINE HYPERPARAMETERS OPTIMIZATION

Model	Hyperparameters
K-NN	n_neighbors=5, Euclidean distance
LR	solver='linear'
DT	max_depth=100, criterion='entropy'
NB	var_smoothing=1e−04
RF	n_estimators=100, max_depth=50
AdaBoost	n_estimators=20, learning rate=0.5 3.3
XGB	n_estimators=100, learning rate=0.1
SVM	Kernel= RBF, C=3.3

IV. RESULTS AND DISCUSSION

A. Dataset

1) Collected data: In this paper, authors have collected a music audio dataset consisting of 440 recordings for different four classes of Arabic music sounds has been collected.

The collected data became available for free at: (https://www.kaggle.com/datasets/emanatyaesmaeil/zekrayati-dataset).

2) GTZAN dataset: GTZAN [53] was one of the first widely available datasets for MGC and is well-known within the scientific community. This dataset contains (1000) music clips from (10) western music genres. GTZAN's ten genres ('blues', 'classical', 'country', 'disco', 'hip-hop', 'jazz',' metal', 'pop', 'reggae', and 'rock') are evenly distributed, each featuring '100' clips. Each music clip is (30) seconds long. Table III shows the description of the dataset.

TABLE III. DESCRIPTION OF THE DATASET

Class Name	No of file	Minimum duration (s)	Maximum duration (s)
Zekrayati (P_1)	120	12	60
Zekrayati (P_2)	120	26	41
Zekrayati (P_3)	80	12	35
Zekrayati (P_4)	120	46	91

B. Performance Evaluation

The results related to ML models have been measured using the following indicators: accuracy, precision, recall, and F1-score. Equations (14 through 17) employ the confusion matrix to calculate these values. Where (TP=true_positive), (FP=false_positiv), (FN=false_negative), and (TN=true_negative) [54].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$Recall = \frac{TP}{TP+TN} \quad (15)$$

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (17)$$

C. K-Fold Cross Validation

In this method, epochs are split up into k groups at random, each with roughly the same amount of features. The remaining groups are used to test the learning, while the K-1 groups are used for training. The technique is repeated (k) times, every time with a different group of tests [55]. For performance evaluation, a 5-fold cross-validation procedure is used, and the result is calculated as the 5-fold average.

D. Experiment 1: ML Methods with Proposed Dataset

In this paper, all samples in the proposed dataset converted to Mel-spectrogram images, then three levels of DWT with haar mother wavelet function were used to represent each Mel-spectrogram images, then chosen sub-band of level three for feature extracted by using GLCM, LBP, and HOG and reducing feature vector using PCA, and finally used eight machine learning techniques SVM, K-NN, DT, RF, LR, NB, XGB, and AdaBoost for classification Mel-spectrogram images dataset using 5 k- Fig. 6 depicts the splitting of Mel-spectrogram pictures into 5 k-folds for training and testing. Table IV and Fig. 7 show the level of accuracy ratings of models after applying 5 k-fold. Fig. 8 to Fig. 12 depicts the confusion matrix for various ML methods.

According to Table IV, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 99.54% for accuracy, 99.44% for F1 score, 99.41% for recall, and 99.51% for precision.

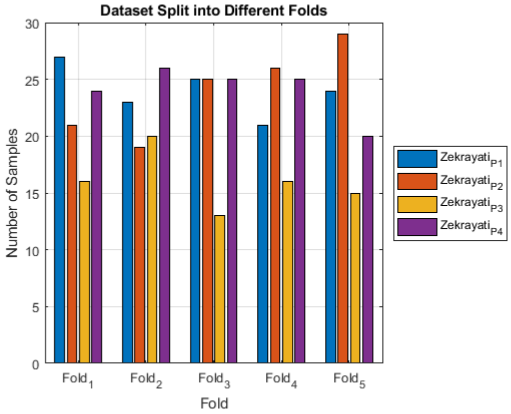


Fig. 6. The split of Mel-spectrogram images into train and test using 5 k-fold.

TABLE IV. SHOWS THE ACCURACY RESULTS OF VARIOUS ML MODELS

Model	Fold	Accuracy	F1-Score	Recall	Precision
K-NN	1	98.86	98.77	99.07	98.53
	2	97.73	97.52	97.60	97.49
	3	98.86	98.56	99.00	98.21
	4	97.73	97.94	97.85	98.15
	5	98.86	99.04	98.96	99.17
	Mean	**98.41**	**98.37**	**98.50**	**98.31**
LR	1	96.59	96.67	96.84	96.56
	2	95.45	95.57	95.39	95.81
	3	96.59	96.54	97.00	96.25
	4	97.73	97.94	97.85	98.15
	5	97.73	97.47	98.10	97.06
	Mean	**96.82**	**96.84**	**97.04**	**96.77**
SVM	1	98.86	98.94	98.81	99.11
	2	98.86	98.83	98.91	98.81
	3	100	100	100	100
	4	98.86	98.72	98.44	99.07
	5	100	100	100	100
	Mean	**99.32**	**99.30**	**99.23**	**99.40**
DT	1	86.36	86.20	86.62	86.14
	2	87.50	87.46	87.35	87.61
	3	86.36	85.50	86.15	85.08
	4	86.36	86.11	86.00	86.37
	5	87.50	86.35	86.55	87.01
	Mean	**86.82**	**86.32**	**86.53**	**86.44**
NB	1	88.64	88.35	88.48	88.73
	2	89.77	90.08	89.97	90.45
	3	88.64	88.19	88.15	88.27
	4	89.77	89.72	89.61	90.06
	5	88.64	88.65	87.68	90.43
	Mean	**89.09**	**89.00**	**88.78**	**89.59**
RF	1	98.86	98.88	98.81	99.00
	2	97.73	97.60	97.43	97.82
	3	97.73	96.94	96.15	98.08
	4	97.73	97.93	98.04	97.86
	5	98.86	98.86	98.96	98.81
	Mean	**98.18**	**98.04**	**97.88**	**98.31**
AdaBoost	1	98.86	98.71	98.96	98.53
	2	97.73	97.70	97.79	97.71
	3	100	100	100	100
	4	98.86	98.63	98.81	98.53
	5	100	100	100	100
	Mean	**99.09**	**99.01**	**99.11**	**98.95**
XGB	1	98.86	98.71	98.96	98.53
	2	100	100	100	100
	3	98.86	98.51	98.08	99.04
	4	100	100	100	100
	5	100	100	100	100
	Mean	**99.54**	**99.44**	**99.41**	**99.51**

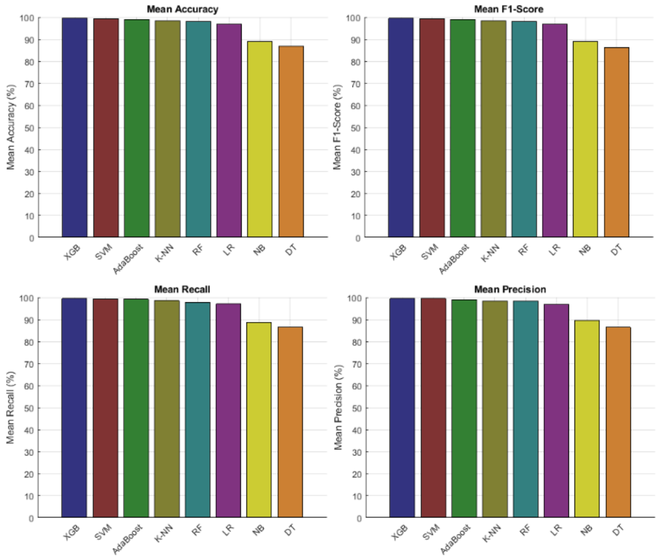


Fig. 7. Result for ML models.

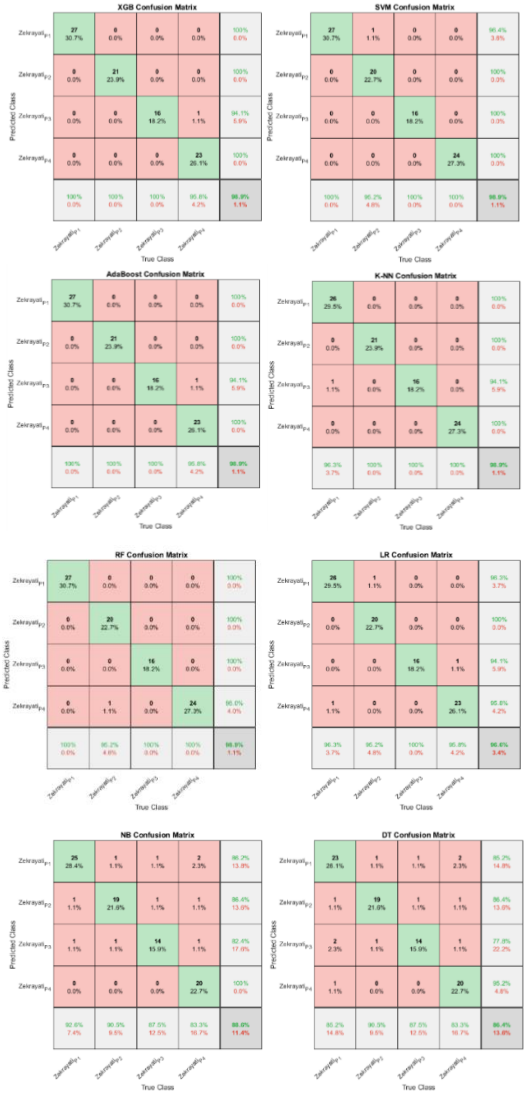


Fig. 8. Shows the confusion matrix of various ML models for fold-1.

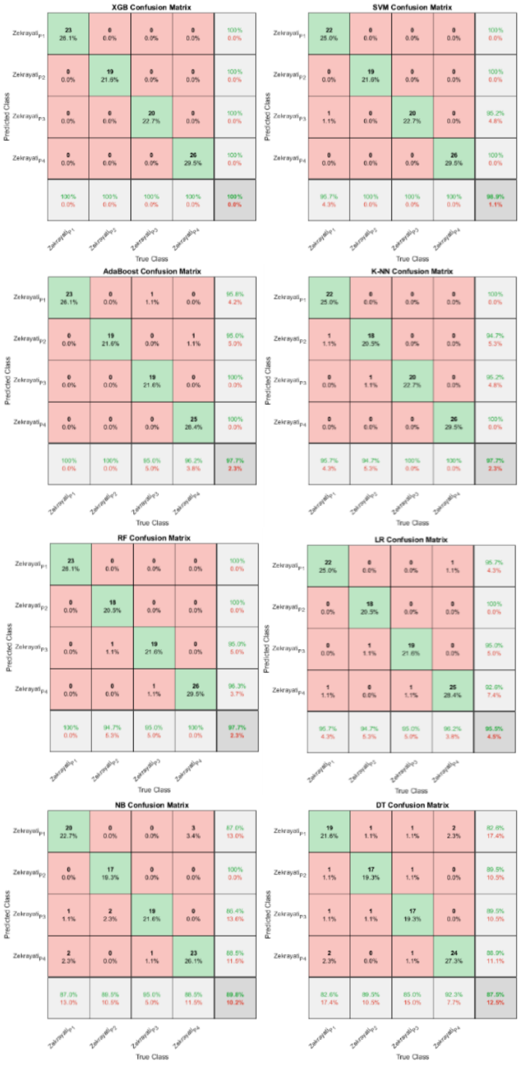
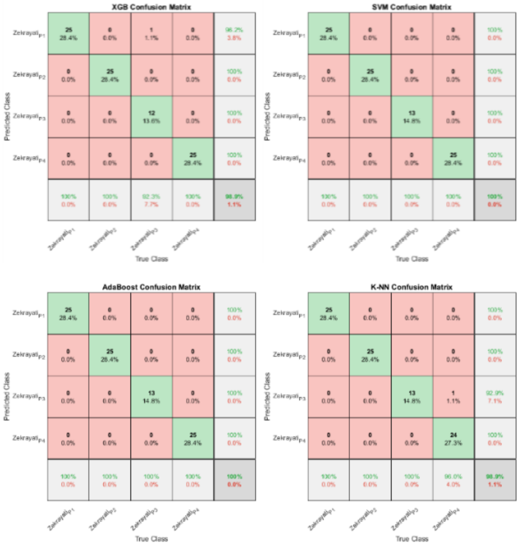


Fig. 9. Shows the confusion matrix of various ML models for fold-2.



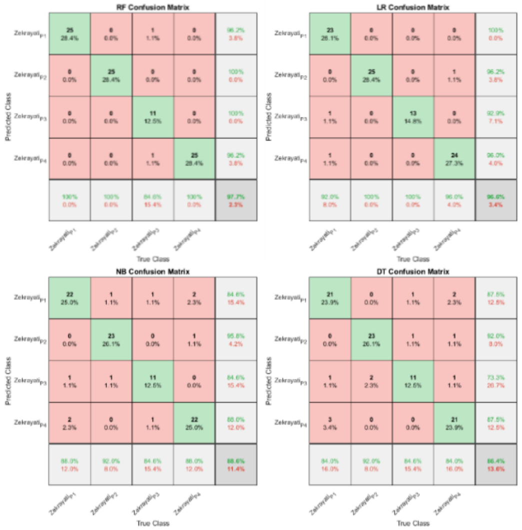


Fig. 10. Shows the confusion matrix of various ML models for fold-3.

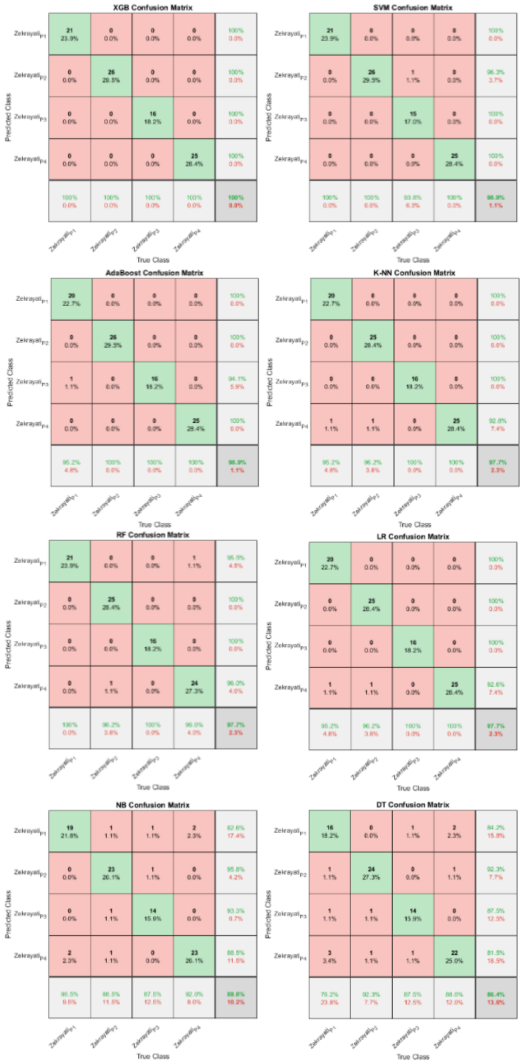


Fig. 11. Shows the confusion matrix of various ML models for fold-4.

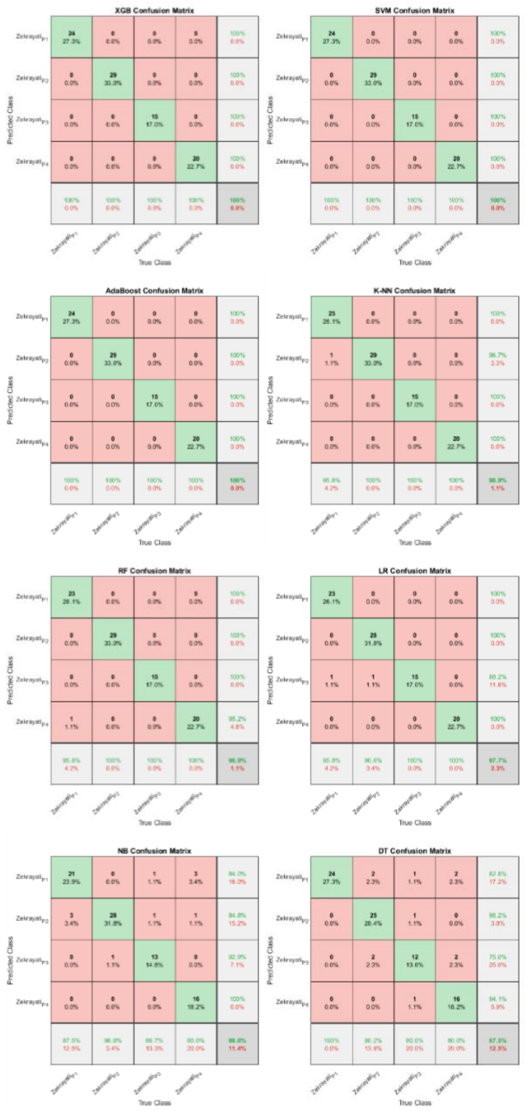


Fig. 12. Shows the confusion matrix of various ML models for fold-5.

E. Experiment 2: ML Methods with GTZAN Dataset

the approach suggested had been compared to state-of-the-art models that used the GTZAN dataset, comprising DL models, specifically convolutional neural networks, bottom-up broadcast neural networks (BBNN), deep unsupervised representation learning from acoustic data auDeep, and ML SVM for categorizing music genres using MEL-spectrogram images and the GTZAN dataset. The steps of the comparison process can be summarized as follows

Step 1: Convert all class to MEL spectrogram images using window length 1024 with overlap length 512, FFT length 4096 and number bands 64 Fig. 13 shown some MEL spectrogram images for GTZAN dataset.

Step 2: DWT is applied using three levels and calculates GLCM, HOG, and LBP for level 3 for each sub-band, and finally reduces the feature vector to 1000 samples and 100 features using PCA.

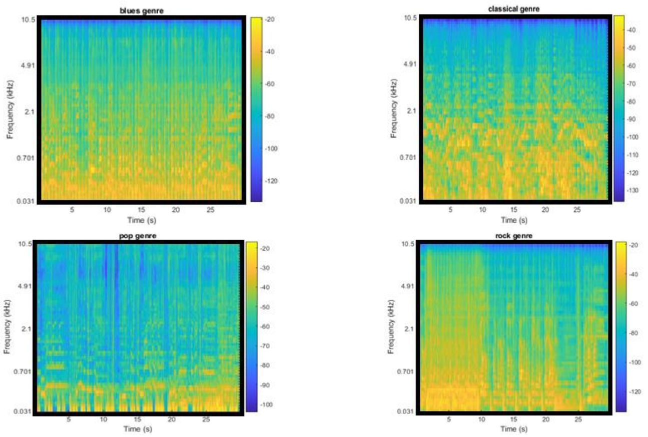


Fig. 13. Mel-spectrogram for some GTZAN dataset.

Step 3: Classification genres using ML models proposed in this paper and the model's performance using Accuracy, F1-Score, Recall and Precision using 5k-fold Fig. 14 shows 5 K-fold split of the GTZAN dataset 20% for testing and 80% for training. Fig. 15 shows the results of ML models for GTZAN dataset. Table V illustrates the accuracy scores and Fig. 16 to 20 shown confusion matrix's.

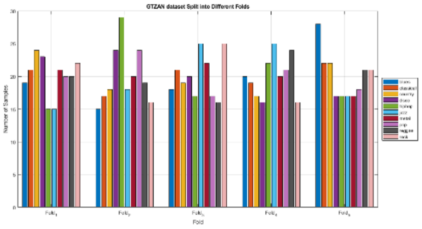


Fig. 14. Using five k-folds, GTZAN splits Mel-spectrogram pictures into train and test.

TABLE V. THE ACCURACY SCORES OF DIFFERENT ML MODELS FOR GTZAN

Model	Fold	Accuracy	F1-Score	Recall	Precision
K-NN	1	95.50	95.29	95.32	95.49
	2	95.00	94.66	94.67	94.87
	3	95.00	94.88	95.27	94.90
	4	95.00	94.69	94.97	95.02
	5	94.50	94.35	94.45	94.60
	Mean	**95.00**	**94.77**	**94.94**	**94.98**
LR	1	93.00	92.91	92.76	93.24
	2	92.50	92.02	92.37	91.92
	3	92.50	92.40	92.46	92.96
	4	92.50	92.42	92.98	92.92
	5	93.00	92.91	92.86	93.23
	Mean	**92.70**	**92.53**	**92.69**	**92.85**
SVM	1	97.50	97.39	97.53	97.30
	2	97.00	96.82	97.03	96.74
	3	97.50	97.53	97.72	97.46
	4	97.00	96.91	97.01	97.13
	5	97.50	97.49	97.30	97.82
	Mean	**97.30**	**97.23**	**97.32**	**97.29**
DT	1	82.50	82.70	82.67	83.43
	2	82.00	81.63	81.73	82.76
	3	82.50	82.01	82.50	82.38
	4	82.00	81.58	81.34	83.39
	5	83.00	81.94	82.52	83.48
	Mean	**82.40**	**81.97**	**82.15**	**83.09**
NB	1	86.50	86.31	86.49	87.05

	2	87.00	87.16	87.82	89.10
	3	87.00	86.53	86.55	88.04
	4	86.50	86.23	86.59	86.63
	5	87.50	86.65	86.37	88.30
	Mean	**86.90**	**86.58**	**86.76**	**87.82**
RF	1	96.50	96.20	96.04	96.47
	2	96.00	95.86	96.00	95.95
	3	96.00	95.86	95.96	95.95
	4	96.00	95.87	95.89	96.18
	5	96.00	96.02	95.94	96.14
	Mean	**96.10**	**95.96**	**95.97**	**96.14**
AdaBoost	1	97.00	96.88	96.92	96.98
	2	97.00	96.80	97.03	96.85
	3	96.50	96.33	96.34	96.49
	4	96.50	96.29	96.24	96.55
	5	97.00	96.99	97.18	96.98
	Mean	**96.80**	**96.66**	**96.74**	**96.77**
XGB	1	98.00	97.91	97.90	97.98
	2	97.50	97.35	97.38	97.47
	3	98.00	97.95	97.97	97.98
	4	97.50	97.45	97.64	97.49
	5	98.00	97.96	97.87	98.13
	Mean	**97.80**	**97.72**	**97.75**	**97.81**

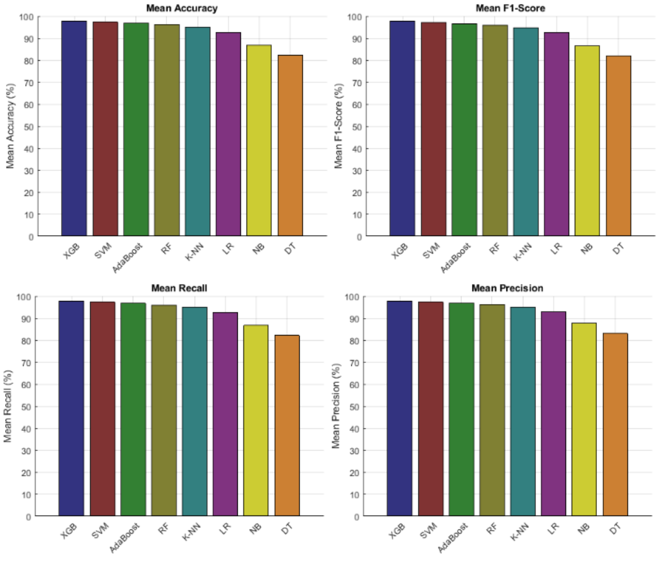
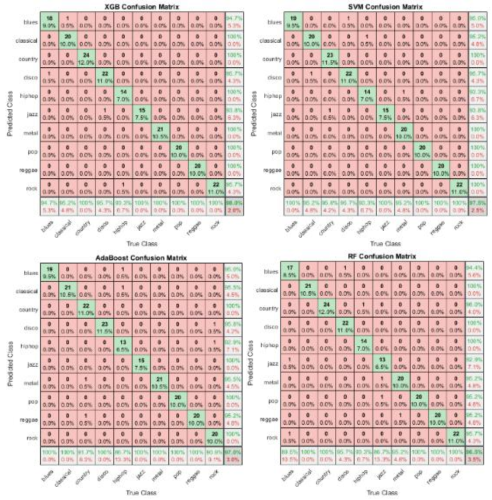


Fig. 15. Results of ML Models for GTZAN dataset.



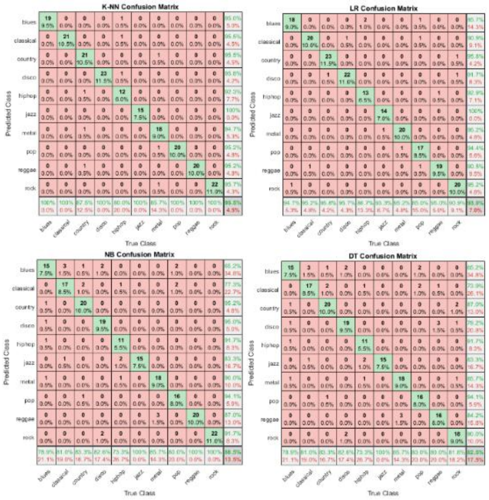


Fig. 16. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-1.

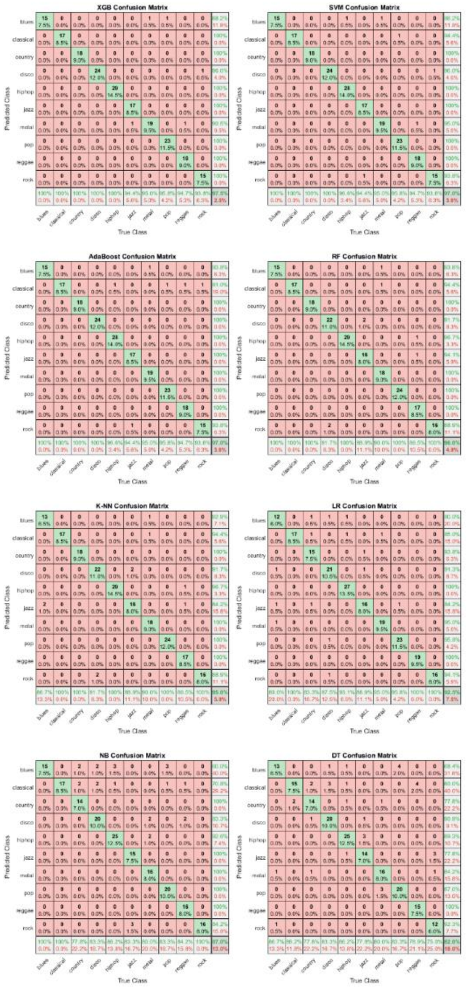


Fig. 17. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-2.

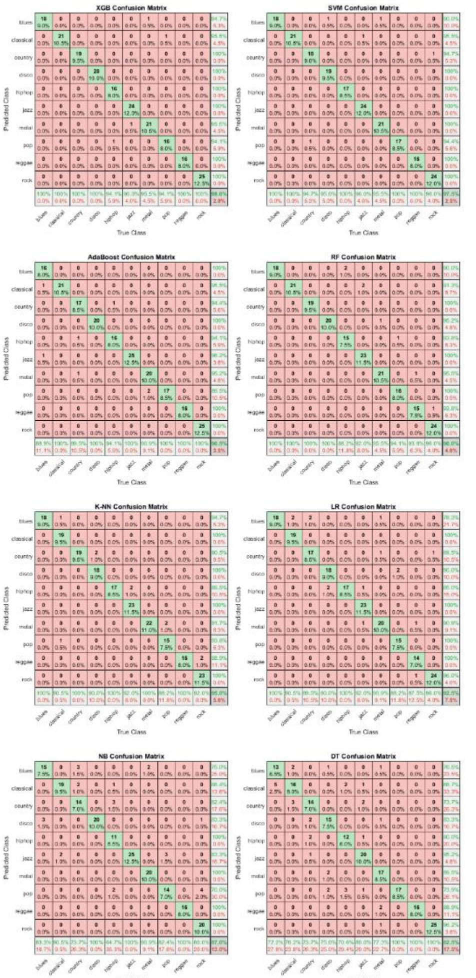
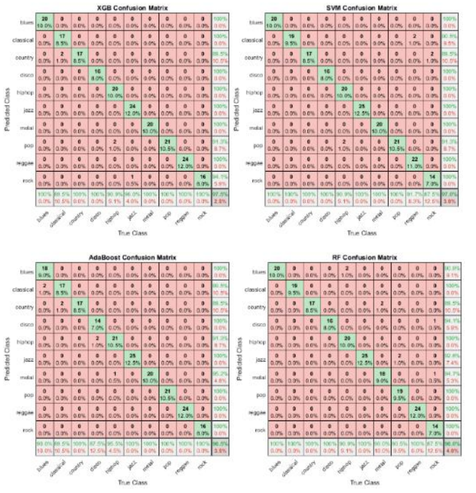

XGB Confusion Matrix
SVM Confusion Matrix
AdaBoost Confusion Matrix
RF Confusion Matrix
K-NN Confusion Matrix
LR Confusion Matrix
NB Confusion Matrix
DT Confusion Matrix

Fig. 18. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-3.


XGB Confusion Matrix
SVM Confusion Matrix
AdaBoost Confusion Matrix
RF Confusion Matrix

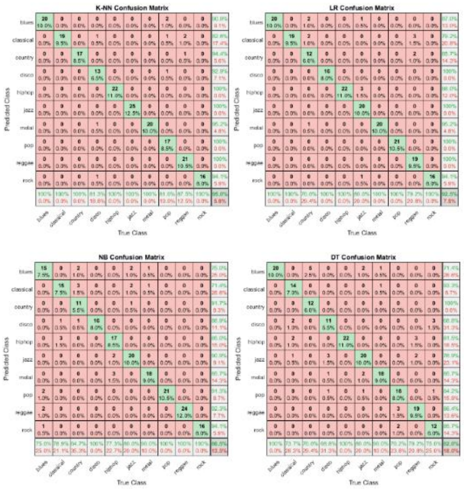

K-NN Confusion Matrix
LR Confusion Matrix
NB Confusion Matrix
DT Confusion Matrix

Fig. 19. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-4.

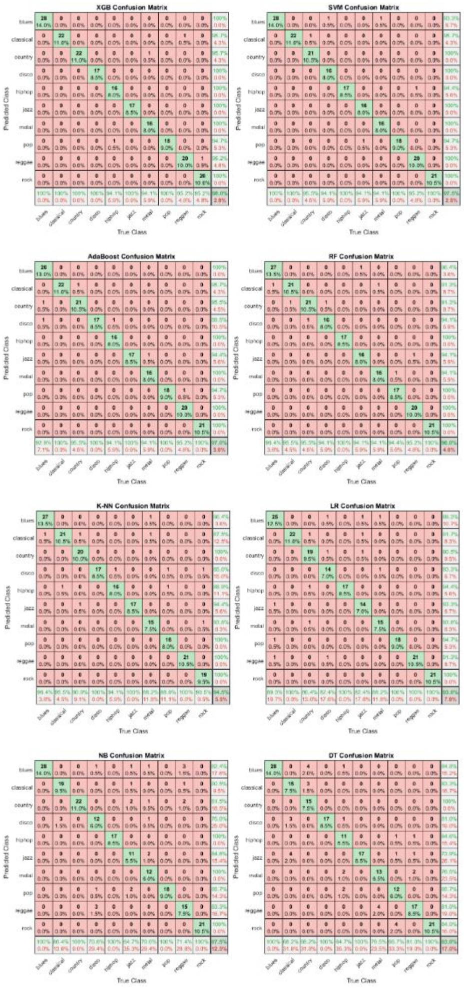

XGB Confusion Matrix
SVM Confusion Matrix
AdaBoost Confusion Matrix
RF Confusion Matrix
K-NN Confusion Matrix
LR Confusion Matrix
NB Confusion Matrix
DT Confusion Matrix

Fig. 20. Displays the confusion matrix of multiple ML models on the GTZAN dataset fold-3.

According to Table V, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Table VI shown comparison accuracy with the GTZAN dataset using XGB classifier.

V. Discussion

One of the main objectives of this work is to evaluate the performance of proposed ML models using two different dataset: the collected dataset that the authors had gathered and the global GTZAN dataset.

For the collected dataset , the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 99.54% for accuracy, 99.44% for F1 score, 99.41% for recall, and 99.51% for precision.

By using GTZAN , According to Table VI, the XGB model achieves high performance in the music classification when compared to other classifiers. The XGB achieved an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Table VI shown comparison accuracy with GTZAN dataset using XGB classifier.

TABLE VI. Comparison Accuracy with GTZAN Dataset

Reference	Feature	Model	Accuracy (%)
Liu, Caifeng, et al. [56]	MEL spectrogram	BBNN network	93.9
Nanni et al. [57]	MEL spectrogram	SVM	90.9
Ghildiyal et al. [58]	MEL spectrogram	CNN	91.00
Nakashika et al. [59]	MEL spectrogram + GLCM	CNN	72.00
Yang et al. [60]	MEL spectrogram	CNN	90.7
Freitag et al. [61]	MEL spectrogram	AuDeep	85.4
Proposed Method using XGB	MEL spectrogram + GLCM + HOG + LBP	XGB	97.80

VI. Conclusion and Future Work

ML techniques are beneficial for classification tasks, especially music genre classification, in which music is classified into different genres concerning its features. The objective of this paper is the classification of musical sound using ML techniques with texture features and Mel-Spectrogram. In the methodology, the audio data was transformed into a Mel-spectrogram, then texture features were applied to extract the audio features, and finally, a classification task was carried out using six ML classifiers. We performed a complete comparison of six ML classifiers in this study. By using two different datasets, the experimental findings indicated that the XGB exhibited a high accuracy with an average performance of 97.80% for accuracy, 97.72% for F1 score, 97.75% for recall, and 97.81% for precision. Comparing the reviewed related works mainly implemented using various ML and DL algorithms, our method obtained higher accuracy on automatic classification for music.

Future enhancements: Current work processes audio files that are 30 to 90 seconds long. More research should be conducted to handle audio of any length.as well as , Implementing music genre classification for other audio formats can be investigated, as the established ML models perform well for the (.WAV) format, but there are many other formats available, including MP3, FLAC, and others.

Finally, in future work, the authors can combine CNN approaches with texture features to enhance computational efficiency, minimize processing time, and identify music subgenres.

References

[1] A. Kumar, A. Rajpal and D. Rathore, “Genre classification using feature extraction and deep learning techniques,” in 2018 10th Int. Conf. on Knowledge and Systems Engineering (KSE), pp. 175– 180, 2018. doi:10.1109/kse.2018.8573325.

[2] K.-K. R. Choo, M. M. Kermani, R. Azarderakhsh, and M. Govindarasu, “Emerging embedded and cyber physical system security challenges and Innovations,” *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 3, pp. 235–236, May 2017. doi:10.1109/tdsc.2017.2664183.

[3] B. Koziel, R. Azarderakhsh, and M. Mozaffari-Kermani, “Low-resource and fast binary Edwards curves cryptography,” *Lecture Notes in Computer Science*, pp. 347–369, 2015. doi:10.1007/978-3-319-26617-6_19.

[4] L. Liu, “Lute acoustic quality evaluation and note recognition based on the Softmax regression BP neural network,” *Mathematical Problems in Engineering*, vol. 2022, pp. 1–7, Apr. 2022. doi:10.1155/2022/1978746.

[5] L. Almazaydeh, S. Atiewi, A. Al Tawil, and K. Elleithy, “Arabic music genre classification using deep convolutional neural networks (cnns),” *Computers, Materials &amp; Continua*, vol. 72, no. 3, pp. 5443–5458, 2022. doi:10.32604/cmc.2022.025526.

[6] F. Ahmed, P. P. Paul, and M. Gavrilova, “Music genre classification using a gradient-based local texture descriptor,” *Smart Innovation, Systems and Technologies*, pp. 455–464, 2016. doi:10.1007/978-3-319-39627-9_40.

[7] A. S. Girsang, A. S. Manalu, and K.-W. Huang, “Feature selection for musical genre classification using a genetic algorithm,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 2, pp. 162–169, 2019. doi:10.25046/aj040221.

[8] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, “Musical Instrument Sound classification using GoogleNet with SVM and KNN Model,” *Lecture Notes in Networks and Systems*, vol. 300, pp. 230–240, Sep. 2021. doi:10.1007/978-3-030-84760-9_21.

[9] M. Chaudhury, A. Karami, and M. A. Ghazanfar, “Large-scale music genre analysis and classification using Machine Learning with apache spark,” *Electronics*, vol. 11, no. 16, p. 2567, Aug. 2022. doi:10.3390/electronics11162567.

[10] X. Mu, “Implementation of music genre classifier using KNN algorithm,” *Highlights in Science, Engineering and Technology*, vol. 34, pp. 149–154, Feb. 2023. doi:10.54097/hset.v34i.5439.

[11] Y. M. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, “Comparing textural features for music genre classification,” *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, Jun. 2012. doi:10.1109/ijcnn.2012.6252626.

[12] L. Nanni, Y. Costa, and B. Sheryl , “Set of texture descriptors for music genre classification,” In proceeding of the 22nd WSCG International Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 2014..

[13] B. L. Sturm, “An analysis of the GTZAN Music Genre Dataset,” Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, Nov. 2012. doi:10.1145/2390848.2390851.

[14] A. Yadav, S. Gaikwad, T. Kuigade, and A. Patil, “MUSIC CHORD PREDICTION USING MACHINE LEARNING,” *International*

Research Journal of Modernization in Engineering Technology and Science, vol. 5, no. 12, pp. 194–199, 2023. doi:10.56726/IRJMETS46945.

[15] S. O. Folorunso, S. A. Afolabi, and A. B. Owodeyi, “Dissecting the genre of Nigerian music with Machine Learning Models,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6266–6279, Sep. 2022. doi:10.1016/j.jksuci.2021.07.009.

[16] D. D. Himabindu, K. Avaneesh, M. A. Mudiraj, S. M. Reddy, and M. S. Varma, “Music genre classification using XGB Boost,” *Springer Proceedings in Mathematics &amp; Statistics*, pp. 269–276, 2024. doi:10.1007/978-3-031-51167-7_26.

[17] A. Bawitlung and S. K. Dash, “Genre classification in music using Convolutional Neural Networks,” *Lecture Notes in Computer Science*, pp. 397–409, Oct. 2023. doi:10.1007/978-981-99-7339-2_33.

[18] T. Li, “Optimizing the configuration of deep learning models for music genre classification,” *Heliyon*, vol. 10, no. 2, Jan. 2024. doi:10.1016/j.heliyon.2024.e24892.

[19] K. K. Jena, S. K. Bhoi, S. Mohapatra, and S. Bakshi, “A hybrid deep learning approach for classification of music genres using wavelet and Spectrogram analysis,” *Neural Computing and Applications*, vol. 35, no. 15, pp. 11223–11248, Jan. 2023. doi:10.1007/s00521-023-08294-6.

[20] J. Mehta, D. Gandhi, G. Thakur, and P. Kanani, “Music genre classification using transfer learning on log-based Mel Spectrogram,” in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021. doi:10.1109/iccmc51019.2021.9418035.

[21] M. Kiran Kumar et al., “Automated music genre classification through Deep Learning Techniques,” *E3S Web of Conferences*, vol. 430, p. 01033, 2023. doi:10.1051/e3sconf/202343001033.

[22] V. Phulmante, A. Bidkar, Y. Mundada, and P. Kulkarni, “Recognition of music genres using deep learning,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 5, pp. 936–942, 2022.

[23] T. Yin, “Music track recommendation using Deep-CNN and Mel Spectrograms,” *Mobile Networks and Applications*, Jul. 2023. doi:10.1007/s11036-023-02170-2.

[24] P. Ghosh, S. Mahapatra, S. Jana, and R. Kr. Jha, “A study on music genre classification using machine learning,” *International Journal of Engineering Business and Social Science*, vol. 1, no. 04, pp. 308–320, Apr. 2023. doi:10.58451/ijebss.v1i04.55.

[25] S. Chikkamath et al., “Indian music instrument classification using Deep Learning on embedded platforms,” *Lecture Notes in Networks and Systems*, pp. 301–313, 2024. doi:10.1007/978-981-99-9442-7_26.

[26] D. R. Ignatius Moses Setiadi et al., “Comparison of SVM, KNN, and Nb classifier for genre music classification based on metadata,” *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, vol. 475, pp. 12–16, Sep. 2020. doi:10.1109/isemantic50169.2020.9234199.

[27] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, “Convolutional neural networks approach for music genre classification,” in *2020 International Symposium on Computer, Consumer and Control (IS3C)*, Nov. 2020. doi:10.1109/is3c50286.2020.00109.

[28] J. Li et al., “An evaluation of deep neural network models for music classification using spectrograms,” *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 4621–4647, Feb. 2021. doi:10.1007/s11042-020-10465-9.

[29] M. Preetham, J. B. Panga, J. Andrew, K. Raimond, and H. Dang, “Classification of music genres based on Mel-frequency cepstrum coefficients using deep learning models,” *Lecture Notes in Electrical Engineering*, vol. 905, pp. 891–907, 2022. doi:10.1007/978-981-19-2177-3_83.

[30] D. R. Ignatius Moses Setiadi et al., “Effect of feature selection on the accuracy of music genre classification using SVM Classifier,” in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2020. doi:10.1109/isemantic50169.2020.9234222.

[31] T. Ahmed, M. A. Alam, R. R. Paul, Md. T. Hasan, and R. Rab, “Machine learning and deep learning techniques for genre classification of Bangla Music,” in *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, Feb. 2022. doi:10.1109/icaeee54957.2022.9836434.

[32] Q. Zhang, W. Lu, R. Wang, and G. Li, “Digital image splicing detection based on Markov features in block DWT domain*,” Multimedia Tools and Applications*, vol. 77, no. 23, pp. 31239–31260, Jun. 2018. doi:10.1007/s11042-018-6230-z.

[33] N. K. Naik, P. K. Sethy, A. G. Devi, and S. K. Behera, “Few-shot learning convolutional neural network for primitive Indian paddy grain identification using 2D-DWT injection and Grey Wolf optimizer algorithm,” *Journal of Agriculture and Food Research*, vol. 15, p. 100929, Mar. 2024. doi:10.1016/j.jafr.2023.100929.

[34] O. Gheyath and D. Q. Zeebaree, “ The Applications of Discrete Wavelet Transform in Image Processing: A Review ,” *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 31–43, 2020.

[35] H. Shayeste and B. M. Asl, “Automatic seizure detection based on gray level co-occurrence matrix of STFT imaged-EEG,” *Biomedical Signal Processing and Control*, vol. 79, p. 104109, Jan. 2023. doi:10.1016/j.bspc.2022.104109.

[36] R. Anand, T. Shanthi, R. S. Sabeenian, and S. Veni, “GLCM feature-based texture image classification using machine learning algorithms,” *EAI/Springer Innovations in Communication and Computing*, pp. 103–125, Oct. 2022. doi:10.1007/978-3-031-20541-5_5.

[37] M. Lv et al., “Sound recognition method for white feather broilers based on spectrogram features and the Fusion Classification Model,” *Measurement*, vol. 222, p. 113696, Nov. 2023. doi:10.1016/j.measurement.2023.113696.

[38] M. H. Daneshvari, E. Nourmohammadi, M. Ameri, and B. Mojaradi, “Efficient LBP-GLCM texture analysis for asphalt pavement raveling detection using extreme gradient boost,” *Construction and Building Materials*, vol. 401, p. 132731, Oct. 2023. doi:10.1016/j.conbuildmat.2023.132731.

[39] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, “Music genre classification using LBP textural features,” *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, Nov. 2012. doi:10.1016/j.sigpro.2012.04.023.

[40] A. E. Salazar, “CLBP texture descriptor in multipartite complex network configuration for music genre classification,” *Procedia Computer Science*, vol. 222, pp. 331–340, 2023. doi:10.1016/j.procs.2023.08.172.

[41] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002. doi:10.1109/tpami.2002.1017623.

[42] C. Zhu, W. Zhao, and H. Lian, “Image recognition and classification with hog based on nonlinear support Tensor Machine,” *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 20119–20138, Dec. 2022. doi:10.1007/s11042-022-14320-x.

[43] J. N. Hasoon et al., “Covid-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images,” *Results in Physics*, vol. 31, p. 105045, Dec. 2021. doi:10.1016/j.rinp.2021.105045.

[44] E. M. Senan and M. E. Jadhav, “Diagnosis of dermoscopy images for the detection of skin lesions using SVM and KNN,” *Advances in Intelligent Systems and Computing*, vol. 1404, pp. 125–134, 2022. doi:10.1007/978-981-16-4538-9_13.

[45] M. N. Sikder and F. A. Batarseh, “Outlier detection using AI: A survey,” *AI Assurance*, pp. 231–291, 2023. doi:10.1016/b978-0-32-391919-7.00020-2.

[46] X. Zou, Y. Hu, Z. Tian, and K. Shen, “Logistic regression model optimization and case analysis,” in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019. doi:10.1109/iccsnt47585.2019.8962457.

[47] M. Awad and R. Khanna, “Support Vector Machines for classification,” *Efficient Learning Machines*, pp. 39–66, 2015. doi:10.1007/978-1-4302-5990-9_3.

[48] B. Charbuty and A. Abdulazeez, “Classification based on Decision Tree Algorithm for Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021. doi:10.38094/jastt20165.

[49] A. Zolnierek and B. Rubacha, "The empirical study of the naive bayes classifier in the case of Markov Chain Recognition Task," *Advances in Soft Computing*, vol. 30, pp. 329–336, 2005. doi:10.1007/3-540-32390-2_38.

[50] A. Börekci and O. Sevli, "A classification study for Turkish folk music makam recognition using machine learning with data augmentation techniques," *Neural Computing and Applications*, vol. 36, no. 4, pp. 1621–1639, Nov. 2023. doi:10.1007/s00521-023-09177-6.

[51] R. Wang, "AdaBoost for feature selection, classification and its relation with SVM, a review," *Physics Procedia*, vol. 25, pp. 800–807, 2012. doi:10.1016/j.phpro.2012.03.160.

[52] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on XGBoost for driving assessment and risk prediction," *Accident Analysis &amp; Prevention*, vol. 129, pp. 170–179, Aug. 2019. doi:10.1016/j.aap.2019.05.005.

[53] https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.

[54] M. H. Daneshvari, B. Mojaradi, M. Ameri, and E. Nourmohammadi, "Hybrid texture analysis of 2D images for detecting asphalt pavement bleeding and raveling using tree-based ensemble methods," *Alexandria Engineering Journal*, vol. 107, pp. 150–164, Nov. 2024. doi:10.1016/j.aej.2024.07.028.

[55] B. Oltu, M. F. Akşahin, and S. Kibaroğlu, "A novel Electroencephalography based approach for alzheimer's disease and mild cognitive impairment detection," *Biomedical Signal Processing and Control*, vol. 63, p. 102223, Jan. 2021. doi:10.1016/j.bspc.2020.102223.

[56] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7313–7331, 2020. doi:10.1007/s11042-020-09643-6.

[57] L. Nanni, Y. M. G. Costa, D. R. Lucio, C. N. Silla, and S. Brahnam, "Combining visual and acoustic features for audio classification tasks," *Pattern Recognition Letters*, vol. 88, pp. 49–56, 2017. doi:10.1016/j.patrec.2017.01.013.

[58] A. Ghildiyal, K. Singh, and S. Sharma, "Music genre classification using machine learning," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020. doi:10.1109/iceca49313.2020.9297444.

[59] T. Nakashika, C. Garcia, and T. Takiguchi, "Local-feature-map integration using convolutional neural networks for music genre classification," *Interspeech 2012*, pp. 1752–1755, 2012. doi:10.21437/interspeech.2012-478.

[60] H. Yang and W.-Q. Zhang, "Music genre classification using duplicated convolutional layers in neural networks," *Interspeech 2019*, pp. 3382–3386, Sep. 2019. doi:10.21437/interspeech.2019-1298.

[61] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2018.

DeeplabV3+ Model with CBAM and CSPM Attention Mechanism for Navel Orange Defects Segmentation

Guo Jinmei[1], Wan Nurshazwani Wan Zakaria[2]*, Wei Bisheng[3], Muhammad Azmi Bin Ayub[4]
College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia[1, 2, 4]
School of Mechanical and Electronic Engineering, Jiangxi College of Applied Technology, 341000 Ganzhou, China[1, 3]

Abstract—**Accurate defect detection of navel oranges is the key to ensuring the quality of navel oranges and extending their storage life. An improved DeeplabV3+ model integrating attention mechanism is proposed to increase the current low recognition accuracy and slow detection speed of defect detection in navel oranges grading and sorting process. The improved lightweight backbone network HECA-MobileV3 is applied in the DeeplabV3+ model to reduce the amount of computational data and improve the image processing speed. In addition, the Convolutional Block Attention Module (CBAM) and Channel Space Parallel Mechanism CSPM are integrated to the DeeplabV3+ model. ASPP structure is redesigned and the low feature extraction network is optimized to enhance the capture of target edge information and improve the segmentation effect of the model. Experimental results show that the proposed model exhibits a better MIoU and MPA with 89.50% and 94.02%, respectively, while reducing parameters by 49.42M and increasing detection speed by 55.6fps, which are 7.27% and 3.51% higher than the basic model. The results are superior than U-Net, SegNet and PSP-Net semantic segmentation networks. As a results, the proposed method provides better real-time performance, which meets the requirements of industrial production for detection accuracy and speed.**

Keywords—Navel oranges; defect detection; DeeplabV3+; HECA-MobileNetV3; CBAM attention mechanism; CSPM mechanism

I. Introduction

Jiangxi Gannan, home to the world's largest navel oranges plantation, has an annual output of up to one million tons annualy. However, despite its large-scale production, the harvesting and sorting processes still rely heavily on manual methods leading to high labour cost, prominent seasonality and high labour intensity. As the production of navel oranges increases year by year and sales channels continue to expand, the automation of grading and sorting of navel oranges are gradually emerging. After being picked from the trees, navel oranges need to go through disinfection and cleaning, sterilization and waxing, drying and weighing, colour and size grading, skin defect sorting, sugar content density quality analysis, and packaging and labeling before the fruit can be shipped to various parts of the world [1]. At present, the navel orange sorting line can quickly sort navel oranges based on their size and colour, but the recognition of local defects on the skin is not accurate at lower detection rate, which affected the overall quality, delayed the storage time and sorting process of navel oranges. Recent trends in machine vision and advancement in deep learning have led to a proliferation of studies that apply both methods in the field of navel oranges skin defect detection.

The traditional machine vision algorithm is mainly used to grade and classify navel orange defects based on the differences in data such as the RGB colour of the navel orange peel, surface brightness distribution, spectral imaging band curve and edge threshold. Abdelsalam et al. [2] detected the external defects of orange citrus fruits using multi-spectral imaging sensor. They segmented the defects based on the near-infrared (NIR) and RGB images of orange fruits and used threshold technology to detect defects in seven colour components of orange fruits. The overall accuracy of the algorithm exceeded 95%. Rong et al. [3] designed a fast edge detection algorithm for navel oranges surface defects to solve the problem of low defect detection accuracy caused by surface brightness by using the threshold edge segmentation method. Zhang et al. [4] proposed an Otsu threshold segmentation method based on image segmentation according to the different characteristics of navel orange surface defects, and the defect recognition rate is approximately 92.7%. Luo et al. [5] used a visible-near-infrared hyperspectral imaging system with a wavelength range of 325 to 1000 nm to collect citrus hyperspectral images, and used guided soft shrinkage (BOSS) and BOSS-SPA (BOSS-continuous projection algorithm) combined algorithms to optimize the spectral variables. Based on the extracted four defect wavelength images, they proposed a fast multispectral image processing algorithm combined with global threshold theory for rotten orange detection, with an overall classification accuracy of up to 98.6%.

Nowadays, deep learning techniques have been widely applied mainly on 1) object detection and 2) semantic segmentation. Object detection involves recognizing and locating target objects in an image, including algorithms such as R-CNN, YOLO and SSD. Semantic segmentation assigns semantic labels to each pixel in the image, including FCN, U-Net, Deeplab, SegNet and PSPNet. Iqbal et al. [6] determined the difference in fruit surface quality by training the RGB image combination data based on different fruit surface colour data. Asriny et al. [7] applied deep convolutional neural networks to grade the quality of navel oranges, establishing a database of over 1000 navel orange images and achieving a detection accuracy of 96% for different categories of navel oranges. Cai et al. [8] proposed a multi-resolution knowledge distillation strategy by integrating multi-scale pyramid modules and semi-resolution reconstruction branches, training the FastSegfermer model, which effectively improved the segmentation accuracy of the network, achieving a MIoU of 88.78%.

However, it can be concluded that there are several shortcomings in the navel orange defect detection research. The research up to now has been mainly based on traditional machine vision algorithms where the key challenge is the algorithm is too complex with computational burden resulting in difficulty to achieve real-time online detection. Although there are few common types of surface defects such as anthrax, sun spots and scratches, the highest chances to detect the similarity defect are low due to the nature of the fruit. Especially in the same image, precise segmentation is not achievable, particularly for small defects, thus failing to meet the requirements for grading and classification.

With the rapid development of computing power and artificial intelligence, researchers are also developed other fruit and vegetable detection using spectral technology, ultrasonic imaging and deep learning. Da Costa et al. [9] introduced the ResNet50 model into tomato external defect detection with the accuracy rate of 94.6%. Liang et al. [10] proposed a semantic segmentation method based on BiSeNet V2 deep learning for apple defect detection, and its average pixel accuracy MPA value approximately 99.66%. Hao et al. [11] applied the DeeplabV3+ model for kiwi defect recognition, using the lightweight convolutional neural network MobileNetV2 to extract image features, reducing the training time and achieving an average classification recognition rate of 96%. Gu et al. [12] used the phantom network and coordinate attention module to construct CA-ChostNet as the backbone feature extraction network of DeeplabV3+ for tomato target recognition, which reduced the number of network parameters while improving the model's segmentation capacity for small target categories. In order to improve the real-time performance of apple defect detection, Fan et al. [13] reduced the number of channels and network depth in the YOLOV4 network, reducing the size of the network model to 8.82MB with lower detection time of only 8.36ms for each image. It can be concluded that there is growing interest in deep learning application as main method for fruit and vegetable detection.

A. Navel Oranges Defect Types

There is multiple type of navel orange defects which are spots, scars, mildew, damage, blemishes, and enlarged fruit heads. If navel oranges are simply divided into good and defective fruits can cause great waste. Therefore, it is necessary to accurately identify the type and size of defects to better achieve navel orange grading and sorting.

Semantic segmentation algorithms include classic algorithms such as FCN, U-Net, SegNet, as well as modern deep learning algorithms such as PSPNet, Deeplab, and Mask R-CNN. Among them, the Deeplab network is a model with outstanding semantic segmentation performance at present, and has been gradually optimized from DeeplabV1 to DeeplabV3+. In 2018, DeeplabV3+ introduced an encoder-decoder structure, integrated multi-scale information, and improved the accuracy of image segmentation, becoming the most outstanding model for semantic segmentation. In view of the problems existing in the current research on navel oranges defect detection, combined with the problems of DeeplabV3+ model with many parameters and weak extraction of small target boundary features, this study proposes an improved DeeplabV3+ navel oranges defect real-time detection and segmentation model. The main research contributions of this paper are:

1) MobileNetV3 is used to replace the backbone network Xception, and the improved ECA attention mechanism is used to replace the SE mechanism in the MobileNetV3 network, which greatly reduces the amount of calculation parameters and improves the real-time performance of detection.

2) Redesigned the Atrous Spatial Pyramid Pooling (ASPP) structure of the DeeplabV3+ model. The CBAM attention mechanism and the CSPM mechanism are integrated to dynamically adjust the weight share of the feature channel to increase the attention to important areas of the image and comprehensively improve the recognition capacity of different types of navel orange defects.

3) The CBAM attention mechanism is added to the extraction of low-order feature information to make the extracted low-order features more representative and discriminative, and improve the segmentation effect and stcapacity of the model for boundary features.

4) Navel orange images are collected and a database of more than 2,000 navel orange defects is created to reduce model overfitting and provide more accurate and reliable model evaluation.

II. DeeplabV3+ Network Model with Improved Attention Mechanism

A. DeeplabV3+ Model

The Deeplab model was proposed in 2015. Over the years, with the continuous iteration and optimization of algorithms and technologies, DeeplabV1 [14], DeeplabV2 [15], and DeeplabV3 [16] models have been proposed one after another, continuously improving the model structure while improving the image segmentation. To address the issues of reduced image resolution, lower accuracy, and loss of details caused by max pooling and downsampling in deep convolutional neural networks (DCNNs), DeeplabV1 introduces the Atrous convolution algorithm and fully connected CRF structure. This approach expands the receptive field and connects DCNNs with CRF, thereby improving segmentation accuracy. DeeplabV2 improves the model's backbone network from VGG to ResNet and constructs an ASPP structure. This configuration captures information at multiple scales with high accuracy and capacity through parallel sampling. DeeplabV3 enhances the capacity to capture multi-scale information in images by varying the unit dilated rate of Atrous convolutions. To solve the problem of prolonged processing time and incomplete detail information in high-resolution images with DeeplabV3, DeeplabV3+ introduces an encoder-decoder structure, enhancing network capacity while ensuring the accuracy of feature extraction.

DeeplabV3+ consists of DCNN with dilated convolution and ASPP as the main structure of the encoder. Due to pooling and strided convolutions in the feature extraction process, some image details, particularly boundary features, are lost. To address this, the model integrates high-level features from the encoder with low-level features from the DCNN, enhancing boundary segmentation accuracy. The DCNN utilizes the Xception backbone, a complex structure with Entry, Middle, and Exit flow layers, leading to a high number of computational parameters and slower training and inference speeds. The ASPP module includes one 1x1 convolution, three 3x3 convolutions with different dilation rates, and one image pooling layer, aimed at dimensionality reduction, multi-scale context information extraction, and global context capture of the input image, respectively. The decoder connects low-level feature maps from the DCNN through a 1x1 convolution with the encoder's 4x upsampled high-level semantic feature maps. It further refines features with 3x3 convolutions and produces accurate prediction maps after another 4x upsampling. The DeeplabV3+ model architecture is illustrated in Fig. 1.

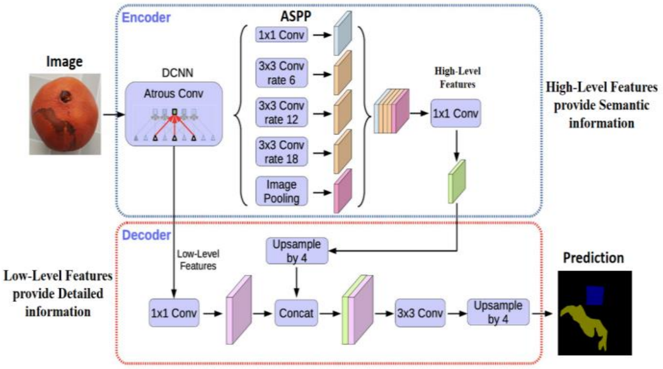


Fig. 1. DeeplabV3+ model structure diagram.

Although the DeeplabV3+ model has demonstrated excellent performance, it still faces challenges such as complex network and large amount of computation, limited capacity to capture details, and strong data dependency. Currently, many scholars have replaced lightweight backbone networks (such as MobileNet and Thin-xception) [17] to reduce the amount of computation and improve real-time performance, or introduced attention mechanisms (such as SE and CBAM) [18] in the model to improve the accurate segmentation of details, or added more multi-scale feature fusion modules (such as FPN [19] and ASPP [20]) to improve the recognition capacity for large-scale targets.

B. *Improved MobileNetV3 Backbone Network*

The DeeplabV3+ model employs the Xception network as the backbone for feature extraction, which inadequate for navel orange grading and sorting since the network required a large number of parameters and not suitable for the real-time application. Therefore, an improved lightweight MobileNetV3 backbone network is proposed to replace Xception.

The MobileNet network was proposed by the Google team in which MobileNetV3 is a lightweight network model that is continuously improved and optimized based on V1 and V2, which achieved excellent performance in tasks such as image classification and semantic segmentation [21]. Since traditional convolutional neural networks have large computational complexity and consume a lot of memory, depthwise convolution and pointwise convolution are combined in MobileNetV1 to construct a deep separable convolution structure, reducing its parameter volume and computational complexity to one square of the convolution kernel. After continuous verification, it was found that most of the computational parameters of V1's depthwise convolution were zero, limiting its effectiveness. Therefore, MobileNetV2 incorporated the inverted residual block and optimized activation functions, resulting in improvements in segmentation accuracy and processing time compared to the V1 structure. MobileNetV3 retains the depthwise separable convolution and inverted residual block from V2, adds the SE attention mechanism, updates activation functions, and redesigns the structure of time-consuming layers [22]. The network structure of MobileNetV3 is shown in Fig. 2.

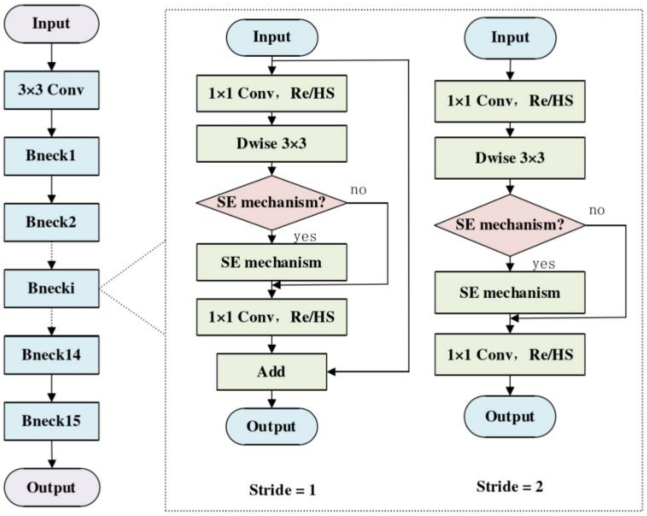
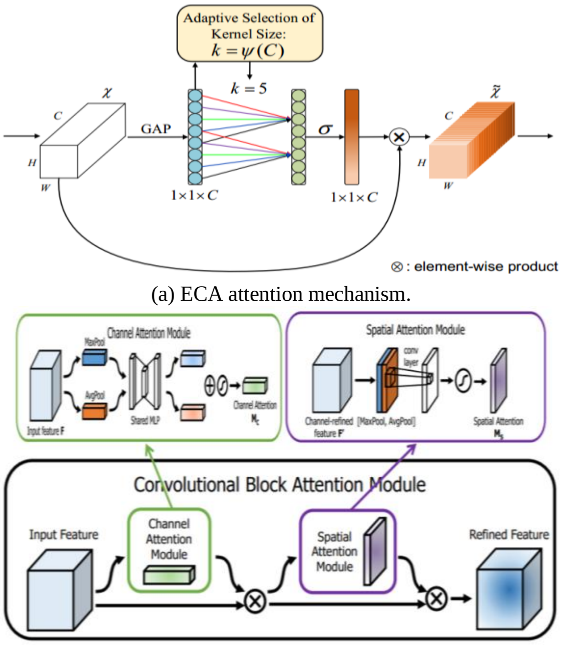


Fig. 2. MobileNetV3 model structure diagram.

The attention mechanism allows the model to improve its representation ability by focusing on different parts or features of the data during processing data. The main attention mechanisms currently include SE (Squeeze-and-Excitation Networks), ECA (Efficient Channel Attention Module), CA (Coordinate Attention) and CBAM (Convolutional Block Attention Module) [23]. The core idea of the SE module is to enhance the feature map by learning the importance of each channel. The SE module first uses a global average pooling operation to capture the global information of each channel, then learns the weight of each channel through two fully connected layers, and finally uses the sigmoid function to normalize the weight of each channel. The ECA module aims to reduce the computational costs by improving the SE model and capturing inter-channel relationships over a larger spatial range. It captures global information by adaptively determining the size of the one-dimensional convolution kernel through the channel dimension function. Since it does not involve global average pooling, it can reduce the computational cost. The CBAM module connects channel attention and spatial attention in series, allowing the model to dynamically focus on important information in the image in both the channel and spatial dimensions. The structures of the ECA and CBAM models are illustrated in Fig. 3.

Hard-sigmoid is used for the ECA attention mechanism instead of Sigmoid linear activation function as the function offers higher computational efficiency and effectively addresses the vanishing gradient problem. The improved H-ECA attention mechanism is shown in Fig. 4. After feature extraction, the Hard-sigmoid activation function is used to obtain the weight *w* of each channel, and finally multiplied with the corresponding element of the original feature map to obtain the final output feature map.



(a) ECA attention mechanism.

(b) CBAM attention mechanism.

Fig. 3. ECA and CBAM model structure diagram.

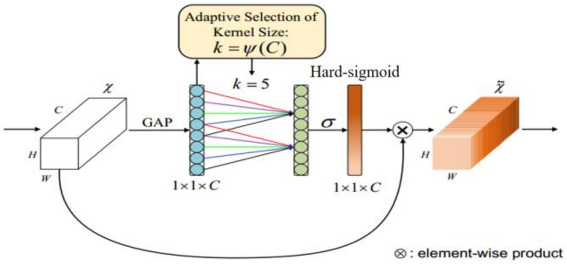


Fig. 4. Improved H-ECA attention mechanism.

To ensure the efficiency of navel oranges sorting, high recognition speed of images during the sorting process is required. The SE attention mechanism with high complexity is used in MobileNetV3, which will affect the response speed and real-time performance of the model. Therefore, to reduce the computational complexity and the number of parameters, the H-ECA mechanism replaces the SE mechanism in *Bneck*, and the feature information of different scales is better captured through local cross-channel convolution operations. The *Bneck* structure diagram of the improved backbone network HECA-MobileNetV3 is shown in Fig. 5.

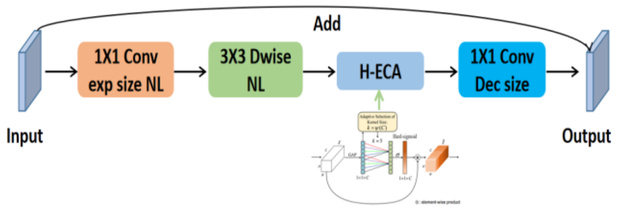


Fig. 5. Improved Bneck structure diagram.

The number of parameters occupied by the attention mechanisms are quantified in both the original MobileNetV3 and the improved HECA-MobileNetV3 networks, as well as their proportion of the total parameters as shown in Table I. The H-ECA attention mechanism replaced the SE attention mechanism in layers 3, 5, 6, 8, 12, 13, 14, and 15. Since H-ECA is not affected by the number of input and output channels, the size of its convolution kernel K is 5, Params = K*1+1, and the calculated Params is 6, accounting for almost 0.00%. It can be seen that the H-ECA attention mechanism occupies only a very small number of parameters.

TABLE I. STATISTICS OF ATTENTION MECHANISM PARAMETERS IN MOBILENETV3 NETWORK BEFORE AND AFTER IMPROVEMENT

Layer	SE		H-ECA	
	Params	Params Proportion	Params	Params Proportion
3	0.003M	0.107%	6	0.000%
5	0.007M	0.228%	6	0.000%
6	0.007M	0.228%	6	0.000%
8	0.105M	3.526%	6	0.000%
12	0.231M	6.722%	6	0.000%
13	0.231M	6.722%	6	0.000%
14	0.460M	13.166%	6	0.000%
15	0.460M	13.166%	6	0.000%
Total	1.504M	43.865%	48	0.000%

C. Improved DeeplabV3+ Network Model

In the actual grading and sorting of navel oranges, a large amount of image data needs to be collected. Even for the same navel orange, images need to be collected from multiple angles, and the detection speed of the results needs to be controlled within the millisecond. At the same time, although the shape of the navel orange does not change much, the details inside the peel are random and variable. The location, size and shape of the defects, the size of the navel and the thickness of the head are all different. Therefore, this study proposes an improved DeeplabV3+ lightweight network model that integrates the attention mechanism. The main improvements are in the following parts:

1) The improved lightweight backbone network HECA-MobileNetV3 is used to replace Xception, which not only reduces the model calculation amount and improves the real-time detection, but also is very easy to add to embedded systems or mobile devices, improving the applicability of the research.

2) In the ASPP structure of the DeeplabV3+ model, the CBAM attention mechanism and the Channel Space Parallel Mechanism (CSPM) were introduced to redesign the ASPP structure. CBAM uses the output of the channel mechanism as the input of the spatial mechanism, with a faster calculation speed and can gradually enhance the expressiveness of the feature map. At the same time, in order to address the key information loss problem that may exist in the CBAM mechanism, a parallel mechanism of channel attention and spatial attention is added to obtain the global dependency information of the input navel orange image in channels and space, and finally the splicing and fusion are passed to the decoding layer.

3) CBAM attention mechanism is added to the backbone network to extract low-level feature information of the image and adaptively adjust the feature map weights that can enhance key low-order features, suppress noise and redundant information, improve the model's generalization ability, and provide clearer and more effective features as shown in Fig. 6.

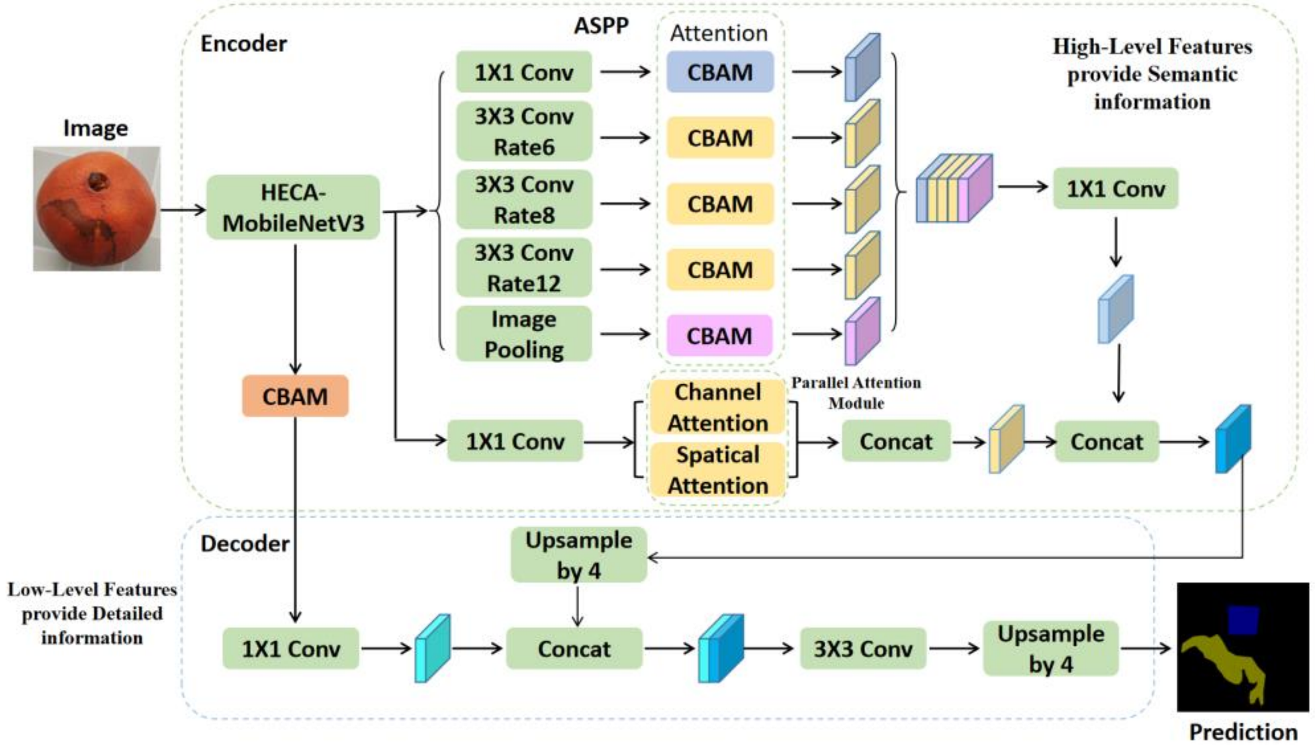


Fig. 6. Improved deeplabV3+ network model.

III. Data Collection and Experimental Evaluation Indicators

A. Data Collection

November to December every year is the harvest season for navel oranges in southern Jiangxi. The shelf life of navel oranges is as long as three to six months. The fruit is available everywhere in the market, which facilitates data collection. In actual fruit and vegetable grading and sorting equipment, multiple industrial depth cameras are often used to collect videos of navel oranges from different angles, convert the videos into pictures, and send them to the controller for intelligent processing [24]. In this study, a smartphone is used to collect image data with a resolution of 3200 x 1440 and 64 mega pixels. All images are collected in a static state, and the size of the collected images is 3472 x 4624. A high-performance computer is used to process image data and optimize, train and analyze deep learning network models. The computer specification is tabulated in Table II.

Various image of navel oranges placed under natural light in the living environment and artificial lighting after harvesting are recorded for database. A total of 800 original images were obtained, including 100 spotted, 100 navel, 100 moldy, 100 thick-heads, 200 damaged, and 200 other defects. Considering the limited number of images, data augmentation techniques were applied using Python's OpenCV and Pillow libraries to rotate, crop, and adjust the brightness of the images to enhance the generalization ability of the model. This process increased the dataset to 2400 images, with the image size is adjusted to 512 x 512.

The collected images were labeled using the *labelme* tool. By segmenting the details of the navel orange peel, six semantic labels were obtained, including spots, mildew, damaged, navel, head hypertrophy and other defects. The *json* file generated by the labeling tool is created as a dataset, and the data was divided into training set and test set in a ratio of 8:2, resulting in 1920 training sets and 480 validation sets.

TABLE II. Computer Hardware and Software Configuration

Hardware and Software	Configuration
Hardware	CPU：Intel(R) Core(TM) i7-14700KF CPU @ 3.4GHz
	GPU：NVIDIA RTX4070Ti Super 16G
	Memory：32G
	Operating System：Windows 11
Software	Deep Learning Frameworks：pytorch 2.2.0
	Image processing software：Open CV 14.2
	Compiled Language：Python 3.12.1

B. Experimental Evaluation Metric Parameters

This study uses the intersection over union (IoU), mean intersection over union (MIoU), and mean pixel accuracy (MPA) parameters to evaluate the accuracy of the model, and uses the parameter quantity and detection speed (FPS) of the model indicators to evaluate the capacity and real-time performance of the model. In all segmentation, recognition, and classification experiments, four types of results exist: true positive (TP), where the actual positive sample is correctly predicted as positive; true negative (TN), where the actual negative sample is correctly predicted as negative; false negative (FN), where the actual positive sample is incorrectly predicted as negative; and false positive (FP), where the actual negative sample is incorrectly predicted as positive. By evaluating the proportions of these outcomes, the effectiveness of the model's predictions can be determined [25]. MIoU is the average IoU of all different semantic categories, where IoU is the ratio of the intersection to the union between the predicted and the ground truth annotations. The formulas for calculating IoU and MIoU are as follows:

$$\mathrm{IoU} = \frac{\text{Predictive Value} \cap \text{True Value}}{\text{Predictive Value} \cup \text{True Value}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \frac{\mathrm{P_{ii}}}{\mathrm{P_{ij}}+\mathrm{P_{ji}}+\mathrm{P_{ii}}}$$

$$MIoU = \frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}}$$

MPA is the mean pixel accuracy, which calculates the proportion of correctly classified pixels for each semantic category and determine the average value. Due to the imbalanced distribution of positive and negative samples in this study, this indicator is can be used to measure the proposed method performance. The calculation formula is as follows:

$$\mathrm{MPA} = \frac{1}{\mathrm{K}+1}\sum\nolimits_{\mathrm{i}=0}^{\mathrm{k}}\frac{\mathrm{p_{ij}}}{\sum_{\mathrm{j}=0}^{\mathrm{k}}\mathrm{p_{ij}}}$$

where *i* represents the true value, *j* represents the predicted value, and *k* represents the number of semantic categories.

The size and complexity of the model have a great impact on the requirements of the device’s hardware performance. The number of model parameters is an important measurement indicator, which is related to the number of input and output channels and the size of the convolution kernel; the number of frames per second represents the images that the model can detect per second, and its calculation formula is as follows:

$$\mathrm{Params} = \sum_{\mathrm{l}=1}^{\mathrm{D}}\mathrm{K_l^2C_{l-1}C_l} + \sum_{\mathrm{l}=1}^{\mathrm{D}}\mathrm{C_l}$$

$$\mathrm{FPS} = \frac{n}{time}$$

where K is the convolution kernel size, C is the number of channels, n is the number of images segmented by the model, and time is the total time required for model segmentation.

IV. Experimental Results

This study uses stochastic gradient descent for training. After repeated optimization, the initial learning rate was set to 0.005 and the batch size was set to 8. The data size was 512 × 512, with a batch size of 4, and the number of training iterations is 100. The first 50 iterations were used for frozen training, during which the backbone feature extraction network was frozen to accelerate training, while the remaining 50 iterations were used for unfrozen training to fine-tune the parameters. The Adam optimizer was selected for this study as this optimizer uses the first-order momentum and the second-order momentum, which can dynamically adjust the learning rate and make the model converge faster.

A. Ablation Experiment

To validate the effectiveness of the various improvements made to the DeeplabV3+ model in enhancing the segmentation accuracy for navel orange defects, ablation studies were conducted by modifying the backbone network, adding the CBAM mechanism, and incorporating the CSPM mechanism. The experimental results of these segmentation effects are presented in Table III.

As shown in Table III, the DeeplabV3+ model with the HECA-MobileNetV3 backbone exhibits the lowest parameter count and the highest frame rate after improvements; after incorporating the CBAM attention mechanism into the basic model, MIoU increased by 3.73% and MPA increased by 2.47%. As CSPM attention mechanism is incorporated into the model, MIoU increased by 3.41% and MPA increased by 1.99%; and after adding both CBAM and CSPM mechanisms to the model, MIoU reached 89.50%, an increase of 8.01%, and MPA reached 94.02%, an increase of 3.86%. These results indicate that the improvements proposed in this study effectively enhance the accuracy and precision of navel orange defect segmentation.

B. Cross-Entropy Loss Function

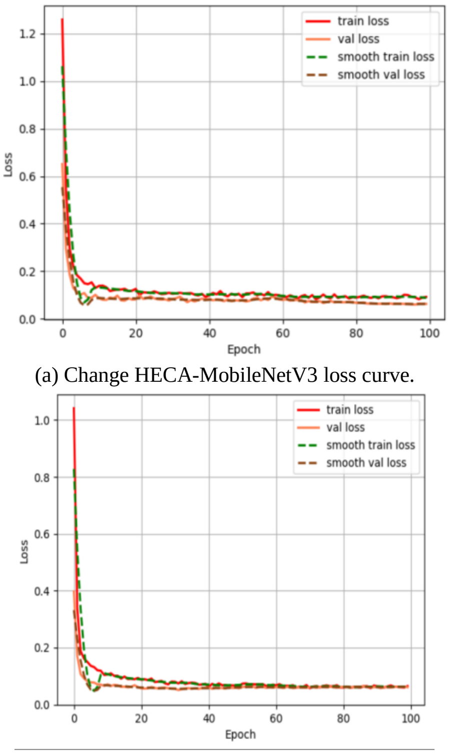
In per-class segmentation experiments, the cross-entropy loss function is often used to check each pixel one by one, and the predicted value is compared with the actual value to average the loss. Since the individual differences of some defective navel oranges are small, it is easy to cause semantic recognition errors of defects. Therefore, the multi-classification cross-entropy loss function is used to measure the segmentation effect. The loss curve for the DeeplabV3+ model only with the HECA-MobileNetV3 backbone is illustrated in Fig. 7(a), while the loss curve for the DeeplabV3+ model incorporating the attention mechanisms is shown in Fig. 7(b).

The horizontal axis represents the number of iterations, and the vertical axis represents the calculated loss value. The train loss represents the loss calculated during training; the val loss represents the loss calculated in the confirmation; the smooth train loss and smooth val loss represent the smooth loss values during training and verification respectively. It was found that the loss values of the two models gradually stabilize with the increase of the number of iterations during the training process. The Deeplabv3+ model improved by the fusion attention mechanism with the smallest loss value, stronger convergence and more stability.

TABLE III. SEGMENTATION ACCURACY OF DIFFERENT BACKBONE NETWORKS AND MODELS WITH ATTENTION MECHANISM

Model	MIoU/%	MPA/%	Params/M	FPS/fps
D+X	82.23	90.51	56.14	15.2
D+M	80.71	89.27	6.61	72.3
D+H	81.49	90.16	**5.83**	**76..5**
D+H+CB	85.22	92.63	6.58	74.1
D+H+CS	84.95	92.15	6.47	74.9
D+H+CB+CS	**89.50**	**94.02**	6.92	70.8

D：DeeplabV3+，X：Xception，M：MobileNetV3，H：HECA-MobileNetV3，CB：CBAM，CS：CSPM



(a) Change HECA-MobileNetV3 loss curve.

(b) Fusion attention mechanism loss curve.

Fig. 7. Loss curve of multi-classification cross entropy loss function.

C. Comparative Experiments

To ensure the validity of the research, several classical networks were applied to this navel orange dataset for comparative experiments. These experiments aim to verify the effectiveness of the improved DeeplabV3+ model incorporating the CBAM attention mechanism in segmenting navel oranges defects. The experimental results are shown in Table IV.

As shown in Table IV, the DeeplabV3+ model with MobileNetV3 as the backbone network has the fewest parameters and the highest frame rate, but lower MIoU and MPA. The improved DeeplabV3+ model significantly outperforms the other five models in terms of MIoU and MPA. Specifically, the MIoU of the improved DeeplabV3+ model is 21.74%, 16.06%, 13.01%, 7.27%, and 8.79% higher than those of Unet, SegNet, PSPNet, DeeplabV3+ with Xception backbone, and DeeplabV3+ with MobileNetV3 backbone, respectively. In terms of MPA, it is 29.31%, 14.35%, 7.57%, 3.51%, and 4.75% higher, respectively. The improved DeeplabV3+ model reduces the number of parameters by 49.42MB and increases the frame rate by 55.6fps compared to the DeeplabV3+ model with Xception backbone. Although its parameter count and frame rate are slightly lower than those of the DeeplabV3+ model with MobileNetV3 backbone, its MIoU and MPA are significantly higher than those of the unimproved DeeplabV3+ model.

To provide a more intuitive comparison of the segmentation performance of different models, this study selected five representative images for visual contrast. The effectiveness of the improved DeeplabV3+ model incorporating the attention mechanism is compared with Unet, SegNet, PSPNet, DeeplabV3+ with Xception backbone, and DeeplabV3+ with MobileNetV3 backbone in navel orange defect segmentation, as shown in Fig. 8.

The result shows the improved DeeplabV3+ model provides the clearest segmentation boundaries and accurately identifies small defects in navel oranges, achieving higher detection performance compared to other models. The unimproved DeeplabV3+ model performs better than the other networks but still exhibits some issues with fuzzy boundary segmentation and misidentification of small targets. Unet, SegNet, and PSPNet networks also suffer from varying degrees of recognition errors and fuzzy boundary segmentation.

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS WITH THE CLASSIC NETWORK

Model	Backbone network	MIoU/%	MPA/%	Params/MB	FPS/fps
U-Net	ResNet50	67.76	64.71	23.9	38.9
SegNet	VGG16	73.44	79.67	21.8	45.8
PSP-Net	ResNet101	76.49	86.45	27.6	42.3
DeeplabV3+	Xception	82.23	90.51	56.14	15.2
DeeplabV3+	MobileNetV3	80.71	89.27	**6.61**	**72.3**
Improved DeeplabV3+	HECA-MobileNetV3	**89.50**	**94.02**	6.72	70.8

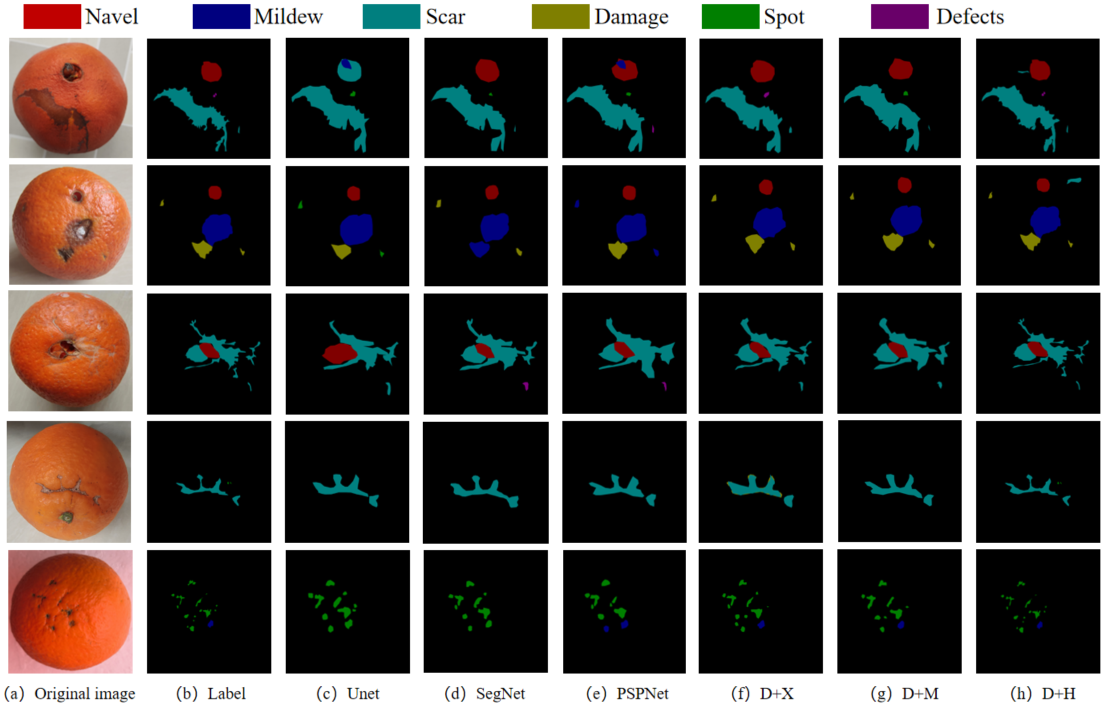


Fig. 8. Effects of navel orange defect segmentation detection using different models.

V. Discussion

Unlike other improved DeeplabV3+ methods, this study does not simply replace the complex Xception network with a lightweight backbone network MobileNetV3, but uses the improved H-ECA mechanism to replace the SE attention mechanism in the MobileNetV3+ structure. The Hard-sigmoid activation function is applied to the ECA structure, which can effectively improve the model calculation efficiency and improve the gradient disappearance problem. ECA is an improved structure based on the SE mechanism. The Hard-sigmoid activation function is combined with ECA and applied to the MobileNetV3 structure, which not only makes the backbone network lighter and ensures the real-time detection, but also effectively improves the ability to extract image features. At the same time, the CBAM and CSPM attention mechanisms are flexibly integrated into the shallow and deep feature extraction networks of DeeplabV3+, and the weights of the feature maps are adaptively adjusted in the two dimensions of channel and space, which improves the sensitivity to navel orange defects, focuses more on high-level semantic information, captures key information that is easily lost in model up and down sampling, integrates global and local features, improves the model's feature representation and generalization capabilities, and achieves more accurate semantic segmentation. Experimental results show that DeeplabV3+ with integrated attention mechanism has faster segmentation speed and higher accuracy.

The improved DeeplabV3+ model integrates attention mechanisms at multiple levels, which plays an important role in the convolution of each layer of the model. It has good segmentation performance for multiple categories of defects on the surface of navel oranges, but for some defects without obvious boundaries and light colors, the segmentation performance of this model is not as good as other defects. In future research, the DeeplabV3+ model will be further improved, such as using a more powerful backbone network, introducing a residual structure, improving the loss function, applying adaptive feature pyramid technology, and proposing a new model structure, which will be applied to navel orange defect segmentation in order to obtain better results, which can further applied to other fruit and vegetable defect detection and other image segmentation fields.

VI. Conclusion

In this paper, an improved new semantic segmentation model DeeplabV3+ model is proposed that incorporates an attention mechanism to solve the problems of low recognition accuracy and slow detection speed of similar defects and small targets in the navel orange defect grading and sorting task. By employing the improved HECA-MobileNetV3 backbone network, the model reduces parameters and enhances real-time detection. The CBAM mechanism is integrated into the ASPP structure and an additional CSPM mechanism is introduced to improve distinguishing and recognition capabilities for similar defect features. Furthermore, CBAM is incorporated into the low-level feature extraction structure to enhance segmentation of small target boundary features. Comparative study with DeeplabV3+ model was conducted resulting in improvement of MIoU of 89.50% and MPA of 94.02%, while reducing parameters by 49.42M and increasing detection speed by 55.6fps. In comparison to other semantic segmentation networks, the proposed model achieves higher detection accuracy and segmentation effectiveness while maintaining advantages in parameter efficiency and speed. The algorithm presented in this paper effectively meets the precision and speed compatibility requirements for navel orange defect grading and sorting in industrial applications.

Acknowledgment

This work was supported by the Natural Science Foundation of Jiangxi Province (Grant No. 20224BAB214056), The Science and Technology Project of Jiangxi Province (Grant No. GJJ214901) and Universiti Teknologi MARA.

References

[1] P. Zhou, J. Wei, T. Zhong and H. Zheng, "The Research on Navel Oranges Detection Systems of Harvesting Robots Based on an Improved YOLOv5," 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), Tianjin, China, 2023, pp. 537-542.

[2] H. Javadikia, S. Sabzi, and H. Rabbani, "Machine vision based expert system to estimate orange mass of three varieties," International journal of agricultural and biological engineering, vol. 10, no. 2, pp. 132–139, Mar. 2017.

[3] D. Rong, Y. Ying, and X. Rao, "Embedded vision detection of defective orange by fast adaptive lightness correction algorithm," Computers and Electronics in Agriculture, vol. 138, pp. 48–59, Jun. 2017.

[4] M. Zhang, T. Wang, P. Li, and Y. Zheng. "Surface defect detection of navel orange based on region adaptive brightness correction algorithm," Chinese Agricultural Science, vol. 52, no. 2, pp. 2360-2370, Jan. 2019.

[5] W. Luo, G. Fan, P. Tian, W. Dong, H. Zhang, B. Zhan. "Spectrum classification of citrus tissues infected by fungi and multispectral image identification of early rotten oranges". Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 279, 121412, Oct. 2022.

[6] S. Md. Iqbal, A. Gopal, P. E. Sankaranarayanan, and A. B. Nair, "Classification of Selected Citrus Fruits Based on Color Using Machine Vision System," International Journal of Food Properties, vol. 19, no. 2, pp. 272–288, May 2015.

[7] D. M. Asriny, S. Rani, and A. F. Hidayatullah. "Orange Fruit lmages Classification using Convolutional Neural Networks," IOP Conference Series Materials Science and Engineering, Vol. 803. No. 1, pp. 12-20, Apr. 2020.

[8] X. Cai, Y. Zhu, S. Liu,and Y. Xu. "FastSegFormer: A knowledge distillation-based method for real-time semantic segmentation of surface defects in navel oranges,Computers and Electronics in Agriculture - X-MOL," X-mol.com, 2024.

[9] A. Z. Da Costa, H. E. H. Figueroa, and J. A. Fracarolli. "Computer vision based detection of external defects on tomatoes using deep learning," Biosystems Engineering, pp. 131-144, Feb. 2020.

[10] X. Liang et al., "Real-Time Grading of Defect Apples Using Semantic Segmentation Combination with a Pruned YOLO V4 Network," vol. 11, no. 19, pp. 3150–3150, Oct. 2022.

[11] J. Hao, Y. Zeng, X. Wang,et al."Research on kiwifruit feature extraction and automatic grading based on DeeplabV3+," Agricultural Machinery and Agronomy, vol. 55, no. 03, pp.49-54, Feb. 2024.

[12] W. Gu, J. Wei, Y. Yin, X. Liu, and C. Ding. "Multi-category segmentation method of tomato images based on improved Deeplabv3+," Transactions of the Chinese Society of Agricultural Machinery, vol. 54, no. 12, pp.261-271, 2023.

[13] S. Fan et al., "Real-time defects detection for apple sorting using NIR cameras with pruning-based YOLOV4 network," Computers and Electronics in Agriculture, vol. 193, p. 106715, Feb. 2022.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," arXiv.org, 2014.

[15] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 1 April 2018.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arxiv.org, Jun. 2017, Available: https://arxiv.org/abs/1706.05587.

[17] H. Peng, S. Xiang, M. Chen, H. Li and Q. Su, "DCN-Deeplabv3+: A Novel Road Segmentation Algorithm Based on Improved Deeplabv3+," in IEEE Access, vol. 12, pp. 87397-87406, 2024.

[18] R. Liu and D. He, "Semantic Segmentation Based on Deeplabv3+ and Attention Mechanism," 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2021, pp. 255-259,

[19] T. Zhang, R. Zhou, L. Zhang and M. Liang, "Research on Multi-scale Feature Fusion Method for Target Detection Based on IN-FPN," 2023 8th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 2023, pp. 94-98.

[20] T. Lei et al., "Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–14, Jan. 2023.

[21] A. Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1314-1324.

[22] S. Qian, C. Ning and Y. Hu, "MobileNetV3 for Image Classification," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 490-497.

[23] H. Yang, L. Lin, S. Zhong, F. Guo and Z. Cui, "Aero Engines Fault Diagnosis Method Based on Convolutional Neural Network Using Multiple Attention Mechanism," 2021 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Weihai, China, 2021, pp. 13-18.

[24] P. Nirale and M. Madankar, "Analytical Study on IoT and Machine Learning based Grading and Sorting System for Fruits," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-6.

[25] L. Yu et al., "A Lightweight Complex-Valued DeepLabv3+ for Semantic Segmentation of PolSAR Image," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 930-943, 2022.

A Framework for Capturing Quality Requirements by Integrating the Requirement Engineering Elements in Agile Software Development Methods

Yuli Fitrisia[1], Rosziati Ibrahim[2]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia[1, 2]
Faculty of Computer Engineering Technology, Politeknik Caltex Riau, Pekanbaru, Indonesia[1]

***Abstract*—The early phase of Agile Software Development (ASD) methods is Requirement Engineering (RE). Quality Requirement (QR) is a type of RE that needs to be captured at the initial development phase to reduce rework, time, and maintenance costs. However, QR is one of the issues mentioned in ASD, namely the need for more capability to elicit, analyze, document, and manage QR. Therefore, this research aims to propose a framework for capturing QR to address QR issues in ASD by integrating RE elements, namely the RE phases, Documentation, Roles, and RE techniques. This research was conducted in four phases: 1) undertaking a theoretical study, 2) conducting an exploratory study to identify the current practices and issues to capture QR in ASD, 3) constructing the framework by using the RE elements, and 4) evaluating the framework by conducting ASD practitioners' view using questionnaires. The questionnaires were then analyzed using descriptive statistics based on the average mean of each element. The result shows the average mean for all elements (4.25), the average mean of each element for the RE phases (4.36), the documentations (4.11), the roles (4.25), and the RE techniques (4.18). The mean distribution of each element is more than 4 out of 5 indicating that the framework to capture QR is verified. Thus, this framework can be used by ASD practitioners as a guideline to capture QR in ASD methods.**

***Keywords*—Quality requirement; requirement engineering; ASD; framework; ASD practitioners**

I. Introduction

In recent years, software practitioners have used many software development methods. Each software development method has its advantages and disadvantages. For example, traditional software development methods are suitable for requirements that are clearly defined from the initial phase of the Software Development Life Cycle (SDLC) [1]. However, these methods have several issues to be aware of, namely limitations in accommodating requirement change during development, lack of interaction with customer, and the client only seeing the product at the end of the project [2]. These issues are covered by Agile Software Development (ASD) methods [2-4].

ASD is a popular software development method widely used in today's business industry [5-8]. ASD methods have advantages in producing software products faster to markets, being flexible to changing requirements, and increasing customer collaboration [9-10]. In addition, based on the Agile annual report [11], ASD increases collaboration, facilitated teamwork, provided better alignment for business needs, better work environment, and better visibility capabilities in application development.

Requirement Engineering (RE) is the early stage of the Software Development Process, including ASD. It has two types of requirements, namely Functional Requirements (FR) and Quality Requirements (QR), also known as Non-Functional Requirements (NFR). QR presents the quality of the software product. However, the tendency to ignore QR, the product quality can be negatively impacted and even cause software product failure [9-11]. Therefore, it is important to capture QR early in the RE phase of the ASD to reduce time, rework, and maintenance cost.

ASD practitioners also realize the importance of documenting QR [15]. It can help with easy analyses and traceability, gives a ready-to-use template, and eases communication among ASD stakeholders. Roles are people who possess responsibilities for every process in software development. Assigning roles with their expertise helps managing QR and making the QR list consistent and unambiguous [16]. Selecting proper techniques for each RE phase is also important to ensure the phase is well-conducted and reduce the complexity of capturing QR.

ASD has issues in terms of lack of capability to handle QR [9,14-18]. For instance, eliciting and analyzing QR challenges, limited techniques for eliciting, modeling, and linking QRs with functional requirements, and inadequate user stories for specifying QR [12]. The other issues are the lack of QR documentation [9,17], and no explicit practice for QR [21]. Additionally, it also lacks ASD capabilities in managing QR through ASD artifacts, namely user stories, prioritizing functional requirements (FR), and the tendency to ignore QR [19-22]. Based on these previous issues, practitioners realize the importance of capturing QR early in the context of ASD [25]. It is also important for capturing QR systematically, which offers how to capture and validate QR [15,23,24,27]. It is also supported by the result of a survey conducted by López et al. [28] reporting that 50% of practitioners follow QR processes in a systematic, well-defined, or ad-hoc process. It can reduce the complexity of managing QR in ASD and improve product quality and customer satisfaction.

Based on literature studies conducted in this research, identified issues related to capture QR in ASD methods can be solved by implementing approaches, frameworks, and guidelines for QR in ASD. Furthermore, several studies have been conducted to capture QR. The QR elicitation guideline proposed by Younas et al. [29] was implemented to tackle QR in ASD methods at the elicitation phase by using several pieces of documentations, namely the QR glossary, historical data, and checklist table between FR and QR. It also involved the developer team, experts, and customers. A study by Younas et al. [30] proposed the Elicitation of Non-Functional Requirements in ASD using a Cloud Computing Environment. This methodology also involves the same RE phase. The documentations used were project history, template for NFR by Kopczy and system type. Then, the roles involved developer team, customers, and experts. According to Jarzębowicz et al. [31] proposed elicitation practices to capture QR in ASD. The roles involved developer team, customers, products owners, and experts. Another study by Behutiye et al. [32] proposed the QR documentation guideline within ASD. In this guideline, they only proposed documentation for QR. Then, Alhaizaey et al. [33] proposed a framework for Reviewing and Improving Non-Functional Requirements in ASD-based Requirements. This framework was implemented at the elicitation, analysis, and validation phase. The documentations used were FR user story, identified NFR, NFR bibliographic source, glossary and standards, and NFR description. It also involved the developer team and customers. Finally, Sherif et al. [24] proposed a framework to manage NFR in ASD. It was implemented at the elicitation, analysis, documentation, and validation phase. The documentations used were FR user story, checklist-based reading template, system type, domain type, project history, mapping sheet between QR, and QR user story. It also involved the developer team, product owners, and customers.

There are still several gaps in proposed previous solutions to capture QR. It shows that three [29] of six of the previous works implemented to capture QR in the elicitation phase only. However, based on the earlier discussion, it is important to capture the QR to construct the RE phase element by using complete phases, namely elicitation, analysis, documentation, and validation, to produce QR properly. Then, the documentation used in the previous study only focuses on partial documentation for each RE phase [33]. Additionally, several studies did not use a software quality model or standard as a reference to identify the QR [24]. However, it is needed because the quality model or standard has the metrics to measure the QR. It can be used to make sure the QR is measurable. One study focuses on QR documentation only [32]. It should need comprehensive documentation by constructing the documentation for each RE phase to give an easy analysis and get a clear QR list.

The other element is the roles involved. Several previous studies report that the roles did not focus on their expertise and involved the roles needed related to each RE phase [24,33]. The roles should be constructed to be suitable for their expertise and responsibility. As a result, it gives an impact on the process of capturing QR more effectively. In addition, several studies do not mention the RE technique used clearly for each RE phase [29, 31]. Furthermore, it also needs to define the RE technique that will be used in implementing the RE phase systematically.

This paper presents research focuses on addressing three issues according to the gap findings, namely (i) lack of clarity on which the Requirement Engineering phase and the techniques used of the capture QR should be implemented, (ii) lack of documentation of QR, and (iii) the inadequate ability of user stories to capture QR by involving roles to address QR effectively. Based on these three issues, the following Research Questions (RQs) are formed. They are RQ1: how to determine the RE elements and each component in ASD methods, RQ2: how to design the framework for capturing QR using the elements and components of RQ1 and RQ3: how to evaluate the QR framework by using practitioner reviews. Thus, the objectives of this research are as follows: (i) to determine the RE elements and each component in the ASD method, (ii) to design a framework for capturing QR, and (iii) to evaluate the QR framework by using practitioner reviews.

The research significantly contributes to the body of knowledge in the field of Software Engineering, particularly in Requirement Engineering, on how QR is being addressed, captured, and documented in an ASD environment. Moreover, ASD practitioners can use this proposed framework as one of the best practices and guidelines for handling QR in an ASD environment, and the Stakeholders in an ASD environment can gain QR status transparency and track the QR. Then, it systematically captures the QR to reduce the complexity of managing QR in ASD, improve product quality, and finally improve customers' satisfaction.

This paper is outlined as follows: Section II presents an overview of related work and the gaps in previous work, Section III presents the research methodology, including theoretical study, exploratory study, framework design, and framework evaluation, Section IV explains the experimental results and analysis, and Section V presents the conclusion and future work.

II. Related Work

Quality Requirement (QR) is one of the RE types and is an important artifact that plays a crucial role in software project success [34, 35]. It can lead to increased costs or longer time-to-market due to the failure to meet QR needs properly [25,26]. Many software quality standards can be adopted in QR for software development. In the 1970s and 1980s, the authors in [36,37] proposed their own QR taxonomies. The ISO/IEC 25010 standard is the most widespread method of defining, categorizing, and managing QR. This standard is widely adopted in today's industry. It has nine quality categories, namely functional suitability, performance efficiency, compatibility, interaction capability, reliability, security, maintainability, flexibility, and safety [38]. Numerous recent studies have conducted extensive reviews on QR in the Requirement Engineering in ASD methods [13]. A study in [29], Younas et al. proposed activities of NFR Elicit guideline as depicted in Fig. 1.

According to Fig. 1, the first phase is the preliminary requirement to collect FRs that have been identified. The next step is to identify the software type and then to identify the QR from Glossary. It involves the historical data to make predictions about new QR. This elicitation guideline encompasses experts' involvement. This process ends when the experts and users finalize it. This guideline also uses a checklist table between FR and QR to manage changing requirements. In contrast, the limitation of this guideline is the mapping between FR and QR, which is done at the end of the process, for it can take time if there is a change, and the process should be repeated from the beginning. The other limitation is that there is no mapping between QRs to check conflict between them. Then, this guideline is only implemented in the elicitation phase, and there is no clear explanation of how to finalize the QR list between the experts and the users.

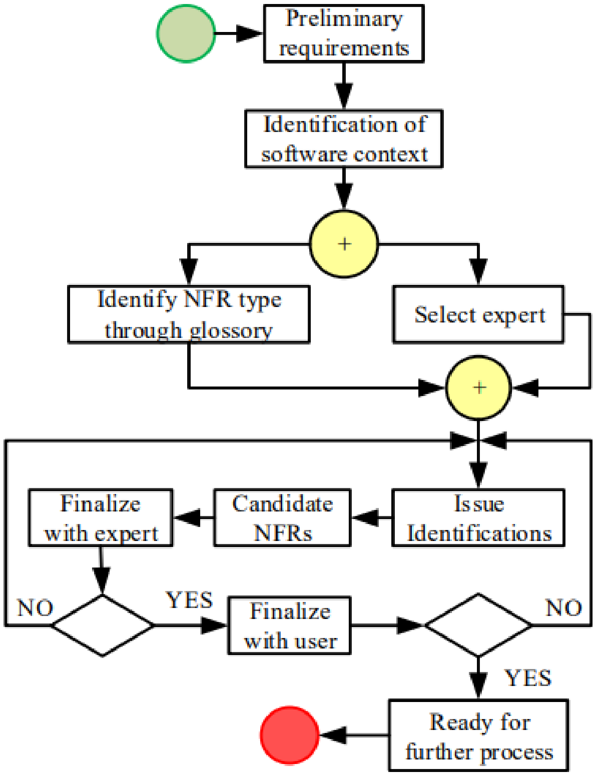


Fig. 1. QR elicitation guideline [29].

Similarly, in a study [30], Younas et al. proposed the Elicitation of Non-Functional Requirements in ASD using a Cloud Computing Environment. This proposed methodology is a continuation of their previous research on the QR elicitation guideline [29] as described in Fig. 2. From Fig. 2, it can be seen the differences of this present study from previous studies are the use of Natural Language Processing (NLP) for QR extraction in the elicitation phase and cloud computing tools for sharing data and communication. However, this methodology has no link between QR to check the conflict, and it is only implemented in the elicitation phase.

Jarzębowicz et al. [31], discussed elicitation practices to capture QR in ASD. The methodologies used in their study were a systematic literature review (SLR) and interviews with ten ASD practitioners. Based on their findings in the SLR study, techniques for elicitation include Customer-Developer meetings, Brainstorming, QR document circulation between Product Owner and QR Stakeholders (e.g., experts), QR catalog, and on the basis of Business Process Models are popular in several studies. The other findings based on interviews with ASD practitioners summarize that ASD practitioners mention that the presence of an analyst role contributes to more thorough QR elicitation. Then, for sources of requirements, several ASD practitioners argue that the Product Owners are capable of giving opinions on QR. In other cases, multiple stakeholders can consider eliciting requirements from their point of view, including other IT systems and document standards. The other sources are software developers and technical experts who will provide inputs based on their expertise. Then, the techniques used for QR elicitation are interviews, and workshops (including brainstorming and other kinds of group work) that are commonly used.

There are efforts to capture QR using the current practice. For instance, the elicit QR is non-verifiable and non-measurable, and not all QR needed in the software development is detected. For example, usability is recognized, but others still need to be defined. Therefore, based on the ASD practitioner's view, they state that it needs an active approach and guidance for all ASD stakeholders.

Another study done by Behutiye et al. [32] proposed the QR documentation guideline within ASD. This guideline, known as ASD QR-Doc, offers 12 recommendations to facilitate the documentation of QR in ASD. The guideline was validated by ASD professionals. They consider the ASD QR-Doc guideline to be straightforward for use in ASD, helpful in the early process and documentation of QR, and not obstructive to the ASD process. However, the guideline is only validated by practitioners, and this guideline needs to be clearly defined when QR is documented in the RE phase.

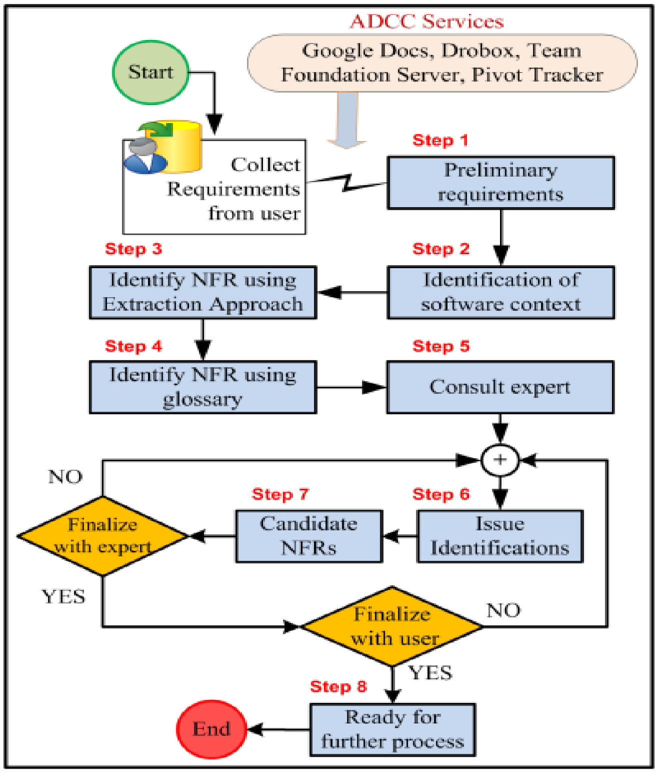

ADCC Services
Google Docs, Drobox, Team Foundation Server, Pivot Tracker
Start
Collect Requirements from user
Step 1
Preliminary requirements
Step 3
Step 2
Identify NFR using Extraction Approach
Identification of software context
Step 4
Step 5
Identify NFR using glossary
Consult expert
NO
Step 7
Step 6
Finalize with expert
Candidate NFRs
Issue Identifications
YES
Finalize with user
NO
Step 8
YES
End
Ready for further process

Fig. 2. QR elicitation methodology using cloud computing [30].

Alhaizaey et al. [33] proposed a framework for Reviewing and Improving Non-Functional Requirements in ASD-based Requirements. It uses NLP and Artificial Intelligence (AI) techniques to automate analyzing QR from user stories. Then, the artifact from this phase is reviewed and inspected for improving the QR. Three artifacts are produced from the previous phase, namely the processed user story, the identified QR, if any, and the description of the QR. The last phase is improving the requirements using four activities, namely rewriting a user story, rewriting [means] part of the user story, including QR as an Acceptance Criteria (AC), and including QR as a Definition of Done (DoD), as depicted in Fig. 3. However, this framework needs to explain how the documentation phase is conducted. It also only involves the developer team and clients to finalize the QR and needs to involve experts in validating the QR list.

Finally, Sherif et al. [24] proposed a framework to manage NFR in ASD called MANoR, which stands for Managing Agile Non-Functional Requirements. It provides two main stages and five main components. The stages are pre-analysis and post-analysis, and the components support various critical functions within requirements engineering, encompassing requirements elicitation, analysis, documentation, and validation. The main steps of the MANoR framework are depicted in Fig. 4. There are four areas for improvement and limitations associated with this approach. Firstly, there is no Quality Model from the glossary that can be used as a source to recommend QR. The quality model, particularly ISO 25010, has an advantage. For instance, it has the metrics to measure the QR lists. Secondly, the expert is not involved in the QR elicitation process, which can help identify QR more effectively. Thirdly, there is no mapping between FR and NFR during the QR analysis process to check for inconsistency. There is also no documentation of QR decisions that will help track them. Fourthly, the validation technique (reading technique) should be mentioned, and QR is only validated by the clients or customers.

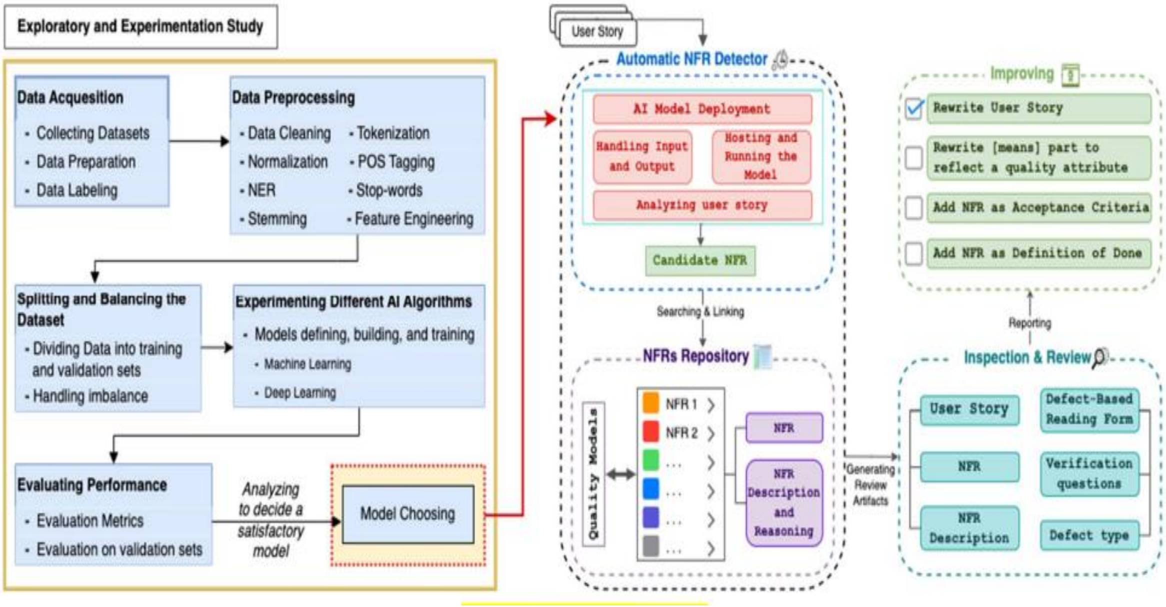


Fig. 3. Framework for elicitation, analysis, and reviewing QR [33].

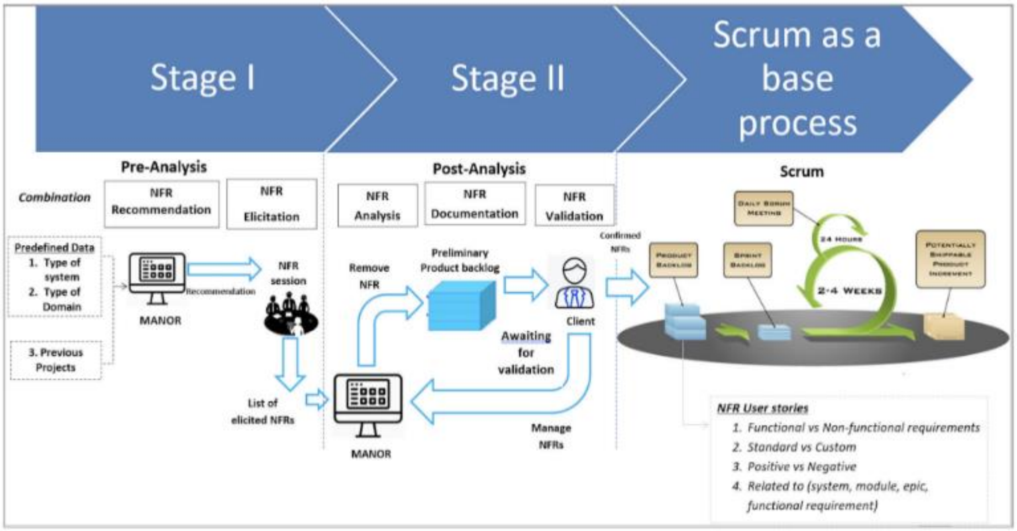


Fig. 4. MANoR framework [24].

In summary, based on related studies discussed, the solutions focus on capturing QR in ASD from different points of view, namely frameworks, approaches, models, and guidelines as previously mentioned. According to the gaps identified in related work, therefore, this research proposes a framework for capturing the QR by integrating the RE elements in ASD methods comprising 1) the RE phase, 2) documentation, 3) the roles involved, and 4) the RE technique.

III. Methodology

Fig. 5 shows that the research methodology consists of four phases. The first phase is theoretical study, the second phase is exploratory study, the third phase is framework design, and the fourth phase is framework evaluation. Each phase has key activities and outputs of the activities.

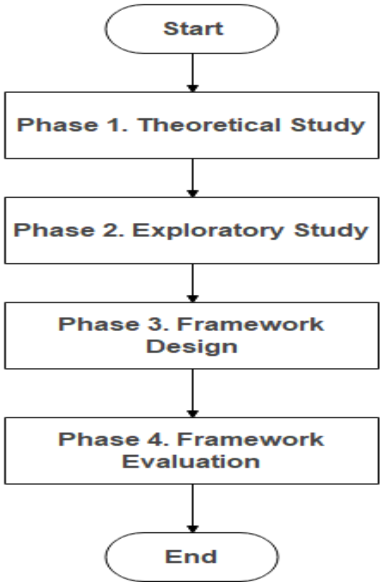

Start
Phase 1. Theoretical Study
Phase 2. Exploratory Study
Phase 3. Framework Design
Phase 4. Framework Evaluation
End

Fig. 5. Research methodology.

A. Theoretical Study

The first phase was conducted by reviewing the literature to explore the concept related to the study, current practices, and issues to capture QR in ASD, identifying the element constructed in the framework, and finding the gap by analyzing the related work in ASD from various sources like journals, conferences, books, and other sources. Then, based on the gap finding from existing related work, this study proposes a new solution for the elements and the components to construct for capturing QR in ASD. The outputs of this phase are a clear explanation of current practices and the issues of capturing QR in ASD. It also gets the elements and the components constructed in the framework.

B. Exploratory Study

The second phase was conducted by doing a qualitative study by interviewing ASD practitioners with a small number of respondents to get detailed information about their opinions [39]. The interview aims to explore current practices to capture QR in the ASD industry and the issues based on practitioners' views. The exploratory study was conducted by an online interview involving 30 ASD practitioners. The summary of current practices when capturing QR in ASD is categorized into three phases, namely RE phase, during development, and during product release to market or customers.

The result percentage indicates that the QR is identified majority during development and release to market is 33.33%. Then, during the RE phase and development is 26.67%, during development is 20%, in RE phase is 10%, and product release to market is 10%. The other findings are that QR is identified the majority during development is 80%. Then, during product release to market is 43.33%, and during the RE phase is 36.67%. On the other hand, it also shows the weaknesses of current practice in capturing QR based on interview findings, namely:

- Identified QR in the RE phase: More time is needed in the planning phase. Therefore, it takes time to start the development process. Only the developer team was involved in identifying the QR in the RE phase. Sometimes, people need to be made aware of QR. Consequently, they should have added one more task for improvement, but it took time.
- Identified QR during development: It needs more rework, cost, and time to adjust the QR because it changes most in software architecture, namely in design and code, adding a new story to fix the bug in Product Backlog. It is challenging to identify QR during development because the document is not updated based on requirement changes. Sometimes, the QR is ignored because there is no QR documentation. It can change the project timeline, and the workload also increases a lot.
- Identified QR after release to market or customers: It needs more rework because it changes most in software architecture, namely in design and code. The application appearing on the device does not meet the user's requirements. It needs more time, it can change the project timeline, product backlog items increase, and the workload also increases a lot.

C. Framework Design

The third phase is to construct the framework based on a theoretical study and exploratory study findings. The theoretical study is based on the literature review, and the exploratory study is based on interviews with ASD practitioners. The output of this phase is the proposed framework for capturing QR by integrating the RE element in ASD based on studying the existing model gaps outlined in related work. It preserves the strengths of these existing models and tries to overcome their limitations. It also constructs various aspects for comprehensive QR.

This present study argues that it is important to capture QR by implementing in the RE phases as the foundation of the software lifecycle, which are elicitation, analysis, documentation, and validation [24]. The QR was identified during an initial iteration of Agile Software Development and then refined in further iteration [36,37]. The authors in [31] also found that early identification of QR at the beginning of a software project is better. It also supports an exploratory study; if QR were defined during the RE phase, it would not require much effort for future tasks. When QR is identified early, it can help to produce QR properly, help to identify project effort, cost, and size, and reduce rework.

The documentation/artifact should need comprehensive documentation, usefulness, relevance, and understandability for supporting QR documentation and its impact on ASD practice [32]. The ASD manifesto focuses on the development of working software over comprehensive documentation [42]. However, a lack of QR documentation can cause misinterpretation and rework [31]. It also supports an exploratory study, which makes it challenging to identify QR during development because the document was not updated based on requirement changes, and sometimes the QR is ignored because there is no QR documentation. Therefore, it is important to include documentation as the element to capture QR. Constructing the documentation for each RE phase can give easy analysis, clear tasks, traceability, clear documentation, and well-documented QR [16]. It can help communication among ASD Stakeholders.

The other element is the roles involved. In several previous studies reviewed, it was found that the roles need to include capturing QR to focus on their expertise [31]. The roles for each RE phase need to be constructed in a way that is suitable for their expertise and responsibility [32]. It also supports an exploratory study result, because only the developer team is involved in identifying the QR, sometimes people are not aware of the QR. Consequently, they should have added one more task for improvement, which took more time. Furthermore, it is important to involve roles with their expertise for each RE phase because it can impact the process of capturing QR more effectively. According to Aljallabi et al. [16] stated that it also provides proper QR results with more reliable results due to invented different points of view.

It also needs to define the RE technique that is used to capture QR, which is in line with ASD practice on direct communication [31]. RE techniques were needed to conduct the RE phase systematically by the team [1]. If we can choose the right techniques, it can be produced and conducted to capture QR more effectively and clearly. It also helps to capture the impact of change of requirement, which is understandable by all stakeholders, and check for errors and inconsistencies. Therefore, this research proposes a framework to integrate the RE elements, namely, 1) the RE phase, 2) documentation/artifact, 3) the roles involved, and 4) the RE technique. These four elements are needed to capture QR in ASD. The following sub-sections define the RE element and its relation to the component of each element.

1) Constructing RE phase element: The RE phase was constructed by using the four-phase component of RE as the first element, namely elicitation, analysis, documentation, and validation [1]. It is important to capture QR by using complete phases to produce QR properly [24]. It is also supported by a study [33], who argue that validating the QR is a crucial requirement process as the last phase of RE. Therefore, it should be conducted in all phases of RE to capture the QR.

The elicitation phase is the first phase of RE that aims to understand the tasks performed by stakeholders and how a new system could support their tasks [1]. This phase is the foundation of project success and aims to explain QR to the stakeholders. This session also determines the QR based on the element used. The output of this phase is the list of elicit QR. The second phase is the analysis phase aiming to find consistency between FR and QR [16]. It is also to make sure there is no conflict between QRs [24].

The purpose of the documentation phase as the third phase is to write down software requirements into a Software Requirements Specification (SRS) document [1]. It can be used to document user requirements and system requirements, namely the FR and QR in the user story written in the product backlog. It can also be used to document the decision on the QR. The last phase is the validation phase, which aims to check the requirements meet the customers' expectations [1]. The checking process consists of a validity check, consistency check, completeness check, realism check, and verifiability check. It is also to ensure that the QR list is clearly defined, that there is no error interpretation, to check areas where clarification may be required, and that there is no missing information. According to Sherif et al. [24], this phase also aims to reach an agreement among stakeholders regarding QR on the same view for the software being developed.

2) Constructing QR documentation element: The documentation element used in this framework can provide a ready-to-use template for easy analysis and traceability. It was constructed at each RE phase by integrating documentation components as a reference for capturing the QR. QR should be documented along with FR [43].

System-type document in elicitation helps users identify relevant QR based on different types of systems [40,41]. Domain-type document in elicitation helps users identify relevant QR based on various application domains [44]. ISO 25010 Quality Model in elicitation has advantages, namely providing a more detailed QR and metrics on how to measure the QR [29,30]. Project history document in elicitation is useful to define the QR for the next project based on historical data. [24]. A mapping sheet between FR and QR document is used in the analysis to check the consistency between FR and QR [16]. A mapping sheet between QR document is used in the analysis to check the conflict between QRs [24].

In this research, a separate user story is used to document QR, which consists of the FR user story and the QR user story. This helps to manage QR during the development process, for example, during the sprint [45]. The functional user story in the documentation aims to document functional requirements in ASD. QR user story in documentation to document the Quality Requirements list in ASD. QR decision in documentation as a history to decide the QR [32]. According to Sabaliauskaite et al. [46] Checklist-based reading document in validation is used to check the properties of documents and what problems or defects should be identified based on the list of questions.

3) Constructing the involvement of roles element: The roles constructed into the element of assigning roles with their expertise are helpful for managing QR and making the QR list consistent and unambiguous. Then, different points of view for the validation phase are involved in working together to produce QR, which results in more reliability and understanding for all stakeholders. It is constructed at each RE phase by integrating role components to capture the QR.

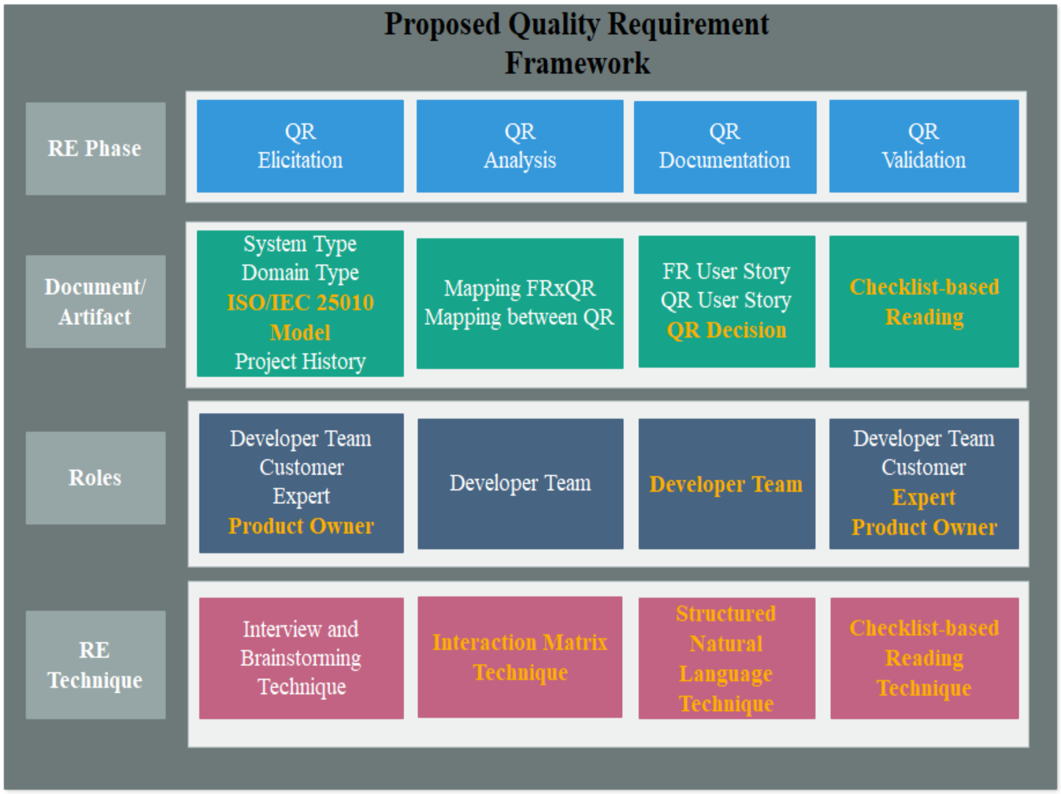
The developer team in elicitation is a person who understands the technical side related to QR that is elicited and is responsible for finishing the QR item in the product backlog [24]. Customers in elicitation are the stakeholders who own the system and need to be explained about QR, which is elicited [29]. An expert in elicitation helps elicit the QR; that is, someone who has more knowledge of QR and is concerned with the fulfillment of QR [29]. The Product Owner in elicitation has the responsibility of managing and optimizing the product backlog to ensure the product value is maximized which aligns with the FR and QR that should be elicited [47]. According to Jarzębowicz et al. [31] QR.

The developer team, in analysis as a technical side, defines consistency and conflict in the mapping sheet according to a clear justification based on the developer's knowledge [16]. The developer team in documentation manages the document if there is a change in these documents. The developer team in validation is a technical team that finishes the QR in the Product Backlog and ensures the QR can be tested. The expert in validation can help refine the QR and interdependencies among QRs and validate the QR. The customer is also involved in validation to make sure the QR list is understandable to the customer. The Product Owner, in validation, ensures the QR list is valid and included in the Product Backlog.

4) Constructing the RE technique element: The RE technique presents how the RE phase was conducted. It was constructed at each RE phase by integrating the components of the RE techniques to capture the QR. The Interview and Brainstorming techniques in elicitation are commonly used and popular techniques in the elicitation phase [31]. The interview aims to discover information and to understand the system to be developed based on asking questions to the stakeholders [1]. Additionally, a study [48] stated that Brainstorming aims to gather information in many creative ways by conducting work group meetings involving roles.

The interaction matrix technique is used in the analysis of two-dimensional requirements to assess the inconsistency between FR vs. QR and the conflict between QR where each requirement is compared to the other [49]. The Structured Natural Language technique in documentation is the documentation technique for writing the FR and QR using natural language on a standard form or template where each field provides information on the requirements [1]. The checklist-based reading (CBR) technique in validation aims to detect defects in the requirements based on a list of questions [49].

There are four major elements constructed in the proposed framework to capture QR, namely: 1) RE Phase, 2) Documentation/Artifact, 3) Roles Involved, and 4) RE technique. Each element has its components for the proposed framework, as depicted in Fig. 6.


Proposed Quality Requirement Framework

RE Phase	QR Elicitation	QR Analysis	QR Documentation	QR Validation
Document/ Artifact	System Type Domain Type **ISO/IEC 25010 Model** Project History	Mapping FRxQR Mapping between QR	FR User Story QR User Story **QR Decision**	**Checklist-based Reading**
Roles	Developer Team Customer Expert **Product Owner**	Developer Team	**Developer Team**	Developer Team Customer **Expert** **Product Owner**
RE Technique	Interview and Brainstorming Technique	**Interaction Matrix Technique**	**Structured Natural Language Technique**	**Checklist-based Reading Technique**

Fig. 6. The proposed QR framework.

D. Framework Evaluation

A practitioner evaluated this framework to verify it. This phase was conducted using quantitative research through questionnaires. The phases are instrument design, instrument validation, pilot study, sampling, data collection, and feedback analyses [50]. The questionnaires consist of two sections: demographic information and elements and the components of the QR framework. These instruments are used to verify the proposed framework to confirm whether the ASD practitioners' review agrees or disagrees with the proposed elements and the components of the framework to capture QR in ASD [50], [51]. The result of the framework evaluation is explained in the following Section.

IV. RESULTS

This section explains the findings of the evaluation by using questionnaires for ASD practitioners classified into sub-sections: A) the Pilot Study, B) Data Collection and Sampling, C) Data Analysis, and D) the Result.

A. Pilot Study

A pilot study was conducted by involving 30 ASD practitioners to check the reliability of the instrument and get feedback for the questionnaires before conducting the sampling survey. The pilot study result shows that it achieves the reliability threshold (> 0.7) based on Cronbach's alpha coefficient [52]. It consists of all elements (0.81), RE phase (0.72), document (0.91), roles (0.90), and RE technique (0.82), which means that the questionnaires are acceptable. The other result is questionnaire feedback from the respondents. The questionnaires were refined to improve the quality and avoid misinterpretation of the questions.

B. Data Collection and Sampling

The respondents of this questionnaire survey are ASD practitioners gathered using Snowball Sampling as one of the non-probability sampling techniques. Then, the questionnaires were distributed using an online survey from various channels, namely alumnae and their networking, colleagues in the industry, and post questionnaires in ResearchGate, LinkedIn Group for ASD, and ASD Community Indonesia. The total number of respondents who filled out the questionnaire was 170 people for one month, as presented in Table I.

TABLE I. QUESTIONNAIRE RESPONSE RATE

Description	Frequency	Percentage
Total questionnaires received	170	100%
Total rejected questionnaires	12	7.06%
Total usable questionnaires	158	92.94%

Table I shows that the total number of questionnaires received was 170, only 12 (7.06%) questionnaires were rejected due to outliers, and 158 (92.94%) questionnaires were used in this research.

C. Data Analysis

The data were analyzed using descriptive statistics. Data preparation aims to ensure that the data are free from errors by processing the cleansing data. It is important to make sure the result is accurate. According to Ibrahim et al. [50], the steps for data preparation start with screening the data and coding data, as well as checking missing data, suspicious response rates, outliers, and normality by using SPSS software.

1) Screening and coding data: This step was done by ensuring the data types were numeric and changing the measure data to scale. It was done at the variable view by changing the code manually.

2) Checking for missing data: This step aims to analyze the data that the respondents fill out. If there is missing data, it can cause an error. Based on the data, no missing values for all items were found. According to Hair et al. [53], if there are missing values, responses can be excluded by less than 10%, and the result is acceptable.

3) Suspicious response rate: This step is to identify the answer pattern if the respondents fill in the same values for all questions [53]. It can be excluded from the data that needs to be analyzed. The result is that no suspicious response rate was found.

4) Outliers: When using parametric or non-parametric tests, outliers can state the error rates and substantial distortions of parameter and statistic estimates [54]. Standardized Z-scores can be used to analyze the outliers based on the variables, and the values are then examined. The acceptable value of the Z-score is between -3.29 and +3.29, which indicates no outliers [55]. According to the Z-scores analysis, 12 outliers were found in this dataset, and it was removed from the dataset.

5) Normality: Before the data can be analyzed, it must meet the normal distribution, where each construct item must meet the normality [56]. The values of skewness and kurtosis can estimate the symmetry and data distribution. The authors in [53] stated that the value of the standard error of skewness close to zero is acceptable. While the authors in [56] stated that the acceptable value of the standard error of kurtosis should not exceed 10. The result shows that the standard error of skewness is 0.19, and the standard error of kurtosis is 0.38. This means that the data are close to the normal distribution and can be used to analyze the dataset.

D. Results

This section explains the results of the framework evaluation using the questionnaires by ASD practitioners. It consists of two sub-sections, namely demographic information and descriptive statistics.

1) Demographic information: This section describes the respondent's background and the organizational background.

a) Respondents' background: This section indicates their position in the organization and years of experience in ASD. According to respondents' positions, programmers are the majority of the respondents, about 46.20%. The second position is System Analyst, about 13.92%, and Quality Assurance (QA)/Tester, about 11.39%, followed by Team Leader, about 10.13% and Project Manager, about 6.33%, Product Owner, about 5.06%, Scrum Master, about 1.90%, and Others about 5.06%. The respondents' experiences with Agile Software Development methods are depicted in Table II.

TABLE II. EXPERIENCE IN AGILE SOFTWARE DEVELOPMENT METHODS

Positions	<1 Year	1-5 Years	6-10 Years	>10 Years	Total
Programmer	2	65	6	0	73
System Analyst	1	19	2	0	22
Quality Assurance/Tester	0	18	0	0	18
Team Leader	1	13	2	0	16
Project Manager	1	3	5	1	10
Product Owner	0	6	1	1	8
Scrum Master	0	1	2	0	3
Others	1	7	0	0	8
Total	6	132	18	2	158

Table II reports that the majority of the respondents' experiences were between 1 and 5 years, comprising 132 respondents, and among them, 65 respondents are programmers. The ASD experiences of more than 10 years are 2 respondents, namely Project Manager and Product Owner presented in Table II.

b) Organizational background: This section describes the organization sector in the industry. The majority of respondents in the organization sector are banking/financial/insurance, about 29.11%. Then, the percentage of Software Houses is 27.22%, the percentage of the Oil and Gas and other mining industries is 13.29%, and other sectors are depicted in Table III.

2) Descriptive statistics: Descriptive statistics is used to measure the tendency and frequency of each item. Table IV presents that the respondents mostly agree (4) and strongly agree (5) for all items based on the Likert scale, which consists of Strongly Disagree (1), Disagree (2), Neutral (3), Agree (4), and Strongly Agree (5).

From Table IV, it can be seen that the average mean of all elements is 4.25. The average mean of the RE phase element is 4.36 confirming that the first research problem in terms of lack of clarity on which the Requirement Engineering phase to capture QR should be implemented is answered. Then, the average mean of the RE technique is 4.18 confirming that the first research problem in terms of lack of clarity on which the techniques used to capture QR implemented is also answered.

The average mean of documents used is 4.11 confirming that the second research problem in terms of lack of documentation of QR is answered. Then, the average mean of roles involved is 4.25 confirming that the third research problem in terms of the inadequate ability of user stories to capture QR by involving roles to address QR effectively is answered.

Table IV presents that our finding extends the previous work of [29-31] on how to capture QR, which was only implemented in the elicitation phase. According to Table VI, each of the RE phases is needed to implement for capturing QR early. It also extends the previous work of [24], according to Table IV, in terms of documentation used, roles involved, and the RE techniques are also needed for each RE phase. Furthermore, it extends the previous work of Alhaizaey et al. [33], according to Table IV, in terms of documentation used and the roles involved are also needed for each RE phase. On the other hand, according to Table IV, our findings are in line with the documentation used [32] and in terms of implementation for capturing QR in all the RE phases [24], as described in the research gap. In summary, this study fills the gaps from previous works by implementing all the RE elements. Additionally, it also uses comprehensive documentation, the roles involved, and the RE techniques for each RE phase. According to the statistical results, the mean distribution is more than 4 out of 5 suggesting that the framework (Fig. 6) used to capture QR is verified and shows a positive impact on dealing with capturing QR in ASD. However, this framework still needs to prove its effectiveness by conducting validation using case studies.

TABLE III. ORGANIZATION SECTORS

Organization Sectors	Frequency	Percentage
Banking/Financial/Insurance	46	29.11%
Software House	43	27.22%
Oil and Gas, and other mining industries	21	13.29%
Education/Training	9	5.70%
Telecommunication	6	3.80%
E-Commerce	5	3.16%
Manufacturing	5	3.16%
Transportation & Storage	4	2.53%
Healthcare	4	2.53%
Agriculture, Hunting & Forestry	4	2.53%
Construction	3	1.90%
Others	8	5.06%
Total	158	100%

TABLE IV. DESCRIPTIVE STATISTICS AND VARIABLES

Item		Median	Mode	Mean	Average Mean
Element	RE Phase	4	4	4.32	4.25
	Document	4	5	4.41	
	Roles	4	4	4.16	
	RE Technique	4	4	4.09	
RE Phase	Elicitation	4	4	4.04	4.36
	Analysis	5	5	4.42	
	Documentation	5	5	4.46	
	Validation	5	5	4.50	
Document (elicitation phase)	System-type document	4	4	4.04	4.11
	Domain-type document	4	4	4.09	
	ISO/IEC 25010 Model	4	4	4.11	
	Project History	4	5	4.29	
Document (analysis phase)	Mapping sheet between FR vs QR document	4	4	4.04	
	Mapping sheet between QR documents	4	4	4.00	
Document (documentation phase)	FR user story	4	5	4.29	
	QR user story	4	4	4.08	
	QR Decision	4	4	4.08	
Document (validation phase)	Checklist-based Reading document	4	4	4.04	
Roles (elicitation phase)	Developer Team	5	5	4.53	4.25
	Customer	4	5	4.27	
	Expert	4	5	4.29	
	Product Owner	5	5	4.46	
Roles (analysis phase)	Developer Team	4	4	4.15	
Roles (documentation phase)	Developer Team	4	4	3.80	
Roles (validation phase)	Developer Team	4	5	4.29	
	Customer	4	4	4.10	
	Expert	4	4	4.22	
	Product Owner	4	5	4.37	
RE Technique (elicitation phase)	Interview and Brainstorming Technique	5	5	4.46	4.18
RE Technique (analysis phase)	Interaction Matrix technique	4	4	4.03	
RE Technique (documentation phase)	Structured Natural Language technique	4	4	4.03	
RE Technique (validation phase)	Checklist-based reading technique	4	4	4.20	

V. CONCLUSION AND FUTURE WORK

This research proposes a framework for capturing QR by integrating the RE elements in Agile Software Development methods. The research result based on the framework verification using the questionnaire confirms that the ASD practitioners' review agrees with the proposed element and the component to capture QR in ASD. Thus, this framework offers a comprehensive way to handle QR in ASD while aligning with ASD Practice to reduce rework, time, cost, and even project failure. Furthermore, this framework also emphasizes that the RE phase should be iterative along with the process of capturing functional requirements, even requirements (either FR or QR) that arise in the middle of development to accommodate requirement change. For future research, this framework will be validated by using case studies to evaluate its effectiveness, which is in line with the ASD practice, to see the importance and impact of QR on ASD. The case studies will be implemented in selected companies that have used ASD methods for software development.

ACKNOWLEDGMENT

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research. Communication of this research is made possible through monetary assistance from Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

REFERENCES

[1] I. Sommerville, Software Engineering, 10th ed. England: PEARSON, 2016.

[2] S. Najihi, S. Elhadi, R. A. Abdelouahid, and A. Marzak, "Software Testing from an Agile and Traditional view," in Procedia Computer Science, Niagara Falls: Elsevier, 2022, pp. 775–782.

[3] A. S. Alhazmi, "Integrating Design Thinking Model and Items Prioritization Decision Support Systems into Requirements Management in Scrum," Florida Atlantic University, Florida, 2021.

[4] S. Alsaqqa, S. Sawalha, and H. Abdel-Nabi, "Agile Software Development Methodologies and Trends," International Journal of Interactive Mobile Technologies, vol. 14, no. 11, 2020.

[5] A. Muhammad, A. Siddique, Q. N. Naveed, U. Saleem, M. A. Hasan, and B. Shahzad, "Investigating Crucial Factors of Agile Software Development through Composite Approach," Intelligent Automation and Soft Computing, vol. 27, no. 1, 2021.

[6] N. Govil and A. Sharma, "Validation of agile methodology as ideal software development process using Fuzzy-TOPSIS method," Advances in Engineering Software, vol. 168, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0965997822000357

[7] D. Satria, D. I. Sensuse, and H. Noprisson, "A systematic literature review of the improved agile software development," in 2017 International Conference on Information Technology Systems and Innovation (ICITSI), 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8267925

[8] E.-M. Schön, J. Thomaschewski, and M. J. Escalona, "Agile Requirements Engineering: A systematic literature review," Comput Stand Interfaces, vol. 49, 2017, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0920548916300708?via%3Dihub

[9] P. Serrador and J. K. Pinto, "Does Agile work? — A quantitative analysis of agile project success," International Journal of Project Management, vol. 33, 2015, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0263786315000071?via%3Dihub

[10] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," Inf Softw Technol, vol. 50, 2008, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0950584908000256

[11] V. One, "16th Annual State of Agile Report," 2022. [Online]. Available: https://stateofagile.com/

[12] W. Behutiye et al., "Management of quality requirements in agile and rapid software development: A systematic mapping study," Inf Softw Technol, vol. 123, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S095058491930240X

[13] D. Kumar, A. Kumar, and L. Singh, "Non-functional Requirements Elicitation in Agile Base Models," Webology, vol. 19, 2022, [Online]. Available: https://www.researchgate.net/publication/358057432_Non-functional_Requirements_Elicitation_in_Agile_Base_Models

[14] S. Rahy and J. M. Bass, "Managing non-functional requirements in agile software development," IET Software, vol. 16, 2021, [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/sfw2.12037

[15] W. Behutiye, P. Rodríguez, M. Oivo, S. Aaramaa, J. Partanen, and A. Abhervé, "Towards optimal quality requirement documentation in agile software development: A multiple case study," Journal of Systems and Software, vol. 183, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121221002090

[16] B. M. Aljallabi and A. Mansour, "Enhancement approach for non-functional requirements analysis in Agile environment," in 2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), IEEE, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7381407

[17] W. Alsaqaf, M. Daneva, and R. Wieringa, "Quality requirements challenges in the context of large-scale distributed agile: An empirical study," Inf Softw Technol, vol. 110, pp. 39–55, 2019.

[18] R. Kasauli, E. Knauss, J. Horkoff, G. Liebel, and F. G. de O. Neto, "Requirements engineering challenges and practices in large-scale agile system development," Journal of Systems and Software, vol. 172, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121220302417

[19] L. López et al., "Quality measurement in agile and rapid software development: A systematic mapping," J Syst Softw, vol. 186, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121221002661

[20] E. Sherif, W. Helmy, and G. H. Galal-Edeen, "Managing Non-functional Requirements in Agile Software Development," in International Conference on Computational Science and Its Applications, Malaga, Spain: Springer, 2022, pp. 205–216.

[21] K. Curcio, T. Navarro, A. Malucelli, and S. Reinehr, "Requirements engineering: A systematic mapping study in agile software development," J Syst Softw, vol. 139, 2018, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0164121218300141

[22] D. Ismail and Arviansyah, "A Systematic Literature Review and Delphi Study on Agile Software Development Challenges," in 6th International Conference on Management in Emerging Markets (ICMEM), Bandung: IEEE, 2021.

[23] A. Muhammad, A. Siddique, M. Mubasher, A. Aldweesh, and Q. N. Naveed, "Prioritizing Non-Functional Requirements in Agile Process Using Multi Criteria Decision Making Analysis," IEEE Access, vol. 11, 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10061380

[24] E. Sherif, W. Helmy, and G. H. Galal-Edeen, "Proposed Framework to Manage Non-Functional Requirements in Agile," IEEE Access, vol. 11, pp. 53995–54005, 2023.

[25] W. Behutiye, P. Rodríguez, M. Oivo, S. Aaramaa, J. Partanen, and A. Abhervé, "How agile software development practitioners perceive the need for documenting quality requirements: a multiple case study," in 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Portoroz, Slovenia: IEEE, 2020.

[26] A. M. Almanaseer, W. Alzyadat, M. Muhairat, S. Al-Showarah, and A. Alhroob, "A proposed model for eliminating nonfunctional requirements in Agile Methods using natural language processes," in 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), IEEE, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10009796

[27] S. Kopczyńska, M. Ochodek, and J. Nawrocki, "On Importance of Non-functional Requirements in Agile Software Projects—A Survey," in Integrating Research and Practice in Software Engineering, Switzerland: Springer, 2020, pp. 145–158. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-26574-8_11

[28] L. López, J. Partanen, P. Rodríguez, and S. Martínez-Fernández, "How Practitioners Manage Quality Requirements in Rapid Software Development: A Survey," in 2018 IEEE 1st International Workshop on Quality Requirements in Agile Projects (QuaRAP), IEEE, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8501270

[29] M. Younas, I. Ghani, D. N. A. Jawawi, and R. Kazmi, "Non-Functional Requirements Elicitation Guideline for Agile Methods," Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, 2017.

[30] M. Younas et al., "Elicitation of Nonfunctional Requirements in Agile Development Using Cloud Computing Environment," IEEE Access, vol. 8, 2020, [Online]. Available: https://ieeexplore.ieee.org/document/9178791

[31] A. Jarzębowicz and P. Weichbroth, "A Qualitative Study on Non-Functional Requirements in Agile Software Development," IEEE Access,

vol. 9, 2021, [Online]. Available: https://ieeexplore.ieee.org/document/9371679

[32] W. Behutiye, P. Rodriguez, and M. Oivo, "Quality Requirement Documentation Guidelines for Agile Software Development," IEEE Access, vol. 10, 2022, [Online]. Available: https://ieeexplore.ieee.org/document/9810243

[33] A. Alhaizaey and M. Al-Mashari, "A Framework for Reviewing and Improving Non-Functional Requirements in Agile-based Requirements," in 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal: IEEE, 2023.

[34] L. Chung and J. C. S. do P. Leite, "On Non-Functional Requirements in Software Engineering," in Conceptual Modeling: Foundations and Applications, vol. 5600, Springer, 2009, pp. 363–379.

[35] J. Doerr, D. Kerkow, T. Koenig, T. Olsson, and T. Suzuki, "Non-functional requirements in industry - three case studies adopting an experience-based NFR method," in 13th IEEE International Conference on Requirements Engineering (RE'05), Paris, France: IEEE, 2005.

[36] B. W. Boehm, J. R. Brown, and H. Kaspar, Characteristics of software quality, 2nd ed. North-Holland Publishing Company, 1978.

[37] J. McCall, "Factors in software quality," 1977.

[38] ISO/IEC JTC 1/SC 7, "ISO/IEC 25010," https://iso25000.com/index.php/en/iso-25000-standards/iso-25010.

[39] C. Boyce and P. Neale, CONDUCTING IN-DEPTH INTERVIEWS: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input, vol. 2. Pathfinder International, 2006.

[40] S. W. Ambler, "Beyond functional requirements on agile projects - Strategies for addressing nonfunctional requirements," Doctor Dobbs Journal, vol. 33, no. 10, pp. 64–66, 2008.

[41] V. Sachdeva and L. Chung, "Handling non-functional requirements for big data and IOT projects in Scrum," in 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, IEEE, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7943152

[42] K. Beck et al., "Manifesto for Agile Software Development."

[43] D. Mairiza and D. Zowghi, "Constructing a Catalogue of Conflicts among Non-functional Requirements," in International Conference on Evaluation of Novel Approaches to Software Engineering, Athens, Greece: Springer, 2010, pp. 31–44.

[44] D. Mairiza, D. Zowghi, and N. Nurmuliani, "An investigation into the notion of non-functional requirements," in Proceedings of the 2010 ACM Symposium on Applied Computing, Sierre, Switzerland: Association for Computing Machinery, 2010, pp. 311–317.

[45] A. E. Sabry and S. S. El-Rabbat, "Proposed framework for handling architectural NFR's within scrum methodology," in International Conference Software Engineering Research and Practice, SERP, 2015, p. 238.

[46] G. Sabaliauskaite, F. Matsukawa, S. Kusumoto, and K. Inoue, "An Experimental Comparison of Checklist-Based Reading and Perspective-Based Reading for UML Design Document Inspection," in Proceedings International Symposium on Empirical Software Engineering, Nara, Japan: IEEE, 2002.

[47] N. K. Rad and F. Turley, Agile Scrum Handbook, 2nd ed. Van Haren Publishing, Zaltbommel, 2018.

[48] Miro, "What is brainstorming?," https://miro.com/brainstorming/what-is-brainstorming/.

[49] K. Wiegers and J. Beatty, Software Requirements, 3rd Edition. Redmond, Washington: Microsoft Press, 2013.

[50] R. Ibrahim, N. A. M. Asri, S. Jamel, and J. A. Wahab, "Validation of Requirements for Transformation of an Urban District to a Smart City," Int J Adv Comput Sci Appl, vol. 12, no. 7, pp. 322–328, 2021.

[51] N. A. M. Asri, R. Ibrahim, and S. Jamel, "Designing a Model for Smart City through Digital Transformation," International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 1.3, pp. 345–351, 2019.

[52] J. Frost, "Cronbach's Alpha: Definition, Calculations & Example," https://statisticsbyjim.com/basics/cronbachs-alpha/#:~:text=Cronbach%E2%80%99s%20alpha%20coefficient%20measures%20the%20internal%20consistency%2C%20or,agreement%20on%20a%20standardized%200%20to%201%20scale.

[53] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, "A Primer on Partoa; east Squares Structural Equation Modelling (PLS-SEM)," International Journal of Research & Method in Education, vol. 38, 2016.

[54] D. W. Zimmerman, "A note on the influence of outliers on parametric and nonparametric tests," Journal of General Psychology, vol. 121, no. 4, pp. 391–401, 1994.

[55] B. G. Tabachnick and L. S. Fidell, Using Multivariate Statistics, 7th ed. Boston: Pearson Education, 2021.

[56] Holmes-Smith P, Coote L, and Cunningham E, Structural Equation Modeling. School Research, Evaluation and Measurement Services, 2006.

Data Mining for the Analysis of Student Assessment Results in Engineering by Applying Active Didactic Strategy

César Baluarte-Araya, Oscar Ramirez-Valdez, Ernesto Suarez-Lopez, Percy Huertas-Niquén
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

***Abstract*—To make improvements in the teaching-learning process in educational institutions such as universities, it is necessary to analyse the results obtained and recorded from applying Active Didactic Strategies and, based on this, to propose improvements that will help to achieve the Student Outcomes established for the subject in question; the problem to be solved is thus defined, and the results to be obtained from the analysis are relevant for the improvement of student performance. The objective is to analyse the results of the student assessment, the basis for the calculation of which is based on the recording of the qualification achieved through the performance indicators defined for each criterion, of the competencies involved and aligned with the Student Outcomes of the problems proposed to the student, applying various data mining techniques. Data mining is used to treat large amounts and types of data to obtain hidden information and reveal states, patterns and trends; as well as in Education to study the behaviour of students in terms of their performance. The methodology used for the development of the work is based on the Cross-Industry Standard Process for Data Mining methodological model, which is widely used in data mining projects. The results obtained reveal that the Student's t-test and Snedecor's F-test are highly significant, as well as the determination of the lowest performance indicators in order to plan future improvement actions towards better student performance and achieve a high level of learning. Concluding that if the same teaching and learning process is applied the result will be very similar, therefore, the students have finished learning very well.**

***Keywords*—Student outcome; problem based learning; assessment; performance indicator; data mining**

I. Introduction

Considering that the main benefit of Data Mining (DM) is the power to identify patterns and relationships in the gigantic volumes of data found in numerous sources.

Also, other aspects such as the growth of data from educational sources ranging from university admissions, academic information systems, monitoring, assessment, graduation, as well as IT and teaching and learning support platforms is enormous.

The analysis of the data recorded is essential to obtain information, reach conclusions and therefore also to see the correlation in the performance of students that allows to produce multiple results, such as passes, fails, dropouts, grades, among others, and to make decisions for improvement. This is how nowadays the further development of DM to discover the knowledge that may be hidden in databases, has allowed to demonstrate its effectiveness in different contexts, being the educational one the one that is growing in its application giving rise to Educational Data Mining (EDM).

The objective of the research work is to apply EDM to analyse the results of the evaluation of students in the area of Engineering when applying Active Didactic Strategies, which in the case of the study is Problem Based Learning in the course of Electronic Business in the Professional School of Systems Engineering (EPIS) [1] of the Universidad Nacional de San Agustín de Arequipa (UNSA) [2]; and thus discover trends, patterns, behaviours based on the available data and analyse the performance of the students according to the results of the evaluations reflected in the grades obtained, as well as the grades given to the different performance indicators determined for each defined criterion and included in the Reports of Deliverables and Formative Research developed by the students during the academic semesters covered in this work. Additionally, combined with a severe statistical analysis to determine patterns of usage.

UNSA's Educational Model is based on the professional training of students based on competences and the application of Active Didactic Strategies [3] to achieve the objectives and competences established for training.

The EPIS 2022 achieved official accreditation certification by the Accreditation Board for Engineering and Technology (ABET) [4], which has as one of its objectives that the University demonstrates that its graduates and trainees achieve the desired competencies and student outcomes.

The research produced is applied and descriptive, the methodology used is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodological model [5].

The main result obtained is that the models used have a high significance according to the t-tests and Snedecor's F-tests to which they are subjected.

It is concluded that, if the same teaching and learning process is applied in the course, the result obtained will be very similar, which allows us to visualise that what is planned and executed is well designed; and therefore the student ends up learning very well.

II. Theoretical Framework

A. Active Didactic Strategies

The EPIS professors plan as one of their objectives to apply a certain Didactic Strategy for the development of the course, taking into account that the students are trained through the competence-based approach being the centre of attention as referred to by [6], by applying an established or agreed active didactic strategy contemplated by [7], [8] without leaving aside what is dealt with by [9] which is improved with technology.

The training of students with Active Didactic Strategies implies autonomous and continuous learning, as concluded by [10], ordering their learning, being a permanent work in the professional training.

In the experience of using the active didactic strategy of Problem-Based Learning (PBL), as a learning methodologies, students work in small groups of three or four members executing the planning, organisation, planning the solution, developing it, validating it and thus satisfying the requirements of the problem or problems of reality.

The application of the aforementioned Active Didactic Strategies encourages students to work in teams, increasing their critical capacity, broadening their knowledge of the problem area, developing the competences of the course and therefore those of their professional training, increasing their abilities as well as their procedural skills, valuing their achievements and thus achieving a certain experience in problem-solving for the good of society.

The adaptation and improvement in the application of the Active Didactic Strategy and its complementary techniques are of vital importance to improve students' performance, and teachers are the ones most called upon to adapt them to the teaching and learning methods for the benefit of the students.

B. Indicator-Based Learning Assessment

The evaluation, reflected in a grade that the student obtains to continue to the next level of their study plan, must be based on data and complementary information, to know the student's knowledge deficiencies, weaknesses, the application of methodologies, methods, techniques, tools, as well as understanding the technologies and their limitations in the application of the probable solutions to the problem being dealt with.

In his proposal [11] he contemplates the Evaluation System based on Performance Indicators for Problem Based Learning, which contemplates in its structure: a) General Competence(s) (of the graduate profile), b) Course Competence(s), c) Student Outcomes, d) Process(es), e) Criterion(s), f) Performance Indicator(s), g) Indicator(s) rating scale that allow a more objective, fair evaluation; and have data that allows its analysis to reach conclusions; to its initial structure was adapted the inclusion of the element 'Student Outcomes', which is related to the other elements and application for the rating in the evaluation.

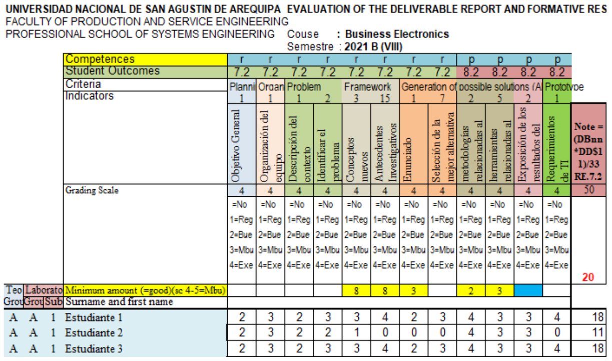
As a result of the application of the same in study [12] from which the set of stored data of the academic semesters involved in the present work is derived.

The Student Outcomes defined for the NE course based on those of the EPIS for the NE course are:

7.2 Acquire new knowledge in learning using appropriate strategies in the context of technological changes.

8.2 Selects appropriate tools, skills, techniques, methods and/or methodologies specific to the discipline with an understanding of its limitations.

Fig. 1 shows some of the performance indicators graded according to a grading scale, from which the mark obtained by the student for each problem proposed for development is calculated.



UNIVERSIDAD NACIONAL DE SAN AGUSTIN DE AREQUIPA EVALUATION OF THE DELIVERABLE REPORT AND FORMATIVE RES
FACULTY OF PRODUCTION AND SERVICE ENGINEERING
PROFESSIONAL SCHOOL OF SYSTEMS ENGINEERING Couse : Business Electronics
Semestre : 2021 B (VIII)

			Competences	r	r	r	r	r	r	r	r	p	p	p	p	
			Student Outcomes	7.2	7.2	7.2	7.2	7.2	7.2	7.2	7.2	8.2	8.2	8.2	8.2	
			Criteria	Planni	Organ	Problem		Framework		Generation of possible solutions (A					Prototype	
			Indicators	1	1	1	2	3	15	1	7	2	5	2	1	
				Objetivo General	Organización del equipo	Descripción del contexto	Identificar el problema	Conceptos nuevos	Antecedentes Investigativos	Enunciado	Selección de la mejor alternativa	metodologías relacionadas al	herramientas relacionadas al	Exposición de los resultados del	Requerimientos de TI	Note = (DBnn *DD$1 1)/33 RE.7.2
			Grading Scale	4	4	4	4	4	4	4	4	4	4	4	4	50
				=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	=No 1=Reg 2=Bue 3=Mbu 4=Exe	20
Teo	Laborato		Minimum amount (=good)(sc 4-5=Mbu)					8	8	3		2	3	🟦		
Grou	Grou	Sub	Surname and first name													
A	A	1	Estudiante 1	2	3	2	3	3	4	2	3	4	3	3	4	18
A	A	1	Estudiante 2	2	3	2	2	1	0	0	0	4	3	3	0	11
A	A	1	Estudiante 3	2	3	2	3	3	4	2	3	4	3	3	4	18

Fig. 1. Grading tool for the deliverable report.

Every evaluation must generate complementary information which must be organised, consolidated, analysed, interpreted, communicated to whoever it may concern, generate knowledge and thus be useful for planning, taking actions and decisions. Thus, it is important to analyse the academic performance of students in order to take timely action individually, as a team and globally for all participants.

C. Data Mining

For study [13] 'Data Mining is the process of extracting useful information from an accumulation of data, often from a data warehouse or from the collection of linked data sets' and also for study [14] 'is the process of discovering patterns and other valuable information in large data sets'.

In the DM process [15] refers to the set of methods, techniques and technologies to explore large databases, automatically, in order to find patterns, trends that help to interpret the behaviour of the data in a given context. Also in research [13] notes that it uses statistical analysis, machine learning algorithms in the search in large databases, analysing the data from different angles and thus identifying trends, patterns and unknown associations.

The application of data mining occurs in different contexts of organisations, as for example in study [16], the Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context is proposed; it is also in the context of education that DM is increasingly taking place to extract information from data and be used to plan projects, improve processes in teaching and learning.

Given the growth of the application of DM in the educational context, it has given rise to Educational Data Mining (EDM) treated as such in study [17] by contemplating that the usual approaches and assessments are not suitable to discover the revealing information of the student's scores. Therefore, EDM deals with cases of student data research used to investigate associations not preliminarily detected in a student database. Also in study [18] state that EDM focuses on developing methods for discovering that Learning Management System (LMS) data is used to understand and make improvements in virtual learning environments, as well as building analytical models to discover patterns and trends. Many works deal with student performance as in studies [19], [20], [21], [22], [23], [24], and also in study [25] where the detection of factors associated with performance from educational assessment databases is discussed; as well as somehow dealing with KPI indicators to map strategies and performance indicators applying predictive modelling as applied by study [29].

What is normally followed in a DM process is the application of the CRISP-DM methodology in its six phases or stages that ensures that the expected results are obtained.

D. Data Mining Techniques and Tools

Data mining techniques are being increasingly adopted in studies related to data processing in organisations, which use them for data processing, analysis, generation of information for decision making [14].

Data mining tools contain powerful statistical, mathematical and analytical capabilities to examine huge data sets to identify trends, associations, patterns, correlations, intelligence and strategic organisational information to support planning and decision making [13].

In study [19], they have used DM techniques to generate predictive models of academic performance: decision trees and multivariate regression.

On the other hand, in study [17] they use the statistical package for social sciences (SPSS V26) for both descriptive and inferential statistics, using ANOVA as a method of statistical analysis; [21] uses IBM SPSS MODELER 18.1 through classification techniques.

Also, in [20] they use the best known, WEKA classifiers; so also in study [22] additionally use it with clustering; and also [25] use it to develop statistical analysis of massive information from evaluation databases using association, clustering, classification.

In study [23] they use classification with machine learning algorithms; also in study [26] they used Machine Learning techniques.

Also, in research [24], the need to consider qualitative and quantitative elements to predict and evaluate the academic performance of students is highlighted. Through machine learning, educational data is refined, discovering valuable patterns and simplifying complexities through feature selection methods using clustering, classification and regression.

III. Related Work

The field of action and application of DM is very broad in many contexts of activity in organisations; thus, [27] in the field of IT project management, when defining a new project, it is necessary to choose the development team considering characteristics, roles, results of previous projects, experience of the developers, suitability and affinity. As well as in study [16] in the e-commerce environment to predict customer buying behaviour.

In the context of education, the study [19] performs an analysis of the academic performance of students entering a degree programme, relating performance to socio-economic and academic characteristics stored in a database; as well as in study [17] they analyse the performance of students in TOEFL reading, listening and writing scores; study [20] investigates whether there are patterns in the available data that can be useful for predicting performance based on personal and pre-college characteristics, in research [21] as the aim is to automatically classify students based on their academic performance and to identify profiles and trends as well as student attrition; the study[22] uses classification techniques to study and analyse student performance, as well as to help to advise students; the study [25] develops statistical analyses of massive information from student assessment databases, also presenting to the scientific community statistical procedures and techniques that can be valuable and replicable in other educational and/or social spaces; also in study [23] proposes a model for predicting student performance, allowing accurate identification of final grades and providing information to guide educational interventions. Also in study [26], the analysis of students' academic performance is carried out using the Power BI tool, evaluated different machine learning techniques and found Random Forest to be the most efficient algorithm, with the highest accuracy of academic performance. In study [24] their systems collect data from various sources: exams, virtual courses, enrolments, e-learning platforms; the analysis of this data leads to the application of machine learning techniques to predict and evaluate student performance.

IV. Methodology

The methodology used in this work is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodological model. For study [13] there are as many approaches to DM as there are data mining workers. The approach defined depends on the type of questions asked, as well as the content and organisation of the database.

In this regard [5] summarises the phases or stages of CRISP-DM implementation developed from the knowledge discovery processes widely used in organisations, responding directly to user requirements; namely: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Implementation. Fig. 2 shows the six phases of CRISP-DM.

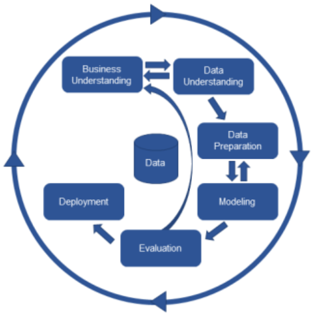


Fig. 2. CRISP-DM model. source (Taylor, 2017).

In many organisations and in research work carried out, the application of CRISP-DM implementation phases is generated according to the problem to be addressed; thus, the study [13] proposes the phases of: Understanding the problem, or at least the research area, Data collection, Data preparation and understanding, User training; also in study [28] in his guide proposes the phases of: : Understanding the business, Understanding the data, Data preparation, Modelling, Evaluation, Virtual, which is closer to the essence of CRISP-DM.

Already in the application of CRISP-DM we have that [20], develops the work with the phases of: Business Understanding, Data Understanding, Data Preprocessing, Modelling, working then the results; in study [18], [21], they use the phases of: Business Understanding, Data Understanding, Data Preparation, Data Preparation, Modelling, working then also the results; in study [26], they deal with the phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Visualisation.

The data mining project uses the tools of: IBM SPSS Modeler V.21 which supports decision tree models, neural networks and regression models; to do statistical analysis of the data, determine relationships between Student Outcomes and their assessments. It is used to perform data capture and analysis, create tables and graphs with complex data; known for its ability to handle large volumes of data as well as perform text analysis among other formats.

IBM SPSS Modeler software offers advanced statistics in addition to many basic statistical functions, including cross-tabulation, frequencies, dual variable statistics such as t-tests and ANOVA, correlation, linear and non-linear models.

RAPIDMINER is a free software tool used for data mining tasks in both research and organisations, supporting single and multidimensional predictive models. It works from nestable operators to compose the models to work with; to do data analysis, determine patterns.

It allows to accelerate the creation, delivery and maintenance of predictive analytics; dealing with large volumes of data from multiple different sources, managed by more business-oriented than technical profiles.

Rapidminer software by the use of its workflow system decreases the use of code for data modelling, speeding up the analysis.

The data mining task is to analyse the university performance of students from their assessments.

V. Development of the Proposal

A. *Research Context*

The experience is developed in the subject of Electronic Business (NE) that correspond to the V semester respectively of the EPIS Curriculum.

From the academic year 2020 to the year 2023, the tool of qualification of the Deliverable Report and Formative Research Report is applied, the result of which by Feedback Report is given to the students so that in work team sessions they analyse, reach conclusions on the qualification of each Performance Indicator, propose improvements and apply them in the development and solution of the following problems.

In the development of the course it is contemplated and ensured that there is alignment between the Student Outcomes, the contents, the evaluation method, the use of the grading tool that encompasses what is proposed by study [11].

B. *Development of the Model*

The data mining work is based on the CRISP-DM research approach as it is a non-proprietary, application-neutral standard for data mining projects and widely used by users.

Phase 1: Understanding the business

The objectives and requirements of the business, which in this case is the education sector, are understood.

The objective of the research is to analyse the results of the student evaluation by applying different data mining techniques.

Among the main requirements are:

- Determine the relationship between Student Outcomes and their Evaluation Process.
- To analyse the results of the application of different data mining algorithms.
- Determine trends or patterns.
- Show the lowest performing performance indicators in order to plan and take future improvement actions.

Phase 2: Understanding the data

It comprises the tasks of initial data collection, establishing its main characteristics such as: structure, quality, identifying likely subsets of the data of interest.

The initial data collection is given by the recording of the ratings made in the respective assessment of each problem addressed through the tool used.

The description of the data is given by the composition of the 1681 records of the grades recorded in the tool in Excel, the data are grouped into two categories:

- Course General Data with the Final Results of the evaluation of each student: Faculty, Professional School, course, credits, type, year, semester, theory group, laboratory group, laboratory subgroup, analysis code, surname and first name, gender; marks for: final, exam and continuous averages, exams 1 2 3, continuous 1 2 3, student results 7.2 and 8.2; dropout.
- Data of the Results of the Grading of the Deliverable Reports made by the students: Problem, problem name, problem score, student performance indicators score, problem assessment criteria scores, student outcomes scores 7.2 and 8.2 of the problem.

The exploration of data by understanding its structure and characteristics.

In the Final Assessment Result 5 categories are considered to appreciate the behaviour of the assessment; Fig. 3 shows the behaviour of the Final Assessment Results variable of the students in the course, appreciating that category 4 with a rating of Very Good with a mark of 14 to 16 prevails, and category 5 with a rating of Outstanding with a mark of 17 to 20 has a significant presence not so in the year 2021.

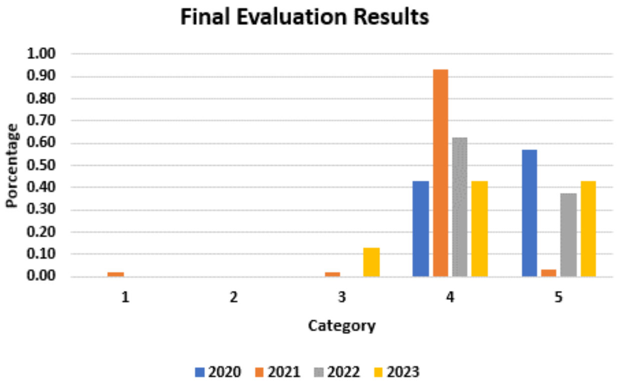


Fig. 3. Final evaluation results.

Fig. 4 shows the behaviour of the variable: Gender of the students, showing that the presence of men prevails, which is derived from the application and entry to study in the study programme.

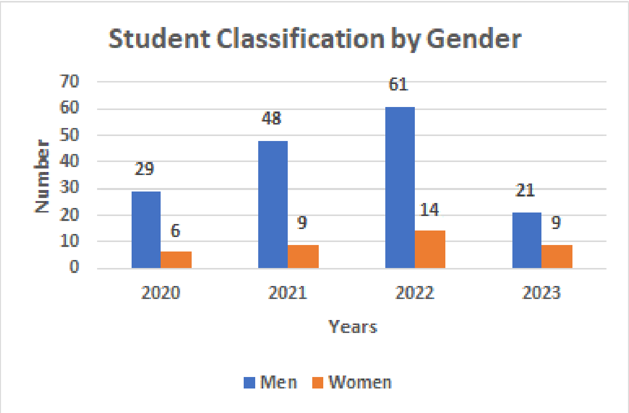


Fig. 4. Classification of students according to gender.

Table I shows the composition of the Theory Groups, Laboratory Groups and Subgroups.

TABLE I. CONFORMATION OF GROUPS AND SUBGROUPS

Year	Groups Theory	Groups Laboratory	Subgroups Laboratory
2020	A	A	5
	B	B	5
2021	A	A	4
	B	B	6
		C	5
2022	A	A	5
	B	B	5
		C	5
		D	5
2023	A	A	4
	B	B	4

Table II shows the problems dealt with in the development of the course by applying Problem Based Learning.

TABLE II. PROBLEMS

Coding	Description
1	Virtual Stores
2	Customer Relations Management - CRM
3	Supply Chain Management - SCM
4	e-Marketplace
5	e-Learning
6	e Employee
7	e-Government
8	m-Business

The data quality check checked the consistency of the data values.

Phase 3: Data preparation

This phase involves the activities to extract the student data from the assessment record and build the final dataset into a dataset described with 112 data for each instance or occurrence that serves for modelling.

The Selection of the data covers the student assessments for the years 2020 to 2023 which are 1681 records of the grades recorded in the grading tool.

Data Cleaning deals with the inclusion of missing data, such as gender; or derived from existing data such as dropout/dropout; and data correction such as replacing blank value with zero in performance indicators.

The Construction and Integration of the data is given by the elaboration of the file containing the two categories: General Course Data with the Final Results of the evaluation of each student and the Data of the Grading Results of the Deliverable Reports made by the students, and thus obtaining the final dataset.

Data Formatting converts the data types for the application of the specific DM technology. The 66 performance indicators are nominal variables with four or five different values.

Phase 4: Modelling

In this phase, the modelling techniques that best enable the analysis of the data of the data mining project are selected. The data are entered into the data mining software, the runs are performed and the results are studied.

In the modelling, the following options have been used:

- Linear Regression, supported by correlation coefficients, Student's t-test and Snedecor's F-test.

The Student's t-test is used when the population does not follow a normal distribution. In this case we do not know the behaviour of the population data so we understand that there is no normal distribution because this t-test is robust to deviations from the normality of the population.

The F-test is used to assess the overall significance of a regression model. This test gives strength to the analysis of variance of linear regression.

According to the systematic review of the research works discussed or reviewed, we did not find a general model that deals with student performance indicators for each of the criteria proposed in an evaluation based on the definition of the rubric that allows us to evaluate the level of performance or performance of a task, which in this case are the proposed problems; Our proposal differs because we reach the level of qualification in the evaluation of the performance indicators defined for each criterion; and we conceptualise a scheme based on periods, groups and subgroups, problems, managing to reach estimated models; therefore, it is determined to use the splines for the determination of patterns.

Phase 5: Evaluation

The models created are measured against the defined objectives; according to the result feedback is given and the resulting modifications are part of the knowledge discovery process.

Phase 6: Implementation and Visualisation

In this phase, the generated models are used by other users to produce business intelligence. Based on the results obtained, there are tables and graphs that allow visualisation, the generation of reports and improvements in the procedures of the processes.

VI. Results

The main objective of the study is firstly to perform the statistical analysis of data to determine the relationship of Student Outcomes and their Assessment Process; secondly to perform the data mining process to determine the existence or not of patterns, correlations or trends in the assessment data.

The results of the statistical analysis of the SPPS Modeler application are shown below.

There is a high relationship between the predictor variables and the R72 model, in the same way, there is similarity with the R82 model, which implies that in the R82 model there is a greater effort on the part of the students to learn. Highly significant results by analysis of variance (F-test). These results are reflected in Table III and Table IV.

On the other hand, the constants of the variables of the estimated models show a highly significant Student's t-test, supporting the result of the correlation coefficient. These results can be seen in Table V.

TABLE III. Results of the Correlation Coefficient for the R72 y R82 Models

Model	R	R^2	R^2 corrected	Predictor variables
R72	0,963	0,928	0,926	(Constant), CO3, EX3, EX1, EX2, CO1, CO2
R82	0,944	0,891	0,887	(Constant), CO3, EX3, EX1, EX2, CO1, CO2

TABLE IV. Analysis of Variance

Model	Sum of squares	gf	Root mean square	F	Significance
R72	547,947	6	91,324	409,448	,000
R82	1002,675	6	167,112	258,653	,000

TABLE V. Student t Test

Predictor variables	Dependent variable			
	Model R72		*Model R82*	
	t	*Significance*	*t*	*Significance*
Constant	3,127	,002	-1,990	,048
EX1	13,743	,000	-3,108	,002
EX2	11,510	,000	1,829	,069
EX3	11,527	,000	2,398	,017
CO1	10,451	,000	10,830	,000
CO2	3,724	,000	13,905	,000
CO3	11,895	,000	5,619	,000

From the application of the Rapidminer tool we have the results of applying splines for the determination of patterns.

Fig. 5 shows that in the RE82 model, students are assertive in the handling of techniques, methods and tools, but they have difficulties in the RE72 model in coping with learning. It can be seen that between grades 13 and 17 students make the greatest effort in learning and applying information technologies; it can also be seen that from grade 18 onwards they culminate with their learning.

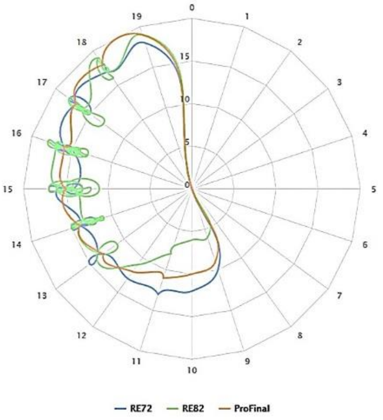


Fig. 5. R72_R82_PFINAL.

Fig. 6 shows the Student Outcome model 7.2 - RE72 which deals with the average exam marks and average continuum marks, indicating that students between marks 13 and 15 have greater practical skills and greater difficulty in theoretical learning; from the upper limit mentioned above, both theoretical and practical aspects are combined to reach the desired level of learning.

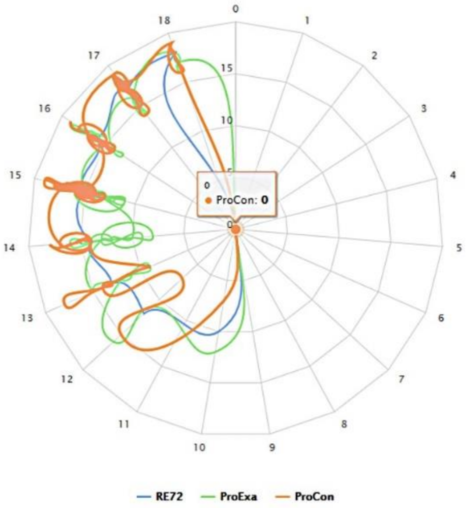


Fig. 6. R72_PEXA-PCON.

On the other hand, Fig. 7 shows the Student Outcome model 8.2 - RE82 which deals with the average marks of exams and the average of continuations, indicating that the student understands the practice better through the use of techniques, methods and information technology tools that work in parallel with the proposed RE82 model; and the theory is adapted according to the practice until mark 18, at which point they are combined with the model, reaching the desired level of learning.

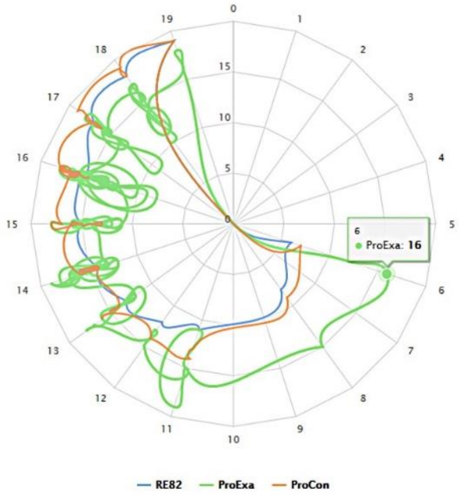


Fig. 7. R82_PEXA-PCON.

From the result of the treatment of the qualification of the performance indicators of the criteria contemplated for the evaluation of the problems proposed and developed by the students, those performance indicators (66 in total) are shown that in the qualification scales have the scale value of Absent (>= to 15%) and/or Regular (>=10%) that allows the teachers to pay attention to reflect weaknesses to improve, and to complement the analysis of the results. and based on the conclusions to raise the improvement actions to raise the level of performance of the students. These actions, when dealt with in the first class, session of the following semester, also encourage students to take into account the fulfilment of the tasks to be completed in all the indicators and thus have a high level of learning.

Table VI shows the performance indicators involved.

TABLE VI. PERFORMANCE INDICATORS INVOLVED

Indicators Involved				
Indicator		***Percentages Scale***		
Cod	***Name***	***Absent***	***Regu lar***	***Total***
12	Organisation of information	6.70	11.30	18.00
13	New concepts	4.50	17.60	22.10
18	Search for models related to the problem	5.80	13.50	19.30
20	Search for methods related to the problem	7.60	14.90	22.50
21	Search for techniques related to the problem	7.40	13.00	20.40
22	Search for tools related to the problem	7.30	18.30	25.60
23	Search for the necessary skills	8.60	13.60	22.20
25	Research background	11.70	15.10	26.80
26	Configuration of tools (SW)	16.60	9.40	26.00
27	Installation of tools (SW)	17.30	6.20	23.50
28	Sharing information worked on	16.00	12.60	28.60
30	Drawing up the comparative table of methodologies	15.10	3.30	18.40
31	Drawing up the comparative table of methods	16.90	3.20	20.10
32	Drawing up the comparative table of techniques	16.80	3.50	20.30
34	Drawing up the table of necessary skills	16.30	7.90	24.20
35	Propose and support the IT collected.	16.40	7.70	24.10
43	Others.	30.40	4.90	35.30
44	Selection of the best alternative	18.70	7.00	25.70
50	IT requirements	15.10	13.20	28.30
51	Configuration	15.10	8.10	23.20
52	Installation	17.10	6.40	23.50
53	Prints Screen of results	12.10	11.50	23.60
56	Discussion or comments	30.60	3.50	34.10
58	Writing consistent with objectives	11.50	18.00	29.50
62	Other	20.60	14.80	35.40

It is also emphasised that many efforts are being made to increase the learning levels of Peruvian students, such as the effort being made by UNSA to apply Active Didactic Strategies in the teaching-learning process in order to increase student learning levels.

VII. DISCUSSION

There are several works in Education such as that of [27] in the area of Data Mining; in the selection of computer project teams; as well as in the area of those related to students' academic performance from those of [19], [20], [25], [23], [26],

[24], or to extract academic behavioural profiles such as that of [21] or also that of [22] to help advise students and predict their academic performance.

Having that the prediction of academic performance shows challenges due to the varied factors that influence it as discussed by [24] by referring to [H.A.A. Hamza, P. Kommers, 2018, and M.M.A. Tair, A.M. El-Halees, 2012], having that through prediction by machine learning is instrumental in improving education in various ways, by allowing early identification of deficiencies, academic difficulties, enabling timely interventions and personalized learning plans.

In view of the above and nowadays the availability of information technologies with sufficient capacity to process large and varied types of data has allowed Data Mining techniques to evolve and allow obtaining, processing and detecting information from large amounts of data tools Nghe et al., 2007, cited in [25].

This study uses IBM SPSS Modeler v. 21 software for statistical analysis of the data to determine relationships and Rapidminer software to determine the presence of patterns, showing the potential of techniques and algorithms that respond to the stated objective by applying simple data analysis tools.

Continuous monitoring promotes quality assurance, accountability and a competitive advantage for institutions. Overall, it empowers educators to provide targeted support to plan and make improvement decisions, leading to better student outcomes and a more responsive education system as manifested in [24].

A variety of techniques are also used that based on different algorithms allow the processing of data from a given storage source and the evaluation of the results provided to make appropriate decisions.

VIII. Conclusion

The following conclusions are reached:

The R72 model considers the Student's t-test which is highly significant, which means that the prognosis that can be obtained, the results are totally valid; that is to say that if the same teaching-learning process is applied the result is going to be very similar. Because there is a high relationship between the RE72 (Student Outcome 7.2) and its evaluation process, where the Snedecor's F-test is highly significant. That is, in the teaching and learning process the students have finished learning. Similarly, the R82 model shows similarity with respect to the RE72 model.

According to the result of the analysis students have some difficulty in creating results (in the Deliverable Report) of creating the product. There is a problem in the starting point of the course; and they start to understand in Co1 (Continuous Assessment 1) that deals with the development of the first two (2) problems; which is not enough to pass the EX1 (Exam 1); and from CO2 (Continuous Assessment 2) they correct themselves noticing the improvement in the learning and they manage to finish the learning.

The students in the RE82 model show assertiveness in the handling of techniques, methods and tools; however, in the RE72 model they present certain difficulties in dealing with theoretical learning and at the end of the period they complete their learning.

The Student Outcome model 7.2 - RE72 PEXA-PCON, indicates that the student, in certain sections, presents greater skills for practice and in other sections greater difficulty for theoretical learning; however, from the upper limit, both the theoretical and practical aspects are combined, reaching the desired level of learning.

The Student Outcome Model 8.2 - RE82 PEXA-PCON, indicates that the student understands the practice better through the use of techniques, methods and information technology tools that are performed in parallel with the proposed RE82 model in many sections; and the theory is adapted according to the practice until the upper section, where they are combined with the model, reaching the desired level of learning.

It is possible to determine the highest and lowest performance indicators to which attention should be paid by complementing the factors that influence the lower performance of students and thus plan future improvement actions that allow better performance of students and their level of learning to be at a high level. This is an advantage over other forms of assessment, as it leads to the rating of each Performance Indicator involved in each Criterion, which in its analysis and evaluation makes it possible to determine the status level according to its rating scale and to make decisions to improve student performance.

It allowed us to see the additional particular consideration that can be given to a system by considering, for example, the Hierarchy of levels that the evaluation system contemplates from Faculty, Professional School, Course, Academic Year and Semester, Theory Group, Laboratory Group and Subgroup, Problem.

It allowed to visualise the future applicability of the analysis that can be carried out by the use of other Active Didactic Strategies that are used in other courses in the teaching-learning process.

It is recognised that the scope of future application can be extended to other courses of the EPIS Syllabus, to other Professional Schools of the Faculty, as well as Professional Schools of other Faculties and thus be able to obtain results at Faculty Level, University of the application of the Evaluation by Performance Indicators determined for each of the Criteria contemplated, which differs as a working method with respect to the treatment of students‘ performance dealt with by other works, for the development of students’ Competences when applying Active Didactic Strategies in the teaching-learning process.

Acknowledgment

Thanks to the Universidad Nacional de San Agustín de Arequipa for the support it provides for the development and implementation of proposals that benefit the continuous improvement of student performance through the application of Active Teaching Strategies, with beneficial results for students as well as the teaching-learning process.

REFERENCES

[1] Escuela Profesional de Ingeniería de Sistemas. http://www.episunsa.edu.pe.

[2] Universidad Nacional de San Agustín de Arequipa. http://www.unsa.edu.pe.

[3] C Ma. Cristina Sánchez Martínez, M. Aguilar Venegas, J.L. Martínez Durán and J. L. Sánchez Ríos, Estrategias didácticas en entornos de aprendizaje enriquecidos con tecnología (antes del Covid-19), UNIVERSIDAD AUTÓNOMA METROPOLITANA-XOCHIMILCO, México, 2020, pp. 11-15.

[4] ABET. Why ABET Accreditation Matters. https://amspub.abet.org/aps/category-search?countries=PE.

[5] Ultimo acceso julio 2023.

[6] Connolly, Thomas; Begg, Carolyn, Sistemas de Bases de Datos, cuarta edición, Pearson Educación S.A., España, 2005, págs.1120-1122.

[7] B. Restrepo Gómez, Aprendizaje basado en problemas (ABP): una innovación didáctica para la enseñanza universitaria, Educación y Educadores, vol. 8, 2005, pp. 9-19, Universidad de La Sabana, Cundinamarca, Colombia.

[8] Subdirección de Currículum y Evaluación, Dirección de Desarrollo Académico, Vicerrectoría Académica de Pregrado, Universidad Tecnológica de Chile INACAP. (2017). “Manual de Estrategias Didácticas: Orientaciones para su selección”, Santiago, Chile: Ediciones INACAP.

[9] R. Rodriguez, "Compendio de estrategias bajo el enfoque por competencias", Instituto Tecnológico de Sonora, México, 2007. http://www.itesca.edu.mx/documentos/desarrollo_academico/compendio_de_estrategias_didacticas.pdf.

[10] Ma. C. Sanchez, M. Aguilar, J.L. Martinez and J.L. Sanchez, “Estrategias didácticas en entornos de aprendizaje enriquecidos con tecnología (antes del Covid 19)”, Universidad Autónoma Metropolitana, No. 146 series académicos, Mexico, 2020.

[11] A. León, E. Risco del Valle, and C. Alarcón, "Estrategias de Aprendizaje en educación superior en un modelo curricular por competencias", Revista de la Educación Superior, Vol. XLIII (4); No. 172, pp. 123-144, octubre-diciembre del 2014. ISSN: 0185-2760.

[12] C. Baluarte-Araya, " Proposal of an Assessment System based on Indicators to Problem Based Learning – IEEE Conference Publication, Published in: 2020 39th International Conference of the Chilean Computer Science Society (SCCC), 16-20 Nov. 2020, Coquimbo, Chile. DOI: 10.1109/SCCC51225.2020.9281203.

[13] César Baluarte-Araya, Ernesto Suarez-Lopez and Oscar Ramirez-Valdez, “Problem based Learning: An Experience of Evaluation based on Indicators, Case of Electronic Business in Professional Career of Systems Engineering” International Journal of Advanced Computer Science and Applications (IJACSA), 12(9), 2021 http://dx.doi.org/10.14569/IJACSA.2021.0120966.

[14] SAP, ¿Qué es la minería de datos?, 2024 https://www.sap.com/latinamerica/products/technology-platform/hana/what-is-data-mining.html IBM, ¿Qué es la minería de datos?, 2024. https://www.ibm.com/es-es/topics/data-mining.

[15] R. Herrera, Myriam; Ruiz, Susana; Romagnano, María; Ganga, Leonel; Lund, María Inés y Torres, Estela. “Aplicando métodos y técnicas de la ciencia de los datos a datos universitarios”. XXI Workshop de Investigadores en Ciencias de la Computación, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina, Abril 2019.

[16] Orieb Abu Alghanam, Sumaya N. Al-Khatib and Mohammad O. Hiari, “Data Mining Model for Predicting Customer Purchase Behavior in E-Commerce Context” International Journal of Advanced Computer Science and Applications(IJACSA), 13(2), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130249.

[17] Khaled M. Hassan, Mohammed Helmy Khafagy and Mostafa Thabet, “Mining Educational Data to Analyze the Student’s Performance in TOEFL iBT Reading, Listening and Writing Scores” International Journal of Advanced Computer Science and Applications(IJACSA), 13(7), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130741.

[18] Johan Calderon-Valenzuela, Keisi Payihuanca-Mamani and Norka Bedregal-Alpaca, “Educational Data Mining to Identify the Patterns of Use made by the University Professors of the Moodle Platform” International Journal of Advanced Computer Science and Applications(IJACSA), 13(1), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130140.

[19] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent and L. R. Zúnica, Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos, XIII Jornadas de Enseñanza Universitaria de la Informática, España.

[20] D. Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification, CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No 1, 2013, Sofia, Bulgaria. DOI: 10.2478/cait-2013-0006.

[21] Norka Bedregal-Alpaca, Danitza Aruquipa-Velazco, and Víctor Cornejo-Aparicio Tecnicas de Data Mining para extraer perfiles comportamiento Académico y predecir la deserción universitaria, Revista Ibérica de Sistemas e Tecnologias de Informação RISTI N.º E27, 03/2020, Pages: 592–604.

[22] Hosam Alhakami, Tahani Alsubait and Abdullah Aljarallah, “Data Mining for Student Advising” International Journal of Advanced Computer Science and Applications(IJACSA), 11(3), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110367.

[23] Abdellatif HARIF and Moulay Abdellah KASSIMI, “Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques” International Journal of Advanced Computer Science and Applications(IJACSA), 15(7), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150723.

[24] Xi LU, “Modern Education: Advanced Prediction Techniques for Student Achievement Data” International Journal of Advanced Computer Science and Applications(IJACSA), 15(1), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01501126.

[25] Martínez-Abad, Fernando; Hernández-Ramos, Juan Pablo, Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento, REXE. Revista de Estudios y Experiencias en Educación, vol. 2, núm. Esp.2, 2018, Universidad Católica de la Santísima Concepción, Chile, Disponible en: https://www.redalyc.org/articulo.oa?id=243156768012, DOI: https://doi.org/10.21703/rexe.Especial3201812514512.

[26] Salto- Mero, J. & Cruz-Felipe, M., (2023). Análisis del Rendimiento Académico de Estudiantes de las Carreras Economía y Turismo con Power BI en los Periodos (2021). 593 Digital Publisher CEIT, 9(1), 762-772, https://doi.org/10.33386/593dp.2024.1.2162.

[27] I. Wilford R., Minería de datos: herramienta de apoyo en la selección de equipos de proyectos informáticos, Industrial/Vol. XXVII/No. 2-3/2006, Cuba, 2006.

[28] IBM, Guia de CRISP-DM de IBM SPSS Modeler, 2021. https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf.

[29] Ashraf Abdelhadi, Suhaila Zainudin and Nor Samsiah Sani, “A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments” International Journal of Advanced Computer Science and Applications(IJACSA), 13(1), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130156.

Compactness-Weighted KNN Classification Algorithm

Bengting Wan, Zhixiang Sheng*, Wenqiang Zhu, Zhiyi Hu
School of Software and IoT Engineering, Jiangxi University of Finance and Economics, Nanchang 330013, China

***Abstract*—The K-Nearest Neighbor (KNN) algorithm is a widely used classical classification tool, yet enhancing the classification accuracy for multi-feature large datasets remains a challenge. The paper introduces a Compactness-Weighted KNN classification algorithm using a weighted Minkowski distance (CKNN) to address this. Due to the variability in sample distribution, a method for deriving feature weights based on compactness is designed. Subsequently, a formula for calculating the weighted Minkowski distance using compactness weights is proposed, forming the basis for developing the CKNN algorithm. Comparative experimental results on five real-world datasets demonstrate that the CKNN algorithm outperforms eight existing variant KNN algorithms in Accuracy, Precision, Recall, and F1 performance metrics. The test results and sensitivity analysis confirm the CKNN's efficacy in classifying multi-feature datasets.**

Keywords—K-nearest neighbors; feature weight; Minkowski distance; compactness

I. Introduction

In the domains of data science and machine learning, the KNN (K-Nearest Neighbors) algorithm, widely recognized as one of the top 10 classification algorithms [1], plays a crucial role in revealing the inherent patterns and structures of data, effectively grouping data points into distinct categories, especially in market segmentation, social network analysis, bioinformatics, image processing, and other fields [2, 3]. Since Fix and Hodges [4] introduced the KNN algorithm, KNN has emerged as a classic and efficient classification tool widely applied in data mining, data classification, and other fields [5, 6]. For instance, Uddin et al. [5] have applied the KNN algorithm for disease risk prediction, and Han et al. [6] have used it to estimate the photometric redshift of quasars.

However, the KNN algorithm faces several challenges in practice, such as the choice of k value, selection of nearest neighbors, nearest neighbor search, and determination of classification rules [7]. Consequently, researchers have proposed various improvement strategies to enhance the performance of KNN [8-10]. For example, Zhang and Li [8] bolstered the classification performance of the KNN algorithm and reduced computational costs through sparse learning and group lasso techniques. Meanwhile the weighted KNN [11] approach adjusts the influence of each neighbor on the classification decision by assigning different weights, enhancing the efficiency and accuracy of classification. In this method, weights are usually based on the distance or similarity of neighbors to the query point, giving closer neighbors more significant influence in classification decisions. This strategy effectively improves the algorithm's ability to handle uneven distributions or irregular data structures. Nevertheless, weighted KNN algorithms also face challenges. Firstly, selecting and calculating appropriate weights is a crucial issue, as different weight distribution strategies directly affect the accuracy of classification results. Secondly, the algorithm may encounter efficiency issues when handling large datasets, especially in scenarios requiring real-time or near-real-time processing [12]. In response, researchers have proposed various variants of the weighted KNN algorithm, such as the Improved K-Nearest Neighbor rule combining Prototype Selection and Local Feature Weighting (IKNN_PSLFW) algorithm developed by Zhang et al. [13], which combines prototype selection with local feature weighting, and the Option out-of-bag (Opt_OOB), a KNN ensemble learning method based on feature weighting and model selection proposed by Gul et al. [14] Chen and Hao [15] introduced a KNN prediction model based on a feature weight matrix by modifying the standard Euclidean distance. Chen and Gou [16] proposed a series of weighted distance functions for classifying attributes and applied these functions to develop nearest neighbor classifiers.

However, these weighted KNN algorithms, without considering datasets with non-uniform feature distributions, still have room for improvement in classification efficiency for unevenly distributed datasets and may be limited in handling high-dimensional classification problems. Therefore, this paper will consider the compactness of data distribution and introduce a dynamic weight adjustment mechanism based on feature compactness. By reconstructing the feature weights, a new weighted KNN algorithm is proposed. The main contributions of this paper are:

- Inspired by the uneven distribution of data, a weight calculation method based on compactness is proposed;
- Inspired by the weighted KNN, the CKNN algorithm based on weighted Minkowski distance is proposed;
- In real-world datasets, the CKNN algorithm was employed for classification purposes and was subsequently compared and analyzed against recent weighted KNN algorithms. Additionally, the CKNN algorithm underwent a sensitivity analysis along with Friedman’s and Nemenyi’s post-hoc tests. These evaluations demonstrated that the CKNN algorithm possesses specific superior performance characteristics.

The rest of this paper is organized as follows: Section II elaborates on the related research work of KNN; Section III details the construction process of the CKNN algorithm; Section IV implements the CKNN algorithm and compares it with other existing variant weighted KNN algorithms; finally, the advantages and limitations of the CKNN algorithm are analyzed in this section. Finally, the paper is concluded in Section V.

II. Related Work

A. KNN

Based on similarity, the KNN classifier first identifies the nearest k neighbors of an unknown sample [4]. Then, it determines its category based on the most frequently occurring (highest probability) category among the K-Nearest Neighbor. Below is an outline of the basic principles of the KNN algorithm.

Given a set of labeled samples: $D = \{x_1, x_2, \dots, x_n\}$, a training set $D_T = \{(x_i, y_i)\}_{i=1}^{N}$ is constructed, where $x_n \in D$ is a point in n-dimensional space, y_i is the category label corresponding to x_i and N is the number of samples in the training set. For a query point q with an unknown category, KNN first calculates the Euclidean distance between this point and each sample point x_i in the training set, as shown in Eq. (1).

$$D\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{n} (x_{iv} - x_{iv})^2} \tag{1}$$

Within this framework, x_{iv} denotes the coordinate value of the i^{th}sample point along the v^{th}dimension. Subsequently, the computed distances for N points are arranged in ascending order, from which the nearest k neighbors are selected. At this juncture, the distance set comprising the k nearest neighbors to the query point q can be represented as $D_T^* = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{k}$. Subsequently, the category label of the query point q is predicted through the majority vote of its nearest neighbors, resulting in the target equation, as shown in Eq. (2).

$$predicted = argmax_{c \in C} \sum_{i=1}^{k} I(x_i = c) \tag{2}$$

In Eq. (2), C represents the set of all possible categories, k is the number of nearest neighbors, x_i is the category of the ith nearest neighbor, and $I(x_i = c)$ is an indicator function that equals 1 when y_i equals category c, and 0 otherwise. The KNN algorithm is described as shown in Algorithm 1.

```
Algorithm 1: KNN algorithm
Input: A test sample and some training samples
Output: The test sample's category
Process:
1.  For number of training samples do
2.      calculate the similarity between the test sample    and a
        training sample
3.  End for
4.  find the k training samples that are most like the test sample
5.  determine test sample's category
```

However, the classical KNN algorithm has faced several challenges: (1) Noise sensitivity. The KNN algorithm determines a sample's category based on the k nearest neighbors' labels, making it sensitive to noise and outliers. (2) Redundant sample testing. The nearest neighbors are searched for each test sample given, but this is not necessarily optimal for all test cases. (3) Dependence on the hyperparameter k. The performance of the algorithm varies with different k parameters. (4) Poor stability. The algorithm performs well on some datasets and poorly on others.

To address these issues, researchers have proposed various KNN variants. To tackle the challenge of classifying imperfect data in high-dimensional spaces, Gong et al. [17] improved the traditional KNN method by synchronizing neighborhood search and feature weighting, proposing an enhanced KNN algorithm—AEKNN. Bian et al. [18] improved the traditional fuzzy KNN [19] by adaptively selecting the optimal number of nearest neighbors (the k value) for each test sample Gou et al. [20] optimized the classifier by capturing the proximity and geometric characteristics of the K-nearest neighbors and learning the contribution of each neighbor to the classification of the test sample through linear representation methods, thereby reducing the algorithm's sensitivity to k and enhancing its performance.

B. Weighted Distance KNN

In response to issues such as class overlap and the difficulty in choosing the k factor, researchers have embarked on explorations and achieved many research results. For instance, Zhang et al. [13] have developed the IKNN_PSLFW algorithm, which combines prototype selection with local feature weighting. In this method, the prototype selection part divides the training set into multiple pure subsets, where each subset contains instances of only one class label. Following this, the scope of prototype selection and the weights of features are updated through local feature weighting, optimizing the objective function as illustrated in Eq. (3).

$$rw_s = max_{j=1,2,\dots,n} \sqrt{\sum_{j=1}^{d} w_s^j \left(w_s^j - x_i^j\right)^2} \tag{3}$$

In Eq.(3), $w^j = u^j / \sum_{j=1}^{d} u^j$, $u^j = \begin{cases} \frac{1}{v^j}, (v^j \neq 0) \\ 10 \times max\left\{\frac{1}{v^j} \middle| i \neq j, v^j \neq 0\right\}, (v^j = 0) \end{cases}$, and v^j is the variance of the j^{th} feature in the subset. Ultimately, a representative example is chosen from each subset as a prototype, and both the boundary of the subset and the total count of instances it includes are recorded. During the classification phase, depending on the weighted distance between an unknown instance and each prototype, three potential scenarios are identified: the instance falls within the range of a single prototype, within an overlap area, or outside the range of all prototypes, with different rules applied to predict its category accordingly. Throughout the process, there is no need to predetermine the value of k, and the time complexity of the IKNN_PSLFW algorithm is $o(n^2)$. This method effectively reduces the number of instances and overlap areas, thereby enhancing the accuracy and efficiency of classification.

Gul et al. [14] introduced an ensemble learning method for the k nearest neighbor, termed Opt_OOB, predicated on feature weighting and model selection. This method selects the best model by leveraging the out-of-bag prediction error to assemble the ultimate ensemble classification model. This methodology establishes an objective function utilizing the distance Eq. (4).

$$D_w\left(X'_{1\times p'}, X_{1\times p'}\right) = \left\{\sum_{j=1}^{p'} w\left(x'_j - x_j\right)^2\right\}^{\frac{1}{2}} \tag{4}$$

In Eq. (4), feature weights w are determined by $w = argmax_w[min\{d_H(\psi(x))\}]$, to identify Out-of-Bag (OOB) observations during the bootstrap sampling process. Subsequently, the prediction error for each base model on its corresponding OOB observations is calculated. Models are then ranked according to the magnitude of OOB error, and a certain proportion of the best-performing models are selected to form the final ensemble classification model. This method reduces the dependency on the parameter k and ensures the diversity and accuracy of the ensemble model, although its performance may decrease on tiny datasets.

Chen and Hao [15] have proposed a K-nearest neighbors predictive model based on a feature weighting matrix by modifying the standard Euclidean distance. The core of this algorithm lies in improving prediction accuracy by altering the positional relations of sample points. The feature weighting matrix is shown in Eq. (5).

$$P = \left\{\begin{matrix} InfoGain(f_1) & 0 & \dots & 0 \\ 0 & InfoGain(f_2) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & InfoGain(f_n) \end{matrix}\right\} \tag{5}$$

In Eq. (5), $InfoGain(A) = \sqrt{Info(\mathrm{D}) - Info_A(D)}$, $Info(D) = -\sum_{i\in\{-1,\ +1\}} \frac{|C_{\{i,D\}}|}{|D|} log(\frac{|C_{\{i,D\}}|}{|D|})$, and $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info\left(D_j\right)$. Here, D represents the dataset, $|D|$ denotes the size of the dataset, and $C_{\{i,D\}}$ represents the subset of the dataset D that belongs to class C_i. This method has certain advantages for large-scale or complex datasets, but the effectiveness of the model depends on the accuracy of the feature weighting matrix, which requires sufficient prior knowledge or data analysis to determine appropriate weights.

In pursuit of an optimal distance metric for precisely quantifying the dissimilarities among classified samples, Chen and Gou [16] introduced a series of weighted distance functions tailored for categorical attributes, which have been applied to advance nearest neighbor classifiers. The Global Gini K-nearest neighbors (GGKNN) incorporate a weighting scheme as depicted in Eq. (6).

$$\omega_d^{(GG)} = e^{-\frac{M}{M-1}\sum_{s_d\in s_d} P(S_d)\times GG(S_d)} \tag{6}$$

In Eq. (6), $GG(s_d) = -\sum_{m=1}^{M} P(m|s_d) \log_2 p(m|s_d)$, $p(s_d) = \frac{1}{N}\sum_{(X,y)\in tr} I(x_d = s_d)$,and $p(m|s_d) = \frac{\sum_{(X,y)\in c_m} I(x_d=s_d)}{\sum_{(X,y)\in tr} I(x_d=s_d)}$. Herein, tr represents the training dataset, M denotes the number of classes contained within the training dataset, $|s_d|$indicate the discrete values of the d^{th} attribute. The weighting for Global Entropy K-nearest neighbors (GEKNN) is shown in Eq. (7).

$$\omega_d^{(GE)} = e^{-\frac{1}{log_2 M}\sum_{s_d\in s_d} P(S_d)\times GE(S_d)} \tag{7}$$

In Eq. (7), $GE(s_d) = 1 - \sum_{m=1}^{M}[P(m|s_d)]^2$. These methods use global statistical approaches to weight attributes, considering the information from all data points to determine the importance of each attribute. Conversely, the weighting for Local Gini K-nearest neighbors (LGKNN) is shown in Eq. (8).

$$\omega_{md}^{(LG)} = e^{-\frac{|s_d|}{|s_d|-1}\times LG(m,d)} \tag{8}$$

In Eq. (8), $LG(m,d) = 1 - \sum_{s_d\in s_d}[P(s_d|m)]^2$,$p(s_d|m) = \frac{1}{|c_m|}\sum_{(X,y)\in c_m} I(x_d = s_d)$,and c_m represents the m^{th} class within the training dataset tr. The weighting for Local Entropy K-nearest neighbors (LEKNN) is shown in Eq. (9).

$$\omega_{md}^{(LE)} = e^{-\frac{1}{log_2|s_d|}\times GE(m,d)} \tag{9}$$

In Eq. (9), $LE(m,d) = -\sum_{s_d\in s_d}(s_d|m) \log_2 p(s_d|m)$. This employs a local method for computing weights, meaning that it adjusts attribute weights based on the local information surrounding each data point. This method achieves soft feature selection for categorical data, thereby improving the quality of classification.

Furthermore, researchers have introduced multiple variations of the KNN algorithm that utilize weighted distance measures for optimization. Açıkkar and Tokgöz [21] have enhanced the conventional KNN algorithm by introducing a new weighted voting mechanism and adaptive k-value selection techniques. These modifications have improved the performance of the KNN algorithm in scenarios with complex or nonlinear decision boundaries, especially in the context of processing datasets with noise or outliers.

III. Compactness-Weighted K-NN Classification Algorithm

This section delineates an improved K-Nearest Neighbors algorithm (CKNN) predicated on compactness and local feature weighting, devised to augment the classification efficacy of the KNN algorithm. The research methodology unfolds in two pivotal steps: Initially, the compactness for each feature is ascertained, forming the groundwork for the recalibration of feature weights. After that, a CKNN algorithm, hinged on compactness, is introduced.

In most pattern recognition tasks, the relevance of different features differs, particularly in classification tasks. Even if all features in the dataset are relevant, their degrees of relevance may vary. To address this, we propose the concept of feature compactness, as illustrated in Fig. 1.

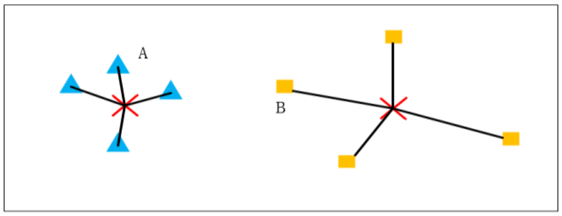


Fig. 1. Feature Compactness, with Feature A having greater compactness than Feature B.

Fi. 1 shows that the sum of distances between the elements in set A and their centroid is significantly less than that in set B, leading to the conclusion that feature A possesses greater compactness than B Various distance metrics are utilized to measure feature compactness, such as Euclidean distance, Manhattan distance, and Minkowski distance. The Minkowski distance [22], in particular, allows for adjusting the parameter p according to different scenarios and is widely employed. Inspired by this, the article adopts the Minkowski distance to measure the distances of features. Within a dataset $X = \{x_1, x_2, \dots, x_n\}$ comprising n samples, each with m features, x_i represents a feature vector of dimension m. Assuming the centroid vector c represents the arithmetic mean of all sample point feature vectors, the compactness c_j for the j^{th}feature, based on the Minkowski distance, is defined as the Minkowski distance between the values of all sample points for that feature and the value of the centroid for that feature. The calculation is as shown in Eq. (11).

$$C_j = \sum_{j=1}^{n} \left| x_{ij} - c_j \right|^p \tag{11}$$

In Eq. (11), n represents the total number of samples in the cluster, C_j denotes the compactness of the j^{th} feature, x_{ij} is the value of the j^{th} feature for the i^{th} sample, C_j is the value of the j^{th} feature of the cluster centroid, and p is the exponent parameter of the Minkowski distance. From Eq. (11), it is inferred that, under the conditions of a given number of samples and a defined centroid, a smaller value of C_j indicates greater compactness, and vice versa.

Specifically, for a given dataset and its corresponding centroid, the initial step involves calculating the difference between each data point and the centroid across all dimensions. Subsequently, these differences are raised to the p^{th} power using the Minkowski formula, where p is a predefined parameter. The steps for solving compactness will be detailed in Algorithm 2.

```
Algorithm 2: Calculate Compactness (CL, P)
Input:  CL: feature vector
        P: Minkowski index
Output: A one-dimensional array containing the disper- sion
        of each feature $S = \{s_1, s_2, \dots, s_k\}$
Process:
1.  Set $C$←The arithmetic mean of the eigen vectors of all
    points in the cluster
2.  For each feature in the feature space do
3.      Add $S$←Discrete degree calculated by Eq. (11)
4.  End for
```

To address the discrepancy that arises from the assumption in the canonical KNN algorithm, where each feature is assigned an equal weight reflecting an assumption of equal contribution to the decision-making process—a scenario often divergent from real-world applications where the importance of features can vary significantly. This study introduces a methodology grounded in compactness to determine the weights of different features. Inspired by the findings in study [23] and assuming a given dataset is presumed to contain K categories, with each category corresponding to a distinct cluster, this paper proposes a novel objective function. This function represents the sum of weighted averages across different classification sets, as calculated in Eq. (12).

$$\text{J} = \sum_{c=1}^{K} \sum_{v=1}^{V} w_{cv}^{\beta} C_{cv} \tag{12}$$

where, K represents the total number of categories, V represents the total number of features, w_{cv} is the weight of the v^{th} feature in the c^{th} category, and β is a weight adjustment parameter. The adjustment parameter β is used to control the extent of the weight influence. When β>1, it indicates a higher emphasis on features with high weights, when β=1, the model degenerates to a traditional equal-weight model. C_{cv} is an indicator measuring the compactness of the v^{th} feature in the c^{th} category, calculated by $\sum_{v=1}^{n} ||x_{cv} - c_v||^p$,, where x_{cv} is the v^{th} feature value of the c^{th} sample, and n represents the total number of samples in the cluster.

By assigning different weights to various classes, the goal is to minimize the weighted average compactness within each class, thereby improving the compactness of classification. To this end, Eq. (12) can be converted into Eq. (13). Within this framework, the degree of classification compactness can be obtained, and the weights of each feature, w_{cv}, can be solved.

$$\sum_{c=1}^{K} \sum_{v=1}^{V} w_{cv}^{\beta} C_{cv} = \sum_{c=1}^{K} \sum_{v=1}^{V} \{ w_{cv}^{\beta} \sum_{v=1}^{n} ||x_{cv} - c_v||^p \} \tag{13}$$

Considering the weights of different features satisfy the constraints: $\sum_{v=1}^{V} w_{cv} = 1$ and w_{cv}≥0, it is evident that Eq. (13) represents a nonlinear programming equation while also satisfying convex function constraints. To enhance the compactness within classes by optimizing feature weights, the Lagrangian function L is employed to minimize Eq. (14):

$$L = \sum_{v=1}^{V} w_{cv}^{\beta} C_{cv} + \lambda \left(1 - \sum_{v=1}^{V} w_{cv}^{\beta} \right) \tag{14}$$

Taking the partial derivative of w_{cv}^{β} in Eq. (14), and then setting it to zero to find the extremum, as shown in Eq. (15):

$$\frac{\partial L}{\partial w_{cv}} = \beta w_{cv}^{\beta - 1} C_{cv} - \lambda = 0 \tag{15}$$

Solving Eq. (15) yields the weight w_{cv}, as shown in Eq. (16):

$$w_{cv} = \left(\frac{\lambda}{\beta C_{cv}} \right)^{\frac{1}{\beta - 1}} \tag{16}$$

Given the weight constraints $\sum_{v=1}^{V} w_{cv} = 1$ and w_{cv}≥0, Eq. (16) can be further derived to obtain Eq. (17).

$$\sum_{v=1}^{V} \left(\frac{\lambda}{\beta C_{cv}} \right)^{\frac{1}{\beta - 1}} = 1 \Leftrightarrow \left(\frac{\lambda}{\beta} \right)^{\frac{1}{\beta - 1}} = \frac{1}{\sum_{v=1}^{V} \left(\frac{1}{C_{cv}} \right)^{\frac{1}{\beta - 1}}} \tag{17}$$

Simplifying Eq. (17) gives the formula for solving weight w_{cv}, as shown in Eq. (18). From Eq. (18), it can be seen that under compact classification, the weight of feature v in category C can be obtained by solving the Minkowski distance.

$$w_{cv} = \frac{1}{\sum_{u=1}^{V} \left(\frac{C_{cv}}{C_{cu}} \right)^{\frac{1}{\beta - 1}}} \tag{18}$$

Building on Eq. (18), it can be determined that the weights of various features can be calculated given a classification. However, in the KNN classification process, both the classification and the weights are the objectives to be determined. Inspired by the varying importance of different features and the concept of compactness as discussed in references, a weighted Minkowski distance based on compactness weights is proposed, as shown in Eq. (19).

$$d_w\left(x_i, x_j\right) = \sqrt[p]{\sum_{v=1}^{V} w_{cv}\left(x_{iv} - x_{jv}\right)^p} \tag{19}$$

In Eq. (19), for given data samplesx_i,x_j∈C, where β is a user-defined parameter, $w_{cv}$is the weight of feature weight v. In this case, the weight of each feature no longer depends on a specific cluster but is based on the feature distribution across the entire dataset. Feature weights should be non-negative and satisfy $\sum_{v=1}^{V} w_{cv} = 1$, and $w_{cv} \geq 0$.

Leveraging the concept of compactness-weighted distances within the framework of the KNN algorithm, this section introduces the Compactness-weighted KNN (CKNN) algorithm. The CKNN algorithm begins by calculating the compactness of each feature in the dataset according to Eq. (11). This calculation necessitates using the Minkowski distance measure to ascertain the compactness of each feature relative to its centroid. Drawing on the principle of compactness, the weights for each feature can be determined using Eq. (18). Subsequently, a compactness-weighted Minkowski distance, as delineated in Eq. (19), is constructed to facilitate the computation of distances between samples. Ultimately, the CKNN algorithm replaces the Euclidean distance traditionally employed in KNN with the weighted Minkowski distance, selects the K-nearest neighbors based on this distance, and utilizes a voting mechanism predicated on the category labels of these neighbors to ascertain the category of the target sample. The steps to implement the CKNN algorithm are outlined in Algorithm 3.

Algorithm 3: Compactness-weighted KNN algorithm

```
Input:  $D_T$ :  Training data set,
        $K$:  The number of nearest neighbors
        $T$:  Test Dataset
        $\beta$:  Minkowski index
Output: $Y$:  classification result
Process:
Set $Y \leftarrow \emptyset$, $C$←calculateCompactness(CL,P), $w \leftarrow \frac{1}{V}$
For The weight of each feature $w$ do
    Update the weight $w_{cv}$ of each feature through Eq. (18)
End For
For Each sample in the test dataset $T$ do
    Set list ←  $\emptyset$
For Training data set $D_T$ do
    Compute weighted Minkowski distance by Eq. (19).
    Add distance to list
    Assign test sample category by majority vote from K nearest neighbors.
    Add Classification Results to Y
End For
Return $Y$ as the classification results for all samples in $T$
```

IV. Algorithm Implementation

This section elucidates the datasets employed by the CKNN algorithm, the performance evaluation metrics utilized, and an analysis of the experimental outcomes. The experiments were conducted on a computer with a 12th Gen Intel(R) Core (TM) i7-12700H CPU, clocked at 2.70GHz, and 16.0GB RAM, running the Windows 11 operating system. The Python3.10 programming language executed the implementation.

A. Dataset and Evaluation Metrics

The implementation adopted five datasets from the UCI Machine Learning Repository (Wine, Breast Cancer, Promoters, Mc2, Car) as benchmark datasets (http://archive.ics.uci.edu/). Table I shows the essential characteristics of the five datasets, including the total number of samples, the number of features, and the number of classes. For datasets with some features as strings, traditional label encoding methods will be used. The Car dataset comprises 1728 samples, representing a multi-sample dataset. The Wine, Mc2, Promoters, and Breast Cancer datasets consist of 13, 39, 57, and 30 feature attributes, thus categorizing them as high-dimensional datasets. Wine and Car datasets have three and four categories, respectively.

To evaluate the classification results, this paper uses four evaluation metrics: Accuracy, Recall, Precision, and F1 (F1-measure) to measure the performance of algorithms. Among them, Recall refers to the ratio of correctly predicted positive instances to positive instances; precision refers to the ratio of correctly predicted positive instances to optimistic predictions. Based on the F1 measure, the experiments used macro-F1 (Macro-F1, the average F1 values within classes) for evaluation [22]. All these indicators range from [0, 1], with values closer to 1 indicating better model performance. Accuracy, precision, recall, and F1 are shown in Eq. (20), (21), (22), and (23) respectively. TP, TN, FP, and FN represent the proportions of true positives, true negatives, false positives, and false negatives in the result data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$Precision = \frac{TP}{TP+FP} \tag{21}$$

$$Recall = \frac{TP}{TP+FN} \tag{22}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall+Precision} \tag{23}$$

TABLE I. BASIC CHARACTERISTICS OF THE DATASETS USED IN THE EXPERIMENT

Item	Dataset	Instances	Features	Classes
1	Wine Dataset	178	13	3
2	Mc2 Dataset	161	39	2
3	Car Dataset	1728	6	4
4	Promoters Dataset	106	57	2
5	Breast Cancer Dataset	699	30	2

B. Analysis of Results

The article selects eight existing improved KNN classification algorithms (KNN [4], FWKNN [15], LEKNN [16], LGKNN [16], GEKNN [16], GGKNN [16], IKNN_PSLFW [13], and Opt_OOB [14]) for comparison with CKNN. Among them, KNN represents the classic K-Nearest Neighbor algorithm. FWKNN is a K-Nearest Neighbor prediction model based on a feature weighting matrix. LEKNN calculates feature weights through local entropy, while LGKNN calculates feature weights through local Gini. GEKNN uses global entropy to calculate feature weights, and GGKNN uses global Gini for the same purpose. IKNN_PSLFW is an ensemble learning method based on prototype selection combined with local feature weighting, and Opt_OOB is a K-Nearest Neighbor ensemble learning method based on feature weighting and model selection. The implementation results are shown in Table II (the highest values for each dataset are indicated in bold).

Table II shows that CKNN exhibits superior performance, especially on the Promoters dataset, where its accuracy reached 0.8439, significantly higher than other algorithms. After comparing the performances of different algorithms across multiple datasets, it was observed that the proposed method demonstrates superiority in all evaluation metrics. Specifically, on the Wine dataset, compared to the Opt_OOB algorithm, CKNN showed improvements of 3.71% in Accuracy, 3.69% in the Recall, 3.42% in Precision, and 3.64% in F1 score, the improvement on the MC2 dataset was even more significant, with CKNN surpassing the LEKNN algorithm by 2.04% in Accuracy. Although the improvements in Recall, Precision, and F1 score were closer, they still reflected our algorithm's advantage. On the Car dataset, compared to the FWKNN algorithm, the improvement was particularly notable, with increases of 1.54% in Accuracy, 2.81% in Recall, 11.17% in Precision, and 7.06% in F1 score. On the Promoters dataset, compared to the second-ranked KNN, there were increases of 6.26% in Accuracy, 6.27% in Recall, 6.25% in Precision, and 6.26% in F1 score. Regarding the Breast Cancer dataset, the method also demonstrated its superior performance. Compared to the IKNN_PSLFW algorithm, CKNN improved by 0.58% in Accuracy, 0.47% in Recall, 0.81% in Precision, and 0.62% in F1 score.

To further comprehensively evaluate the performance of the CKNN algorithm, based on the implementation results in Table II, the following will analyze the Sum of Ranking Differences (SRDs) [24], Friedman test [26], Nemenyi test [27], and Bonferroni correction [28].

First, a comparative analysis of the Sum of Ranking Differences (SRDs) was conducted, a multi-criteria decision-making method that achieves evaluation objectives by calculating the sum of absolute differences between each algorithm's actual rankings and reference rankings. Table II presents the values of four evaluation metrics for various algorithms across five datasets; according to the SRDs method, the reference vector contains 20 elements, each of which is the best score among the algorithms. After scaling the SRD values to the [0, 100] interval, their theoretical distribution approximates a normal distribution. Thus, the normal quantiles of each algorithm can serve as the actual SRD values compared to the reference vector, with the implementation results shown in Fig. 2. The scaled Sum of Ranking Differences (SRD) values are plotted on the x-axis and the left y-axis, while the right y-axis displays the relative frequency (black curve). The Gaussian fitting parameters are m=66.72, s=9.87. The SRD values at the 5% probability level (XX1), the median (Med), and 95% (XX19) are also provided.

TABLE II. COMPARISON OF NINE ALGORITHMS ON DIFFERENT DATASETS

Dataset	Methods	Accuracy	Recall	Precision	F1	RANK
Wine	Proposed	**0.963**	**0.963**	**0.961**	**0.963**	**1**
	KNN	0.740	0.726	0.726	0.726	3
	FWKNN	0.740	0.726	0.726	0.726	4
	LEKNN	0.537	0.532	0.538	0.534	8
	LGKNN	0.648	0.644	0.652	0.644	7
	GEKNN	0.444	0.425	0.488	0.402	9
	GGKNN	0.648	0.645	0.680	0.656	6
	IKNN	0.740	0.730	0.723	0.722	5
	Opt_OOB	0.925	0.926	0.926	0.926	2
Mc2	Proposed	**0.714**	0.613	**0.613**	**0.613**	**1**
	KNN	0.612	0.546	0.537	0.534	6
	FWKNN	0.612	0.546	0.537	0.534	7
	LEKNN	0.693	0.600	0.595	0.597	2
	LGKNN	0.653	0.516	0.517	0.517	8
	GEKNN	0.673	0.558	0.558	0.558	4
	GGKNN	0.673	0.558	0.558	0.558	4
	IKNN	0.673	**0.614**	0.596	0.600	3
	Opt_OOB	0.693	0.543	0.554	0.545	5
Car	Proposed	**0.942**	**0.859**	**0.909**	**0.879**	**1**
	KNN	0.859	0.633	0.766	0.681	4
	FWKNN	0.926	0.831	0.797	0.809	2
	LEKNN	0.778	0.419	0.432	0.415	8
	LGKNN	0.724	0.465	0.606	0.493	6
	GEKNN	0.791	0.485	0.528	0.500	5
	GGKNN	0.791	0.485	0.528	0.500	5
	IKNN	0.774	0.390	0.543	0.414	7
	Opt_OOB	0.890	0.667	0.849	0.715	3
Promoters	Proposed	**0.843**	**0.845**	**0.843**	**0.843**	**1**
	KNN	0.781	0.782	0.781	0.781	2
	FWKNN	0.750	0.752	0.752	0.750	3
	LEKNN	0.718	0.727	0.741	0.716	5
	LGKNN	0.718	0.723	0.726	0.718	6
	GEKNN	0.656	0.660	0.662	0.655	8
	GGKNN	0.656	0.660	0.662	0.655	8
	IKNN	0.687	0.690	0.690	0.687	7
	Opt_OOB	0.750	0.749	0.749	0.749	4
Breast Cancer	Proposed	**0.976**	**0.971**	**0.978**	**0.974**	**1**
	KNN	0.941	0.933	0.933	0.936	5
	FWKNN	0.941	0.933	0.933	0.936	4
	LEKNN	0.614	0.562	0.570	0.561	9
	LGKNN	0.713	0.660	0.695	0.666	7
	GEKNN	0.731	0.674	0.722	0.682	6
	GGKNN	0.660	0.619	0.629	0.621	8
	IKNN	0.970	0.966	0.970	0.968	2
	Opt_OOB	0.953	0.946	0.952	0.949	3

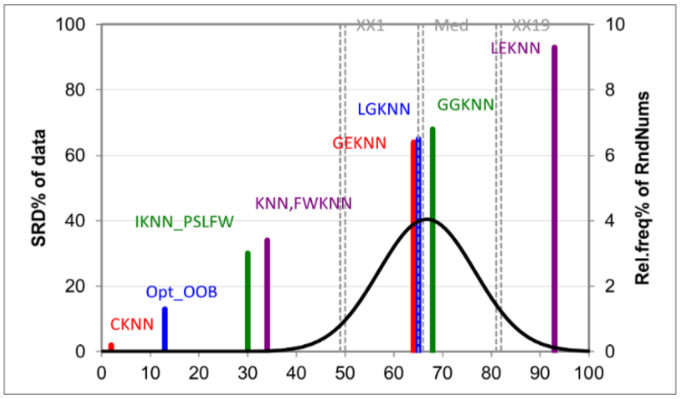


Fig. 2. Evaluation of algorithms using the sum of rank differences.

As can be discerned from Fig. 2, the CKNN algorithm is positioned on the left side of the curve, indicating that CKNN is the algorithm closest to the ideal state. At the same time, CKNN is at a certain distance compared to Opt_OOB and IKNN_PSLFW, signifying a clear advantage of CKNN over Opt_OOB and IKNN_PSLFW. Moreover, aside from GEKNN, LGKNN, GGKNN, and LEKNN, the ranking of the remaining five algorithms shows a significant difference from random ranking (α=0.05).

To highlight the advantages of CKNN, this paper further conducts a Friedman test [26]. Based on the Accuracy, Precision, Recall, and F1 metrics of CKNN, Table III presents the Friedman statistic F_F and the corresponding p-values for KNN, FWKNN, LEKNN, LGKNN, GEKNN, GGKNN, IKNN_PSLFW, and Opt_OOB in terms of accuracy, precision, recall, and F1 metrics. Table III shows that the null hypothesis (i.e., all compared algorithms will have equivalent performances) is significantly rejected at the significance level of α=0.05 for each evaluation metric, meaning there is a significant difference between CKNN and the other algorithms. However, it does not specify which algorithms are superior or inferior.

To further observe the differences among algorithms, this paper uses the Nemenyi test to assess the competitiveness of algorithms. In this test, if the difference in average ranks between two classifiers reaches at least the critical difference $CD=q_{\alpha}\sqrt{\frac{k(k+1)}{6N}}$, it is considered that there is a significant difference in performance between these two classifiers. At a significance level of α=0.05, q_{α} is 3.102, and the CD value is 5.369 (where k=9，N=5). Fig. 3 presents the CD diagram of the nine algorithms under Accuracy, Precision, Recall, and F1 metrics. In Fig. 2, any algorithm whose average rank is within a CD interval of CKNN is highlighted with a red line to show its association; otherwise, it indicates a significant performance difference from CKNN. For example, in recall, CKNN's average rank is 1.20, and with the addition of the CD value, the critical value becomes 6.57. At this point, LGKNN and GEKNN, with average ranks of 7.20 and 6.70 respectively, perform poorly. However, for algorithms within the CD interval, it is currently not impossible to determine the performance gap between them and CKNN.

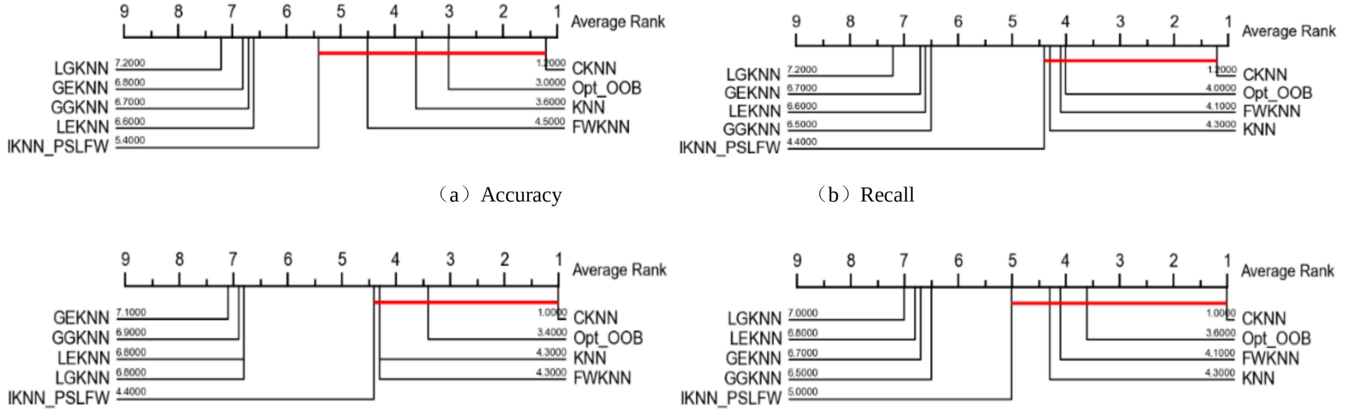
Based on the Nemenyi test, this paper uses the Bonferroni correction [27] to control the type I error (i.e., falsely rejecting a true null hypothesis). Let $\Delta_{\xi}=\bar{\xi}_{algorithm}-\bar{\xi}_{CKNN}$, when Δ_{ξ} is more excellent than CD_{α}, it is marked with "Y", indicating that CKNN outperforms the corresponding algorithm on the respective metric; otherwise, it is not marked. At a significance level of α=0.05, the critical value q_{α} becomes 2.724.

As shown in Table IV, the Bonferroni assessment results indicate that CKNN's performance exceeds that of LEKNN, LGKNN, GGKNN, and GEKNN algorithms.

TABLE III. SUMMARY OF THE FRIEDMAN STATISTIC F_F (K = 9, N = 5)

Evaluation Criteria	F_F	Critical Value (α=0.05)
Accuracy	22.86	15.51
Recall	19.96	
Precision	23.16	
F1 score	21.30	

Note: k represents the number of algorithms being compared; N represents the number of datasets



（a）Accuracy

（b）Recall

（c）Precision

（d）F1 score

Fig. 3. Nemenyi test of CKNN (control algorithm) with other variant KNN algorithms.

TABLE IV. COMPARISON OF CKNN WITH OTHER VARIANT KNN ALGORITHMS

	Accuracy	**Recall**	**Precision**	**F1**
KNN	--	--	--	--
Opt_OOB	--	--	--	--
FWKNN	--	--	--	--
IKNN_PSLFW	--	--	--	--
LEKNN	Y	Y	Y	Y
LGKNN	Y	Y	Y	Y
GEKNN	Y	Y	Y	Y
GGKNN	Y	Y	Y	Y

Confidence intervals [25] are employed to assess the degree of performance improvement among different algorithms. This paper utilizes confidence intervals to evaluate the performance of CKNN against eight compared variant KNN algorithms. Confidence intervals for comparisons among the eight algorithms were constructed to quantify these differences, assuming normality for the ranking differences as depicted in Eq. (24).

$$\frac{\Delta\xi}{\sqrt{\frac{k(k+1)}{6N}}} \sim N(0,1) \qquad (24)$$

At a 95% confidence level, Fig. 4 shows the confidence intervals for Accuracy, Recall, Precision, and F1 metrics for the nine algorithms. From Fig. 4, it is observed that except for KNN, IKNN_PSLFW, and FWKNN, all intervals for Opt_OOB, LEKNN, LGKNN, GGKNN, and GEKNN appear to be less than 0, indicating significant differences between these algorithms and CKNN. For KNN, IKNN_PSLFW, and FWKNN, although the upper bounds of some evaluation metrics' confidence intervals are more significant than or close to 0, the estimated parameter values within the confidence intervals remain below 0, suggesting that CKNN, on the whole, outperforms KNN, IKNN_PSLFW, and FWKNN, with IKNN_PSLFW showing the closest performance to CKNN.

From the analyses based on the Sum of Ranking Differences (SRDs), Friedman test, Nemenyi test, and Bonferroni correction, it is evident that the CKNN algorithm outperforms the compared algorithms, including KNN, FWKNN, LEKNN, LGKNN, GEKNN, GGKNN, IKNN_PSLFW, and Opt_OOB in terms of performance.

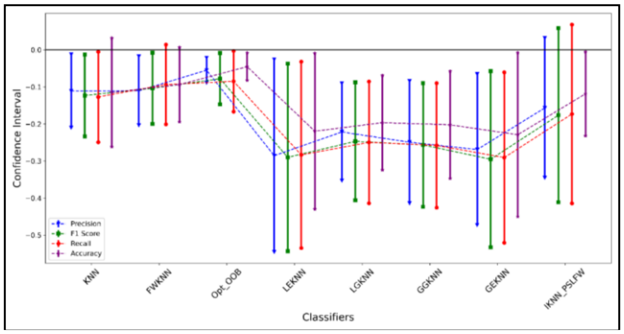
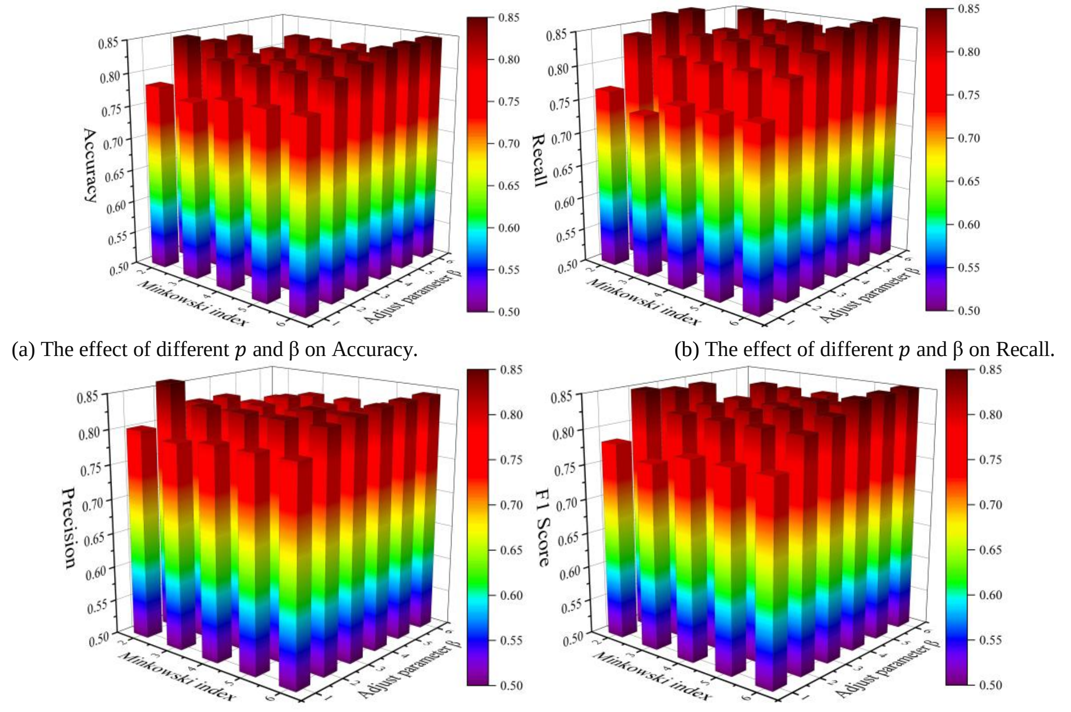


Fig. 4. Confidence intervals for rank differences.

C. Sensitivity Analysis

This section, using the Promoters dataset as an example, will analyze the impact of the Minkowski exponent (p-value) and the tuning parameter β on the performance of the proposed CKNN algorithm. The importance of p and β values in affecting the classifier's performance will be demonstrated through specific experimental results, which are displayed in Fig. 5.



(a) The effect of different p and β on Accuracy.

(b) The effect of different p and β on Recall.

(c) The effect of different p and β on Precision.

(d) The effect of different p and β on F1 score.

Fig. 5. The effect of different Minkowski indices and tuning parameter β on classifier performance.

From Fig. 5, it can be observed that (1) For a specific value of β, the trend of accuracy increasing with an increase in p is quite apparent. The highest accuracy combination occurs at β=6 and p=4, 5, 6, with accuracies all reaching 0.8438. This indicates that a higher combination of β and p values is more likely to produce higher accuracies in this data group. Despite some fluctuations, a general trend can still be seen accuracy tends to increase with an increase in the value of p. The effect of the tuning parameter β seems less direct. However, it can be observed that when the value of the tuning parameter β reaches 6, the accuracy reaches a higher level, especially at higher p values. (2) Recall rates show a certain upward trend with the p increase. Especially at p=3 and subsequent values, recall rates are relatively high, notably at β=2 and β=5, 6, indicating that an increase in p has a positive effect on enhancing recall rates. At β=2, 5, 6 and p=3,4, the recall rates all reached the highest value of 0.8824. Overall, as p increases, there is a trend for an increase in recall rates, although this trend exhibits some fluctuations under different β values. (3) At β =2, precision increases significantly with p, reaching a peak (0.875), then decreasing. For other β values, precision does not vary much across different p values, but overall, when β increases to 6, precision reaches its highest at p=4, 5, 6. This suggests that larger values of β and p might be more beneficial for increasing precision in this specific model. Generally, precision tends to improve with an increase in β, especially at higher p values. (4)F1 score varies under different combinations of β and p. Especially at β=2, the F1 score corresponding to p significantly surpasses other p values, showing the highest score at 0.8485. At β=6 and p=4, 5, 6, the highest F1 scores were observed, each being 0.8571. This finding aligns with previous analyses of precision, suggesting that the model's overall performance may be better with larger values of β and p.

Therefore, CKNN performance metrics (Accuracy, Recall, Precision, and F1 score) generally improve with the increase of the parameter p and perform optimally at larger β values. Reasonable adjustment of the Minkowski index and the tuning parameter β can further optimize the classification performance of the CKNN algorithm.

V. Conclusion

This study proposes an improved K-nearest neighbor (KNN) classification algorithm based on compactness weights, which initially updates feature weights by calculating the compactness of each feature and then employs a compactness-weighted Minkowski distance to calculate the distances between samples, serving as the basis for classification decisions. Experimental results indicate that the CKNN algorithm surpasses traditional KNN and variant KNN algorithms in Accuracy, Recall, Precision, and F1 scores across the selected five datasets, notably showing significant performance improvements on the Promoters dataset.

The analysis of experimental results suggests that when the Minkowski exponent is two and the tuning parameter β is 2, the CKNN algorithm achieves relatively better classification effects. The CKNN algorithm can better balance the local and global information between samples, enhancing classification accuracy. Additionally, the overall results from the SRDs ranking, Friedman test, Nemenyi test, and Bonferroni correction analysis of the CKNN algorithm are superior to those of the compared variant KNN algorithms, confirming the better performance of the CKNN algorithm. Sensitivity analysis results indicate that the performance of the CKNN algorithm is jointly influenced by the Minkowski exponent and the tuning parameter β, and an appropriate selection of these parameters can further enhance the algorithm's performance.

Although the CKNN algorithm proposed in this study enhances the performance of the KNN algorithm, the performance of the KNN algorithm performance remains a key area of research. Therefore, future research will focus on further enhancing the scalability of the K-Nearest Neighbors (KNN) algorithm in large-scale datasets and real-time applications, with an emphasis on exploring parallel processing and distributed computing technologies to improve the efficiency of KNN in big data scenarios. At the same time, by combining the ability of deep learning models to automatically extract features and optimize weights, the KNN algorithm is expected to perform more effectively in handling high-dimensional and unstructured data.

Acknowledgment

This paper is upheld by the Ministry of Education Humanities and Social Sciences Planning Fund Project under Grant 22YJA880051, and in by National Science Foundation of China under Grants 72261016， in part by the Department of Education of Jiangxi Province of China under Grant GJJ2200535, 22YB052.

Data availability

Data will be made available on request.

Conflicts of Interest

These authors state that there have been no competing interests among them.

References

[1] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019, pp. 1255-1260.

[2] Z. Li, H. Wang, S. Zhang, W. Zhang, and R. Lu, “SECKNN: FSS-Based Secure Multi-Party KNN Classification under General Distance Functions,” IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1326–1341, Jan. 2024.

[3] M. M. Abualhaj, A. A. Abu-Shareha, Q. Y. Shambour, A. Alsaaidah, S. N. Al-Khatib, and M. Anbar, “Customized K-nearest neighbors’ algorithm for malware detection,” International Journal of Data and Network Science, vol. 8, no. 1, pp. 431–438, Jan. 2024.

[4] E. Fix and J. L. Hodges, “Discriminatory analysis: Nonparametric discrimination: Consistency properties,” PsycEXTRA Dataset. Jan. 01, 1951.

[5] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, “Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction,” Scientific Reports, vol. 12, no. 1, Apr. 2022.

[6] B. Han, L.-N. Qiao, J.-L. Chen, X.-D. Zhang, Y. Zhang, and Y. Zhao, “GeneticKNN: a weighted KNN approach supported by genetic algorithm for photometric redshift estimation of quasars,” Research in Astronomy and Astrophysics/Research in Astronomy and Astrophysics, vol. 21, no. 1, p. 017, Jan. 2021.

[7] S. Zhang, “Challenges in KNN classification,” IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4663–4675, Oct. 2022.

[8] S. Zhang and J. Li, “KNN Classification with One-step Computation,” IEEE Transactions on Knowledge and Data Engineering, p. 1, Jan. 2021.

[9] J. Hu, H. Peng, J. Wang, and W. Yu, “kNN-P: A kNN classifier optimized by P systems,” Theoretical Computer Science, vol. 817, pp. 55–65, May 2020.

[10] B. Wang and S. Zhang, “A new locally adaptive K-nearest centroid neighbor classification based on the average distance,” Connection Science, vol. 34, no. 1, pp. 2084–2107, Jul. 2022.

[11] N. Rastin, M. Z. Jahromi, and M. Taheri, “A generalized weighted distance k-Nearest Neighbor for multi-label problems,” Pattern Recognition, vol. 114, p. 107526, Jun. 2021.

[12] A.-J. Gallego, J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, “Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation,” Pattern Recognition, vol. 74, pp. 531–543, Feb. 2018.

[13] X. Zhang, H. Xiao, R. Gao, H. Zhang, and Y. Wang, “K-nearest neighbors rule combining prototype selection and local feature weighting for classification,” Knowledge-based Systems, vol. 243, p. 108451, May 2022.

[14] N. Gul, W. K. Mashwani, M. Aamir, S. Aldahmani, and Z. Khan, “Optimal model selection for k-nearest neighbours ensemble via sub-bagging and sub-sampling with feature weighting,” Alexandria Engineering Journal /Alexandria Engineering Journal, vol. 72, pp. 157–168, Jun. 2023.

[15] Y. Chen and Y. Hao, “A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction,” Expert Systems With Applications, vol. 80, pp. 340–355, Sep. 2017.

[16] L. Chen and G. Guo, “Nearest neighbor classification of categorical data by attributes weighting,” Expert Systems With Applications, vol. 42, no. 6, pp. 3142–3149, Apr. 2015.

[17] C. Gong, Z.-G. Su, X. Zhang, and Y. You, “Adaptive evidential K-NN classification: Integrating neighborhood search and feature weighting,” Information Sciences, vol. 648, p. 119620, Nov. 2023.

[18] Z. Bian, C. M. Vong, P. K. Wong, and S. Wang, “Fuzzy KNN method with adaptive nearest neighbors,” IEEE Transactions on Cybernetics, vol. 52, no. 6, pp. 5380–5393, Jun. 2022.

[19] J. M. Keller, M. R. Gray, and J. A. Givens, “A fuzzy K-nearest neighbor algorithm,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 4, pp. 580–585, Jul. 1985.

[20] J. Gou et al., “A representation coefficient-based k-nearest centroid neighbor classifier,” Expert Systems With Applications, vol. 194, p. 116529, May 2022.

[21] M. Açıkkar and S. Tokgöz, “An improved KNN classifier based on a novel weighted voting function and adaptive k-value selection,” Neural Computing & Applications, vol. 36, no. 8, pp. 4027–4045, Dec. 2023.

[22] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, “An evolutionary algorithm based on Minkowski Distance for Many-Objective optimization,” IEEE Transactions on Cybernetics, vol. 49, no. 11, pp. 3968–3979, Nov. 2019.

[23] S. Chowdhury, N. Helian, and R. C. De Amorim, “Feature weighting in DBSCAN using reverse nearest neighbours,” Pattern Recognition, vol. 137, p. 109314, May 2023.

[24] Á. Ipkovich, K. Héberger, and J. Abonyi, “Comprehensible visualization of multidimensional data: sum of Ranking Differences-Based parallel coordinates,” *Mathematics*, vol. 9, no. 24, p. 3203, Dec. 2021.

[25] D. P. Turner, H. Deng, and T. T. Houle, “Understanding and applying confidence intervals,” Headache the Journal of Head and Face Pain, vol. 60, no. 10, pp. 2118–2124, Nov. 2020.

[26] H. Lüpsen, “Generalizations of the Tests by Kruskal-Wallis, Friedman and van der Waerden for Split-plot Designs,” Austrian Journal of Statistics, vol. 52, no. 5, pp. 101–130, Sep. 2023.

[27] L. Štěpánek, F. Habarta, I. Mala, and L. Marek, “A short note on post-hoc testing using random forests algorithm: Principles, asymptotic time complexity analysis, and beyond,” Annals of Computer Science and Information Systems, Sep. 2022.

[28] T. J. VanderWeele and M. B. Mathur, “Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad?,” American Journal of Epidemiology, vol. 188, no. 3, pp. 617–618, Nov. 2018.

SIEM and Threat Intelligence: Protecting Applications with Wazuh and TheHive

Jumiaty, Benfano Soewito
Computer Science Department-BINUS Graduate Program-Master of Computer Science,
Bina Nusantara University, Jakarta 11480, Indonesia

Abstract—**The consequences of cyberattacks on enterprises are highly varied. DDoS assaults can render an organization's website inaccessible; SQL attacks can compromise the integrity of data in a database, and Brute Force attacks can lead to unauthorized users gaining control over a server or application. Hence, it is crucial for enterprises to be aware of these potential dangers and employ solutions capable of monitoring networks, apps, and servers. In this study, the author employs Wazuh, TheHive, Telegram, and CVSS. Wazuh functions as a tool for monitoring applications and identifying potential security risks. TheHive classifies threats according to their level of importance. Telegram is utilized for dispatching notifications to the administrator. The findings indicate that Wazuh can promptly identify security risks by verifying that the date and time configurations on each utilized server align with the Indonesian time standard. Several vulnerabilities in the applications were successfully detected. The Wazuh server monitors two specific apps, namely Kompetensi and ESPPD. Surveillance commenced on March 20, 2024, at 17:49 and concluded on June 20, 2024, at 01:10, effectively amassing a total of 16,580 logs. 11 essential alert categories require follow-up due to their potential to compromise the system's integrity, confidentiality, and availability. To validate the detection results, the Common Vulnerability Scoring System (CVSS) is used. The assessment of vulnerability levels varies depending on the Wazuh level and CVSS. This arises because CVSS assigns scores based on five exploitability characteristics and incorporates the expertise of specialists to determine the assessment category and evaluate the potential impact of a successful threat. The outcome of this assessment, involving professional expertise, is heavily influenced by the unique attributes of each company. As a result, even when evaluating the same threats, the assessment can yield varying results. Evaluations utilizing Wazuh and CVSS are highly efficient in determining the extent of discovered hazards. By integrating these two technologies, the produced findings become more accurate.**

Keywords—Application server security; application vulnerability; threat detection; SIEM; Wazuh; TheHive; Telegram and CVSS

I. Introduction

The exponential growth of information technology will inevitably lead to a corresponding rise in cyber-attacks. Cyberattacks are directed on individuals and government organizations and enterprises [1]. Distributed Denial of Service (DDoS), SQL Injection, and Brute Force are some of the several types of cyberattacks [2]. Cyberattacks have severe consequences, including financial losses, exposure of personal information, harm to reputation, disruption of operations, and the expenses required to manage the event [3]. To mitigate cyber attacks, enterprises are required to establish robust risk management protocols to safeguard applications from such threats. Some risk management frameworks that can be utilized are NIST Cybersecurity, ISO 27000, ISA/IEC 62443, GDPR, and CIS Controls [4].

Ensuring the security of applications is a crucial component of preserving the overall security of a system. Two tools commonly employed for application monitoring are Open Source Security (OSSEC) and Security Information and Event Management (SIEM). SIEM offers immediate log analysis and management, facilitating prompt identification of threats and expedited reaction to security issues. Utilizing these tools can enhance businesses' ability to identify and mitigate possible security threats with more efficiency [5][6].

SIEM includes two categories of tools: commercial tools and open-source tools. Typically, paid SIEM systems have advanced functionalities, but they come with a higher cost. Some examples of paid SIEM tools include Splunk and IBM Qradar [7]. Nevertheless, the functionalities of open source SIEM can be highly efficient, but need more configuration and manual upkeep. When effectively managed, these open source solutions can offer robust and adaptable security measures customized to the organization's requirements.

Some of the open source SIEM tools include Open-source SIEM (OSSIM), Elasticsearch-Logstash-Kibana (ELK) stack, and Wazuh [8]. Wazuh assists in safeguarding the digital assets of businesses and individuals against security threats. The primary elements of Wazuh consist of the Wazuh indexer, Wazuh server, Wazuh dashboard, and Wazuh agent. The Wazuh agent is installed in the target application that is to be monitored [9]. The Wazuh platform includes SIEM management, allowing for real-time monitoring and detection of incidents through the analysis of event or activity reports within the application [10]. This implementation can further enhance the usefulness of Indeks KAMI as a result of SIEM's capability to assess system vulnerabilities, facilitate monitoring and auditing in relevant work units. This implementation has the potential to increase the value of the Information Security Index (KAMI) by utilizing the capabilities of SIEM to evaluate system vulnerabilities and optimize monitoring and auditing procedures in relevant work units [11]. Furthermore, SIEM has the capability to be integrated with SOAR and Honeyport in order to safeguard crucial assets inside an organization [12].

This research aims to identify application security vulnerabilities within an organizational unit by utilizing

Wazuh, which is seamlessly linked with TheHive. Wazuh serves the purpose of monitoring security threats, whereas TheHive serves the purpose of responding to incidents. Based on the analysis results, it can be inferred that Wazuh has a total of 4,372 rules and 16 levels of vulnerability. These rules will inevitably generate a substantial volume of logs on a daily basis. Not all of these logs pertain to the identification of security threats in applications. Only pertinent rules will be employed, and any rules that are not utilized will be disabled.

The Wazuh server collects logs from agents that are installed on the target monitoring application. The amount of agents that can be assigned to a Wazuh server is flexible and can be adjusted based on the organization's requirements. Wazuh agents are compatible with multiple operating systems, including Windows, Linux, Mac, Solaris, AIX, and hpUX [13].

The Wazuh Dashboard will present logs according to the agent. Administrators have the ability to view comprehensive information on threats, including the detection level of each agent. This will provide challenges for administrators to monitor concurrently. So the optimal approach to facilitate administrators' application monitoring is to establish integration Wazuh and TheHive. TheHive will collect logs from Wazuh and provide them on a unified dashboard page for all agents.

Integrating Wazuh with TheHive is a challenging task because of the absence of a shared network between the Wazuh server and TheHive server. To integrate Wazuh and TheHive on the same network, the zerotier custom platform is required as an additional step. Nevertheless, there are many advantages to be gained from effectively combining two technologies to be concurrently utilized in resolving issues within organizational units. In their research, Muhammad Alfian Fahrudi dan I Made Suartana [14] performed a three-stage testing process to integrate Wazuh with Telegram. The stages included vulnerability evaluation, injection attacks, and brute force. The findings demonstrated that the integration of Wazuh with Telegram enables the identification of potential risks and their subsequent transmission to the administrator through the Telegram application. However, the author's predicted detection time is surpassed due to the substantial data kept on the server, resulting in a lengthy procedure lasting approximately 10 minutes. Another research conducted by Muhammad Dehan Pratama, Fitri Nova and Deddy Prayama [15] the integration of wazuh and suricata. Suricata is utilized for threat detection, whereas Wazuh is employed to showcase the logs produced by Suricata on the Wazuh dashboard. The detection process is specifically aimed at identifying denial-of-service (DoS) assaults inside the flood attack category. Out of the five attack attempts, only two were successfully identified by Suricata. This detection rate was influenced by the use of an AWS Amazon server. Some attacks were promptly rejected by the server, preventing Suricata from detecting them.

Therefore, this study aims to combine the Wazuh and TheHive methods to identify and address instances of application security threats. Real-time threat detection will be implemented by Wazuh. To guarantee the real-time detection of threats, it is necessary to configure the time zone on each server. The identified threats will thereafter be transmitted to TheHive according to the pre-established threat level. Based on the identified dangers, TheHive will generate a case, which will then be partitioned into multiple tasks. By examining the specifics of the identified risks, the duties will be subsequently assigned to multiple teams, including the network security team, vulnerability management team, incident response team, and database security team. Following the successful creation of the case, TheHive will promptly transmit a notification to the application administrator using Telegram. The output generated by Wazuh and TheHive Integration will go through validation using the Common Vulnerability Scoring System (CVSS) 4.0 in order to assess the genuine validity and significance of the identified threats by the system. By including Wazuh, TheHive, telegram, and validation using CVSS, it is anticipated that applications inside these organisational units will be adequately protected against cyber threats.

II. Literature Review

Research reviews are performed to validate the chosen technique and uncover areas of research that have not been explored, so opening up new possibilities for this study. Stefan Stanković, Slavko Gajin, and Ranko Petrović [16], did research on the application of Wazuh for identifying security threats. They specifically focused on using Wazuh to identify attacks on web servers. Web servers are highly susceptible to a wide range of threats. Wazuh will provide a comprehensive and real-time display of the detected attacks. Wazuh is utilized not only for detecting security threats, but also for monitoring integrity, policies, and auditing systems.

Rio Pradana Aji, Yudi Prayudi and Ahmad Luthfi [17] conducted an additional study on Wazuh. They utilized Wazuh to enhance the website monitoring system by employing quantitative forensic investigation techniques to identify brute-force attacks. Wazuh assists businesses in the implementation of Centralized Log Management. Based on the research review, the utilization of Wazuh is currently restricted to the detection of a single form of assault, specifically brute force.

A study undertaken by Manju, Shanmugasundaram Hariharan, M. Mahasree, Andraju Bhanu Prasad and H.Venkateswara Reddy [18] investigated the detection of DDOS assaults by the integration of four tools: wireshark, snort, Wazuh, and splunk. Wireshark is utilized to conduct preliminary surveillance through the analysis of network traffic. In addition, the integration of snort with Wazuh will effectively identify and detect potential security threats. The output is transmitted to Splunk and will be presented in a way that is readily comprehensible to the administrator. The duration of this procedure is around five hours, resulting in a higher level of efficiency compared to the prior duration of seven to eight hours, resulting in a time savings of approximately three hours.

Additional study was carried out by Novianda Shafira Suryawatie Yomo, Ahmas Zafrullah Mardiansyah, and I Wayan Agus Arimbawa [19] utilising Security Information and Event Management (SIEM) with the Wazuh system. An experiment was conducted to assess the security of the University of Mataram academic information system by evaluating the Sql Injection attack utilising the Sql Injection payload inputted into the Burp Suite application.

Anand Groenewegen and Joris Shuko Janssen [20] conducted an evaluation and validation of TheHive Project, an open-source security Incident Response Platform. The objective of this study is to assess the level of maturity of TheHive Project as a security Incident Response Platform. Due to its comprehensive documentation, ease of management, and effectiveness in managing security incidents, TheHive Project is regarded as a mature security Incident Response Platform.

A study undertaken by Bharadwaj Mantha, Yeojin Jung, and Borja Garcia de Soto [21] employed the Common Vulnerability Scoring System (CVSS) to evaluate and quantify cyber vulnerabilities inside the construction sector. The CVSS system assigns a numerical score to individual vulnerability characteristics, therefore enabling the quantification of the security risk level for project participants including owners, contractors, and labour. The susceptibility of numerous leading construction firms was methodically evaluated using CVSS version 3.1, employing criteria including base, temporal, and environmental factors.

Based on the findings of the literature research, there are several efficient frameworks for real-time detection of threats to servers or apps. One such tool is Wazuh, which possesses the capability to be seamlessly incorporated with numerous other tools. This information provides background for the author to undertake research on the integration of Wazuh with TheHive for the purpose of detecting security threats and implementing incident reactions in the case of such threats. This integration not only identifies a single kind of threat, but is anticipated to identify all categories of threats that provide a risk to applications and servers. This research also has the implementation of a Telegram Bot to transmit real-time notifications to administrators. The last phase of the research is the verification process utilising CVSS 4.0.

III. Methodology

A. Phases of Research

This project involves the installation and evaluation of integrating Wazuh with TheHive to detect security vulnerabilities on application servers. This procedure involves configuring Wazuh, TheHive, and integrating Wazuh with thehive, as shown in Fig. 1. Subsequently, it is necessary to incorporate rulesets based on the specific requirements of the research. Wazuh is designed to carry out real-time monitoring of logs. Logs that are identified will be shown on the Wazuh dashboard for the purpose of analysis. Logs categorized as vulnerability levels 5 to 15 will be given to TheHive for additional analysis. TheHive will generate a case for the identified threat and break down the case into multiple tasks to be collectively analyzed with the security team, enabling administrators to respond promptly and effectively to the issue.

B. Network Topology

Fig. 2 depicts the functioning of the system, wherein the administrator gains access to the competency application and e-sppd through a VPN connection, using a unique login and password. Both applications require the installation of a wazuh agent, which is responsible for monitoring the logs of application activities. The application is susceptible to cyber attacks. The installed Wazuh agent on the application will transmit logs to the Wazuh server. In addition, the Wazuh server collects logs from the agent and presents them on the dashboard for the administrator to watch. Logs that are being monitored and have high vulnerability ratings will be transmitted to TheHive Server for additional examination. In addition, TheHive will generate a case based on the log and then split it into several tasks. Once the case is created, the administrator will automatically be notified via telegram.

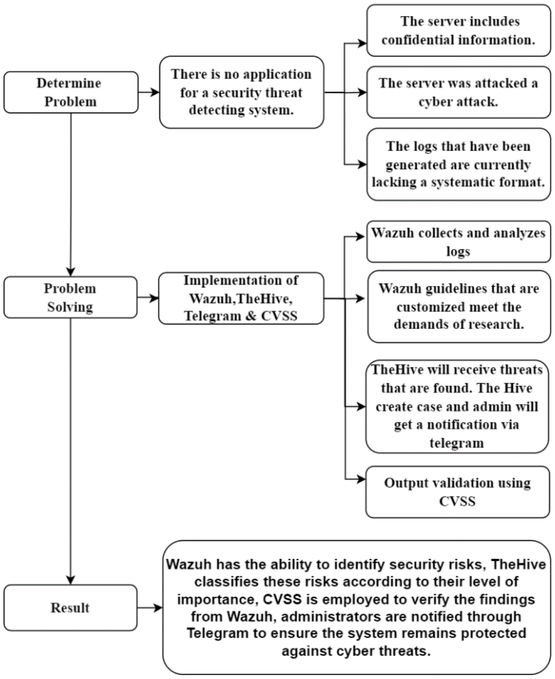

Determine Problem
There is no application for a security threat detecting system.
The server includes confidential information.
The server was attacked a cyber attack.
The logs that have been generated are currently lacking a systematic format.
Problem Solving
Implementation of Wazuh,TheHive, Telegram & CVSS
Wazuh collects and analyzes logs
Wazuh guidelines that are customized meet the demands of research.
TheHive will receive threats that are found. The Hive create case and admin will get a notification via telegram
Output validation using CVSS
Result
Wazuh has the ability to identify security risks, TheHive classifies these risks according to their level of importance, CVSS is employed to verify the findings from Wazuh, administrators are notified through Telegram to ensure the system remains protected against cyber threats.

Fig. 1. Phases of research.

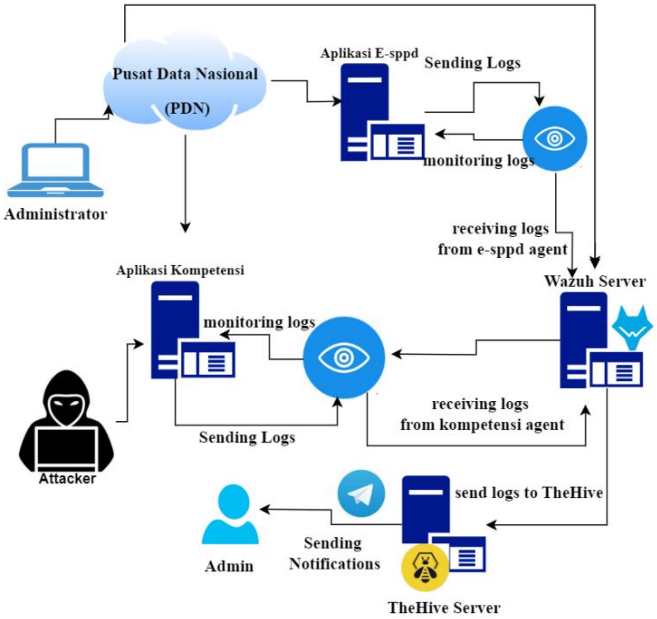


Fig. 2. Network topology.

C. Integration Flow of Wazuh and TheHive

Fig. 3 depicts the sequential steps involved in integrating Wazuh and TheHive. The initial step involves setting up the Wazuh server, which is then followed by configuring the thehive server. The subsequent step involves generating an integration file to facilitate the exchange of data between the two servers. Since Wazuh and TheHive are not connected to the same network, it is imperative to utilize zeroTier in order to establish a unified network for both. The integration file additionally establishes the threat level that will be transmitted to TheHive. TheHive will generate a case and partition it into multiple tasks to facilitate the analysis of the danger by administrators. The specifics of the Integration Flow are as follows: transmitted to TheHive. TheHive will generate a case and partition it into multiple tasks, hence facilitating the analysis of the threat for administrators.

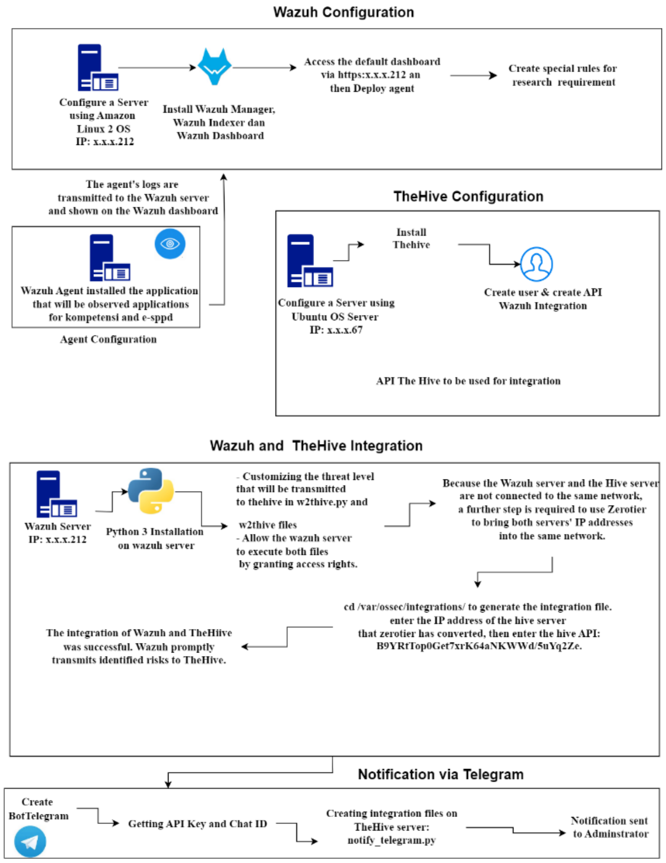

Wazuh Configuration
Configure a Server using Amazon Linux 2 OS IP: x.x.x.212
Install Wazuh Manager, Wazuh Indexer dan Wazuh Dashboard
Access the default dashboard via https:x.x.x.212 an then Deploy agent
Create special rules for research requirement
The agent's logs are transmitted to the Wazuh server and shown on the Wazuh dashboard
Wazuh Agent installed the application that will be observed applications for kompetensi and e-sppd
Agent Configuration
TheHive Configuration
Configure a Server using Ubuntu OS Server IP: x.x.x.67
Install Thehive
Create user & create API Wazuh Integration
API The Hive to be used for integration
Wazuh and TheHive Integration
Wazuh Server IP: x.x.x.212
Python 3 Installation on wazuh server
- Customizing the threat level that will be transmitted to thehive in w2thive.py and w2thive files
- Allow the wazuh server to execute both files by granting access rights.
Because the Wazuh server and the Hive server are not connected to the same network, a further step is required to use Zerotier to bring both servers' IP addresses into the same network.
cd /var/ossec/integrations/ to generate the integration file. enter the IP address of the hive server that zerotier has converted, then enter the hive API: B9YRtTop0Get7xrK64aNKWWd/5uYq2Ze.
The integration of Wazuh and TheHiive was successful. Wazuh promptly transmits identified risks to TheHive.
Notification via Telegram
Create BotTelegram
Getting API Key and Chat ID
Creating integration files on TheHive server: notify_telegram.py
Notification sent to Adminstrator

Fig. 3. System overview.

Based on Fig. 3, it can be concluded that the process details include:

- A server is configured with Amazon Linux OS 2 for the purpose of running Wazuh. This server consists of the Wazuh Manager, Wazuh Indexer, and Wazuh Dashboard. To access the Wazuh dashboard, use the designated IP address. Wazuh agents will be deployed on the application that needs to be monitored. The purpose of the agent is to transmit logs to the server, which will subsequently be exhibited on the Wazuh dashboard.
- The Ubuntu OS will host the installation of TheHive server. Once the installation is finished, the dashboard of TheHive can be accessed by using a designated IP address. First, add TheHive user to the system. Then, proceed with the creation of an API for the integration process.
- The integration between Wazuh and TheHive is accomplished using Python 3. The Wazuh server will have Python 3 installed.
- Wazuh and TheHive need to be connected within the same network utilizing zero tiers.
- The integration file will utilize the API provided by TheHive server and the IP generated by zero tier.
- Wazuh will send the logs to TheHive. The logs received by TheHives will be uploaded to the case, at the same time the administrator will receive detailed notifications of the threat.

IV. Result

A. Configuration and Modification of Wazuh Rules

System configuration and integration refer to Fig. 3. To implement Wazuh and TheHive, two servers are needed, namely the Wazuh server and TheHive server. The wazuh server will be installed on the Amazon Linux 2 operating system and TheHive server using ubuntu operating system. To start monitoring applications using agents, the Wazuh server needs to add as many agents as the number of applications to be monitored. The Wazuh server will monitor logs from two agents, the competency agent and the e-sppd agent. An illustration of the Wazuh server can be seen in Fig. 4.

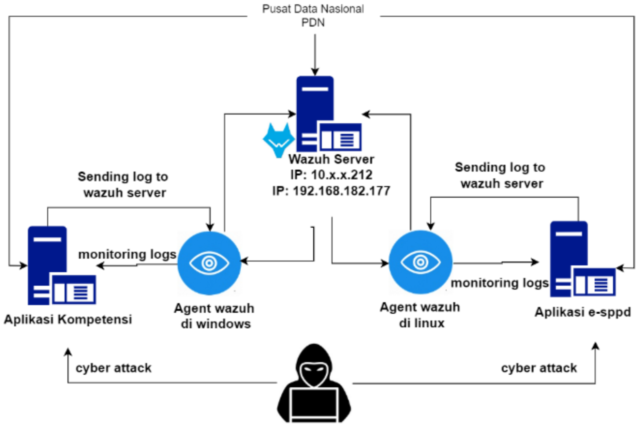


Fig. 4. Server Wazuh illustration.

Configuration stage of the wazuh server:

- Installation of Wazuh package on Amazon Linux server 2 : curl -sO https://packages.wazuh.com/4.4/wazuh-install.sh && sudo bash ./wazuh-install.sh -a
- Checking the Wazuh package that has been installed on the amazon linux server 2: sudo yum list-installed | grep Wazuh.

- After the installation process is complete, the Wazuh dashboard can be accessed using a dedicated ip via https://10.30.x.212.
- The next step is to add an agent. Add agent is done on the server side and the application side will be monitored. On the server side, add agent is done through the Wazuh dashboard then select deploy new agent and adjust the operating system used by the application that will be paired with the agent. The first agent is a competency application using the windows operating system, the agent installation stage from the windows-based application side can be done using windows powershell, the instructions are as follows:
  - Invoke-WebRequest -Uri https://packages.wazuh.com/4.x/windows/wazuh-agent-4.7.1-1.msi -OutFile ${env.tmp}\wazuh-agent; msiexec.exe /i ${env.tmp}\wazuh-agent /q WAZUH_MANAGER='10.30.x.212' WAZUH_AGENT_GROUP='default' WAZUH_AGENT_NAME='kompetensi' WAZUH_REGISTRATION_SERVER='10. 30.x.212'
  - NET START WazuhSvc

The second agent is the e-sppd application that uses the Linux operating system, the agent installation stage from the Linux-based application side can be done using the Linux terminal, the instructions are as follows:

- wget https://packages.wazuh.com/4.x/apt/pool/main/w/wazuh-agent/wazuh-agent_4.7.1-1_amd64.deb && sudo WAZUH_MANAGER='10.30.x.212' WAZUH_AGENT_NAME='esppd' dpkg -i ./wazuh-agent_4.7.1-1_amd64.deb
- sudo systemctl daemon-reload
- sudo systemctl enable wazuh-agent
- sudo systemctl start wazuh-agent

Threats will be identified by the installed Wazuh agent using Wazuh rules. Wazuh's 16 rule classification tiers are organized according to how seriously systems and applications are threatened. These levels are numbered 0 through 15, with each level denoting a distinct degree of intensity. Every rule is intended to identify a range of potentially harmful or suspicious activity on a network or system, from very minor risks to more significant and destructive assaults. Wazuh gives administrators flexibility in responding to various threats by offering 16 levels of rule classification. This allows administrators to tailor actions based on the severity and criticality of each security event that is identified. Furthermore, users may more easily group and arrange security responses thanks to a clear hierarchy in rule classification, which increases handling efficiency for both threats and security incidents as a whole.

Wazuh has 4. 372 ID rules in addition to rule classification, which are grouped according to their functions: syslog, firewall, ids (intrusion detection system), web-log, squid (proxy server), Windows, Wazuh, sysmon (system monitor), powershell, cloudflare (web application firewall), audit detections, Amazon Security Lake, ms-graph (Microsoft Graph), multiverse, sshd (Secure Shell Daemon), fireeye, and unbound (DNS Server). These groups make it simple to arrange and sort rules according to the kind of log or activity being monitored.

In accordance with the requirements of research, Wazuh also makes it easier to add additional decoders and rules. Using the Wazuh dashboard, accessible via the Management menu, one may create custom rules by first selecting rules and then searching for "local_rules.xml" and the specific custom rule that one wants.

B. *Installation Stage of TheHive Server:*

- wget -q -O /tmp/install.sh https://archives.strangebee.com/scripts/install.sh ; sudo -v ; bash /tmp/install.sh. then select 2 which are Install TheHive.
- After the installation process is complete, Thehive dashboard can be accessed using a dedicated ip via http://10.10.x.67:9000/login.
- The next step is to add user mimi@thehive.local and create API Key B9YRtTop0Get7xrK64aNKWWd/5uYq2Ze which will be used in the integration process. after the integration process is complete, the user mimi@thehive.local will receive logs from the Wazuh.

C. *Integration of Wazuh and TheHive Servers*

The Integration Stages are:

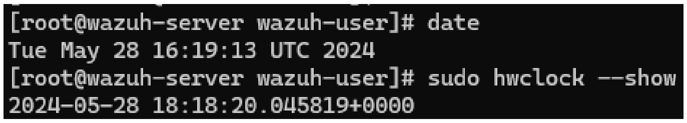
- Installing python 3 on Wazuh server
- Install the Hive Python using PIP (Python Package Index)
- create custom-w2thive.py. and custom-w2thive file which will be used for Wazuh and TheHive integration.
- Customize the custom-w2thive.py and determine the logs that will be sent by Wazuh to TheHive are logs with at least rules level 5 to level 15.
- Continued by customizing the custom-w2thive file.
- Provide access to Wazuh to run the custom-w2thive.py file. and w2thive.:
  - sudo chmod 755 /var/ossec/integrations/custom-w2thive.py
  - sudo chmod 755 /var/ossec/integrations/custom-w2thive

  - sudo chown root:wazuh-user/var/ossec/integrations/custom-w2thive.py
  - sudo chown root:wazuh-user /var/ossec/integrations/custom-w2thive

- The Wazuh Server and TheHive Server are not connected to each other within the same network. An effective resolution is to employ zerotier. The procedure consists of the following steps:
  - Create an network name : mythesis_network
  - Create network_ID : 363c67c55a3f3ff1.
  - Select the format of the desired IP, namely: 192.168.x.x.
  - Install zerotier on both server: curl -s htttps://install.zerotier.com.
  - Connect thehive server with zerotier network ID, with the: zerotier-cli join 363c67c55a3f3ff1.
- After the Wazuh server and Wazuh thehive are in the same network and get the IP from the zero tier. The next step is to create integration files. Integration files can be created through the Wazuh dashboard on the Management menu and then select configuration. IP yang di input pada file merupakan IP dari zerotier. The API used is an API from TheHive server.
- The integration process is completed, TheHive will receive logs from any server starting from rule level 5 to rule Level 15.

D. Threat Detection Results Based on Logs

Based on the application threat detection results from the Wazuh server, there is a discrepancy in the timing between the attack attempt and the log generated by the server. This occurs due to a discrepancy in the date settings between the Wazuh server and the Wazuh dashboard. Presented in Fig. 5 are the unsynchronised date settings.

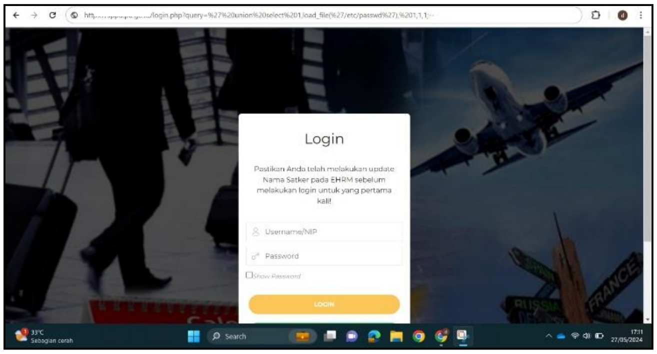


```
[root@wazuh-server wazuh-user]# date
Tue May 28 16:19:13 UTC 2024
[root@wazuh-server wazuh-user]# sudo hwclock --show
2024-05-28 18:18:20.045819+0000
```

Fig. 5. Unsynchronized time display.

The date instruction displays the date and time of the Wazuh server, while the sudo hwclock --show instruction displays the actual date and time. The following is a Sql Injection experiment by inputting the command : https://sppd.pu.go.id/login.php?query=%27%20union%20select%201,load_file(%27/etc/passwd%27),%201,1,1;-- on the web url that will be accessed.

The Sql Injection test in Fig. 6 was conducted on May 27, 2024, at 17:11 but Wazuh recorded the attack at 16:5516.55. Fig. 7 shows Sql injection detection.



http......../login.php?query=%27%20union%20select%201,load_file(%27/etc/passwd%27),%201,1,1;--

Login

Pastikan Anda telah melakukan update Nama Satker pada EHRM sebelum melakukan login untuk yang pertama kali!

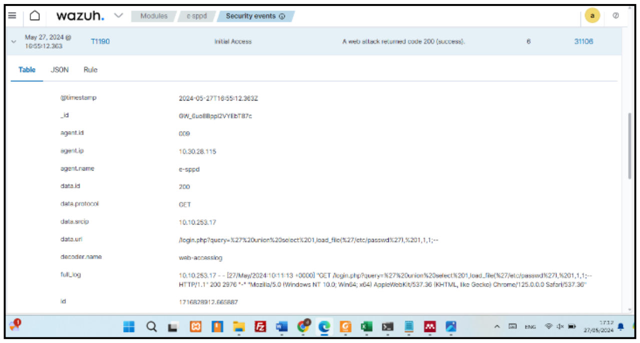
Username/NIP

Password

□ Show Password

LOGIN

Fig. 6. Sql injection testing.



May 27, 2024 @ 16:55:12.363	T1190	Initial Access	A web attack returned code 200 (success).	6	31106

@timestamp	2024-05-27T16:55:12.363Z
_id	GW_6uoBBppl2VYEbT87c
agent.id	009
agent.ip	10.30.28.115
agent.name	e-sppd
data.id	200
data.protocol	GET
data.srcip	10.10.253.17
data.url	/login.php?query=%27%20union%20select%201,load_file(%27/etc/passwd%27),%201,1,1;--
decoder.name	web-accesslog
full_log	10.10.253.17 - - [27/May/2024:10:11:13 +0000] "GET /login.php?query=%27%20union%20select%201,load_file(%27/etc/passwd%27),%201,1,1;-- HTTP/1.1" 200 2976 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/125.0.0.0 Safari/537.36"
id	1716828912.665887

Fig. 7. Sql injection detection.

The attack was carried out using IP 192.168.56.1 but the detection results were noted ip10.x.x.17. If viewed from the IP recorded by the Wazuh as if the attempted attack came from within the organization. This is because of the error in the organization's internal network configuration, this finding will be an input to the internal network improvement. The solution used for time synchronization on wazuh servers and wazuh dashboards is as follows:

- Install ntp: sudo yum install ntp -y
- Change ntp server: Sudo nano /etc/ntp.conf .
- Install ntpdate: sudo yum install ntpdate -y
- Synchronize system time with ntp server: sudo ntpdate -u pool.ntp.org
- If still not synchronized, verify the server time zone: timedatectl.
- Set the time zone to match the server location: sudo timedatectl set-timezone Asia/Jakarta
- Then re-synchroniz: sudo ntpdate -u pool.ntp.org

Furthermore, in addition to performing Sql Injection tests. The Wazuh server additionally monitors two specific apps, namely the competence application and the esppd application. The surveillance commenced on March 20, 2024, at 17:49 and concluded on June 20, 2024, at 01:10, effectively amassing a total of 16,580 logs. There are 11 essential alert categories that require follow-up due to their potential to disturb the system's integrity, confidentiality, and availability. Table I displays a summary of the 11 alert categories.

TABLE I. DETECTED THREATS

No	Alert	Level	Rule.mitre.technique	Amount
1	URL too long. Higher than allowed on most browsers. Possible attack.	13	Endpoint Denial of Service	2
2	Multiple web server 400 error codes from same source ip.	10	Vulnerability Scanning/Reconnaissance	5.388
3	High amount of POST requests in a small period of time (likely bot).	10	Network Denial of Service	857
4	Multiple web server 500 error code (Internal Error).	10	-	2
5	Multiple Windows logon failures.	10	Brute Force	11
6	SQL injection attempt.	7	Exploit Public-Facing Application	100
7	Listened ports status (netstat) changed (new port opened or closed).	7	Netstat Listening Ports	727
8	Host-based anomaly detection event (rootcheck).	7	-	171
9	Integrity checksum changed.	7	Stored Data Manipulation/ Impact	288
10	File deleted.	7	File Deletion/Data Destruction	38
11	Common web attack	6	Process Injection, File and Directory Discovery, Exploit Public-Facing Application	192

Referring to the information shown in Table I, Wazuh will identify a total of 11 risks. These threats will be thoroughly detailed, together with their respective impacts on servers and applications.

1) URL too long. Higher than allowed on most browsers. Possible attack, Identifies a potential attack in which an attacker can take advantage of a vulnerability by transmitting an excessively lengthy URL to the target. The significance of this danger on the system's confidentiality, integrity, and availability is substantial.

2) Multiple web server 400 error codes from same source ip executed across a network that does not necessitate specific circumstances or access privileges and does not need user engagement. Although server availability will be affected by this threat, data confidentiality and integrity will remain unaffected. Yet, these vulnerability screening operations can also serve as the initial stage of more severe attacks.

3) High amount of POST requests in a small period of time (likely bot). The process is executed over the network using POST requests that can be automated by bot programming languages. The present attack does not affect the principles of secrecy and integrity. Nevertheless, this attack will significantly affect availability by inundating the server with a substantial volume of requests, therefore introducing a denial of service (DoS) risk.

4) Multiple web server 500 error code (Internal Error).The exploitation of these vulnerabilities occurs via networks with modest complexity and does not necessitate privileges or user involvement. The aforementioned form of attack has minimal effect on the confidentiality and integrity of data, but significantly affects the availability of services.

5) Multiple Windows logon failures. These vulnerabilities are used on networks with modest complexity and do not necessitate any rights or user involvement. Upon successful execution of this Brute Force attack, the consequences for system confidentiality and integrity are significant, since the attacker gains the ability to view and alter sensitive data. Nevertheless, the effect on the availability of the system is really minimal.

6) The exploitation of Sql Injection attempts over the network can be achieved with minimal complexity and without the need for privileges or user involvement. Should the assault prove successful, the consequences for data confidentiality and integrity are much more pronounced, as the attacker gains access to and can alter sensitive data within the database. However, the effect on system availability is rather minimal, unless the attack results in a server breakdown or overload.

7) Listened ports status (netstat) changed (new port opened or closed). This vulnerability is leveraged across the network with minimal complexity and demands minimal rights without user involvement, rendering it insignificant to system availability, confidentiality, and integrity. Nevertheless, a modification in the state of a port can create new opportunities for attacks, hence enabling an attacker to gain unauthorized access or interfere with services operating on that specific port.

8) Host-based anomaly detection event (rootcheck). This vulnerability is leveraged via local access with minimal complexity and demands minimal rights without actual user involvement. NTFS Alternate Data Streams that are deemed suspicious have the potential to conceal harmful files or content that can be exploited by an attacker. While the effect on data integrity will be significant, the effect on system confidentiality and availability will be quite lesser.

9) Integrity checksum changed. The exploitation of these vulnerabilities occurs via local access with minimal complexity and necessitates elevated privileges without any user involvement. Alteration of system files, such as /usr/bin/cloud-init-per, can significantly affect the integrity and availability of the system, although the effect on system secrecy is rather little.

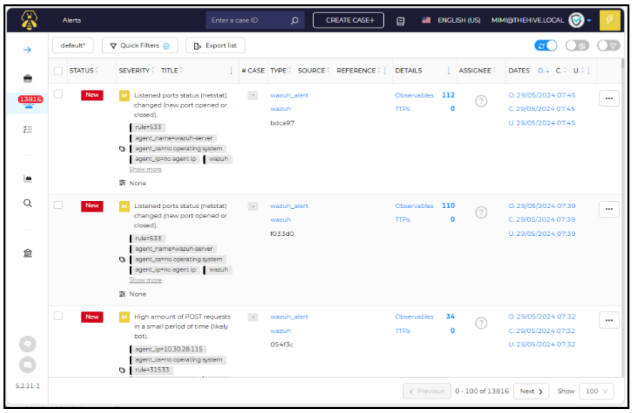
10) File deleted. This vulnerability through local access, this vulnerability demands high privileges without user interaction and is characterised by little complexity. Deleting certificate key files, such as /etc/ssl/pu_go_id/cert3.key, can significantly affect system integrity and availability by compromising the validity of certificates and causing disruptions to services that rely on such certificates. The effect on the confidentiality of the system is really minimal.

*11)*Common web attack. This vulnerability is exploited over the network with low complexity and without requiring privileges or user interaction. This attack attempts to access sensitive files such as /etc/passwd, which can have a high impact on system confidentiality and integrity if successful. However, the impact on system availability is relatively low.

The findings of this work demonstrate that the integration of wazuh with Thehive effectively detects a range of security concerns. Whereas research undertaken by Muhammad Dehan Pratama, Fitri Nova and Deddy Prayama [15] Wazuh and Suricata focuses on detecting a certain kind of threat, namely The detection procedure is deliberately designed to identify denial-of-service (DoS) attacks that fall within the flood attack category. Meanwhile, the research undertaken by Rio Pradana Aji, Yudi Prayudi and Ahmad Luthfi [17] Wazuh is limited to the identification of a singular type of attack, namely brute force.

E. *Threat Analysis*

Wazuh effectively transmits the identified alert records to TheHive. In addition, the logs will undergo further analysis. Fig. 8 displays the Dashboard view of TheHive.



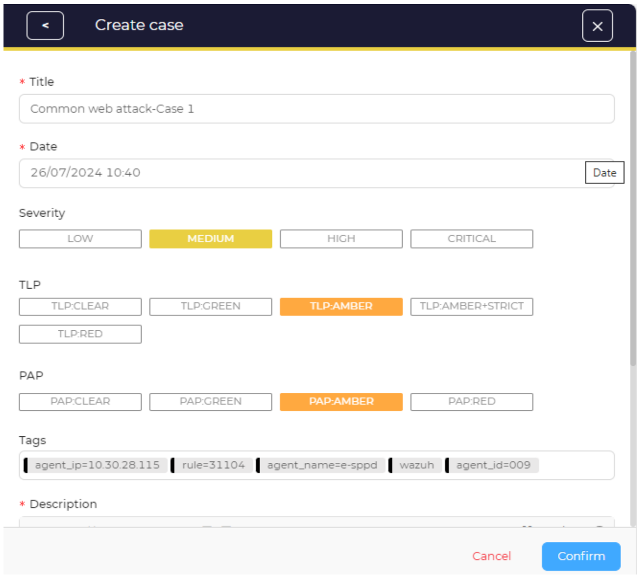
New | M | Listened ports status (netstat) changed (new port opened or closed). | - | wazuh_alert | wazuh | bdca97 | Observables 112 | TTPs 0 | O. 29/05/2024 07:45 | C. 29/05/2024 07:45 | U. 29/05/2024 07:45
rule=533 | agent_name=wazuh-server | agent_os=no operating system | agent_ip=no agent ip | wazuh
Show more
None

New | M | Listened ports status (netstat) changed (new port opened or closed). | - | wazuh_alert | wazuh | f033d0 | Observables 110 | TTPs 0 | O. 29/05/2024 07:39 | C. 29/05/2024 07:39 | U. 29/05/2024 07:39
rule=533 | agent_name=wazuh-server | agent_os=no operating system | agent_ip=no agent ip | wazuh
Show more
None

New | M | High amount of POST requests in a small period of time (likely bot). | - | wazuh_alert | wazuh | 054f3c | Observables 34 | TTPs 0 | O. 29/05/2024 07:32 | C. 29/05/2024 07:32 | U. 29/05/2024 07:32
agent_ip=10.30.28.115 | agent_os=no operating system | rule=31533

< Previous 0 - 100 of 13816 Next > Show 100

Fig. 8. Dashboard view of TheHive.



Create case

* Title: Common web attack-Case 1

* Date: 26/07/2024 10:40

Severity: LOW | MEDIUM | HIGH | CRITICAL (MEDIUM selected)

TLP: TLP:CLEAR | TLP:GREEN | TLP:AMBER | TLP:AMBER+STRICT | TLP:RED (TLP:AMBER selected)

PAP: PAP:CLEAR | PAP:GREEN | PAP:AMBER | PAP:RED (PAP:AMBER selected)

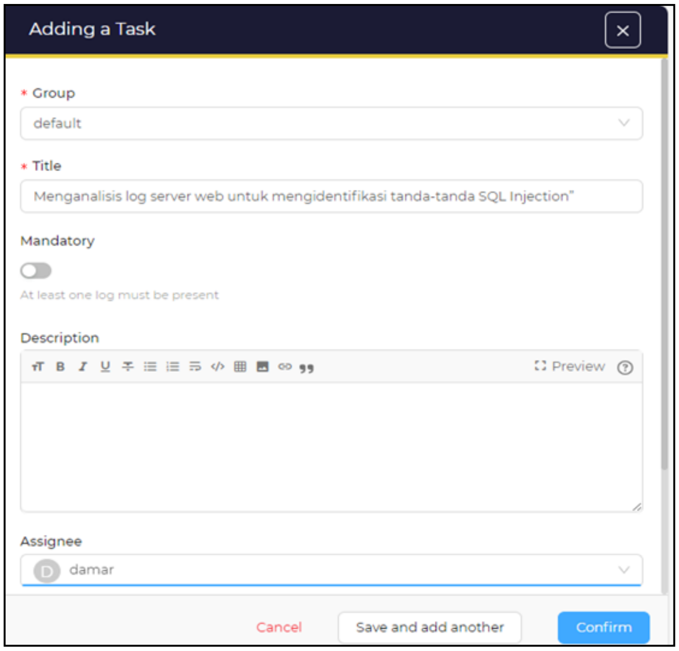
Tags: agent_ip=10.30.28.115 | rule=31104 | agent_name=e-sppd | wazuh | agent_id=009

* Description

Fig. 9. The process of creating a case.

TheHive will generate cases based on the collected logs and allocate them into tasks for each team. By dividing the responsibilities in this way, the analysis process can be expedited and enhanced, allowing for a prompter response to identified security threats. Fig. 9 displays the case view, where as Fig. 10 shows the tasks.

The generated case, including the additional tasks, will be assigned to many teams, such as the network security team, vulnerability management team, incident response team, and database security team, to jointly coordinate and carry out a comprehensive investigation of the risk. The generated case is depicted in Fig. 11. A successful case was set up on July 26, 2024 at 10:55 (WIB).



Adding a Task

* Group: default

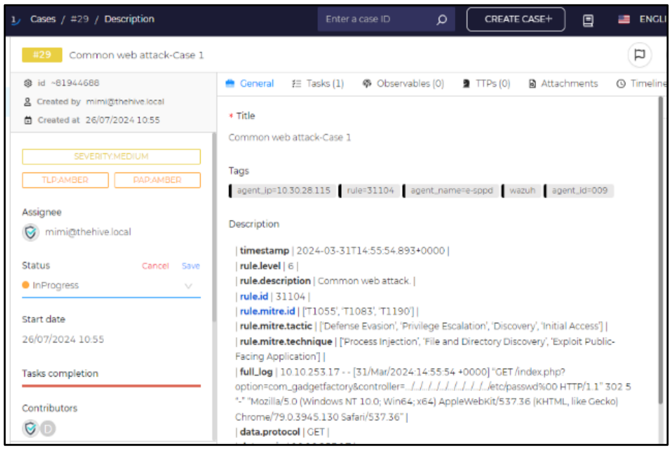
* Title: Menganalisis log server web untuk mengidentifikasi tanda-tanda SQL Injection"

Mandatory: At least one log must be present

Description

Assignee: damar

Fig. 10. The process of adding tasks for cases.



Cases / #29 / Description

#29 Common web attack-Case 1

id ~81944688
Created by mimi@thehive.local
Created at 26/07/2024 10:55

SEVERITY:MEDIUM
TLP:AMBER PAP:AMBER

Assignee: mimi@thehive.local

Status: InProgress

Start date: 26/07/2024 10:55

Tasks completion

Contributors

* Title: Common web attack-Case 1

Tags: agent_ip=10.30.28.115 | rule=31104 | agent_name=e-sppd | wazuh | agent_id=009

Description

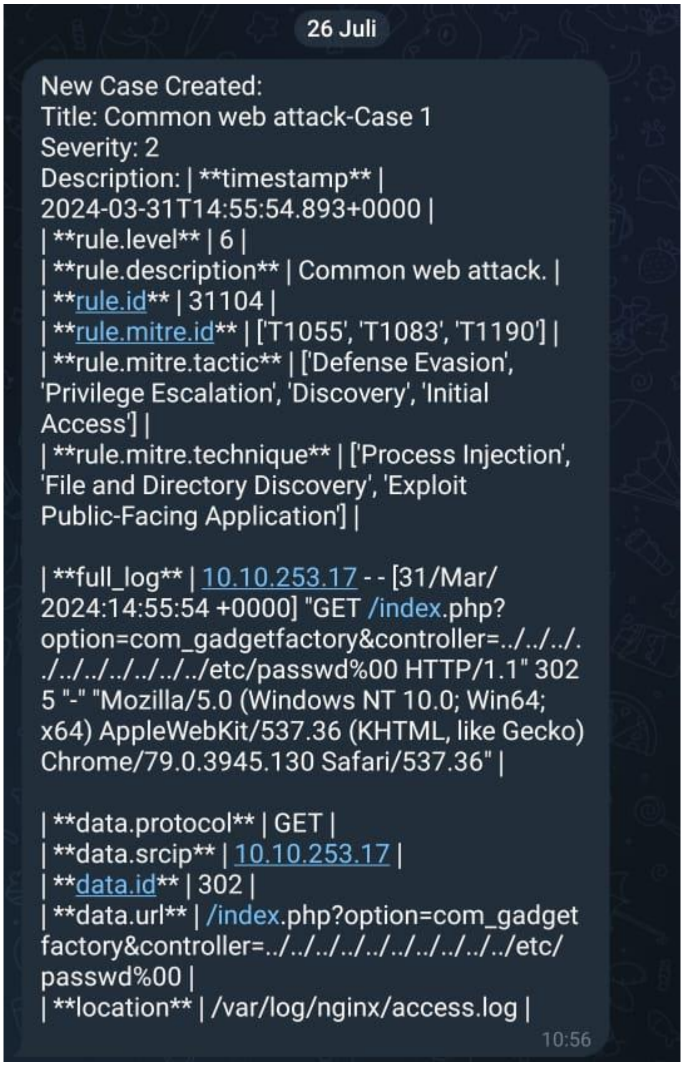
```
| timestamp | 2024-03-31T14:55:54.893+0000 |
| rule.level | 6 |
| rule.description | Common web attack. |
| rule.id | 31104 |
| rule.mitre.id | ['T1055', 'T1083', 'T1190'] |
| rule.mitre.tactic | ['Defense Evasion', 'Privilege Escalation', 'Discovery', 'Initial Access'] |
| rule.mitre.technique | ['Process Injection', 'File and Directory Discovery', 'Exploit Public-Facing Application'] |
| full_log | 10.10.253.17 - - [31/Mar/2024:14:55:54 +0000] "GET /index.php?option=com_gadgetfactory&controller=../../../../../../../../../etc/passwd%00 HTTP/1.1" 302 5 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3945.130 Safari/537.36" |
| data.protocol | GET |
```

Fig. 11. Cases that have been created.

Once the case and task have been established, the subsequent action involves notifying the administrator through the use of a telegraph. The following is a comprehensive breakdown of the steps:

- Develop a Telegram bot utilizing BotFather.
- Enter the command /start, followed by /newbot, and then provide a name for your bot, for as ThesisMimi_bot. Obtain the API Key: 7346165341:AAFqNsOpcqyg8vJ8ScMzYMQrWE8L3 3hpHb4.
- Enter the word "hello" in the chat with ThesisMimi_bot to obtain the Chat_ID. The result is as follows: {"ok":true,"result":[{"update_id":294625573,"message":{"message_id":2,"from":{"id":1134260586,"is_bot":false,"first_name":"Mimi","language_code":"id"},"chat":{"id":1134260586,"first_name":"Mimi","type":"private"},"date":1718511323,"text":"hello"}}]}
- The subsequent action involves generating an integration file on Thehive server, specifically named as the notify_telegram.py file.

An illustration of the notification received by the application administrator or server is shown in Fig. 12.



26 Juli

```
New Case Created:
Title: Common web attack-Case 1
Severity: 2
Description: | **timestamp** |
2024-03-31T14:55:54.893+0000 |
| **rule.level** | 6 |
| **rule.description** | Common web attack. |
| **rule.id** | 31104 |
| **rule.mitre.id** | ['T1055', 'T1083', 'T1190'] |
| **rule.mitre.tactic** | ['Defense Evasion',
'Privilege Escalation', 'Discovery', 'Initial
Access'] |
| **rule.mitre.technique** | ['Process Injection',
'File and Directory Discovery', 'Exploit
Public-Facing Application'] |

| **full_log** | 10.10.253.17 - - [31/Mar/
2024:14:55:54 +0000] "GET /index.php?
option=com_gadgetfactory&controller=../../../.
./../../../../../../etc/passwd%00 HTTP/1.1" 302
5 "-" "Mozilla/5.0 (Windows NT 10.0; Win64;
x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/79.0.3945.130 Safari/537.36" |

| **data.protocol** | GET |
| **data.srcip** | 10.10.253.17 |
| **data.id** | 302 |
| **data.url** | /index.php?option=com_gadget
factory&controller=../../../../../../../../../etc/
passwd%00 |
| **location** | /var/log/nginx/access.log |
```

10:56

Fig. 12. Notification via telegram.

The analysis of Fig. 11 cases that were created on July 26, 2024 at 10:55 (WIB) and Fig. 12 Notification via telegram on July 26, 2024 at 10:56 (WIB) reveals that the time interval between case creation and notification reception by the administrator is a mere one minute. These results demonstrate that the system exhibits a rapid and effective reaction in identifying and issuing alerts to the administrator about possible risks or incidents of security identified.

Whereas research undertaken by Muhammad Alfian Fahrudi dan I Made Suartana [14] The findings demonstrated that the integration of Wazuh with Telegram enables the identification of potential risks and their subsequent transmission to the administrator through the Telegram application. However, this resulted in a lengthy procedure lasting approximately 10 minutes. Meanwhile, the research undertaken by Manju, Shanmugasundaram Hariharan, M. Mahasree, Andraju Bhanu Prasad, and H. Venkateswara Reddy [18] examined the identification of Distributed Denial of Service (DDoS) assaults by combining four tools: wireshark, snort, Wazuh, and spearhead. The duration of this operation was approximately five hours.

F. *Validation of Results using the Common Vulnerability Scoring System (CVSS)*

Exploitability metrics and impact metrics are the primary factors utilized to evaluate the severity of vulnerabilities. Exploitability metrics assess the technical components of vulnerability exploitation, including Attack Vector (AV), Attack Complexity (AC), Attack Requirements (AT), Privileges Required (PR), and User Interaction (UI). The complete description of the Exploitability metrics feature may be found in Table III Impact metrics are employed to evaluate the consequences that would occur if a vulnerability is successfully exploited. These metrics encompass the impact on the confidentiality, integrity, and availability of the vulnerable system. The effect metrics element will be comprehensively explained in Table IV. Table II shows qualitative severity rating scale.

The vulnerability assessment results using CVSS 4.0 will be summarised in Table VI, which includes the Exploitability Metrics from Table III and the Impact Metrics from Table IV. This table additionally incorporates the outcomes of danger detection determined by the rules and degrees of the Wazuh. Prior to that, it is necessary to categorise the 15 levels of risk into five CVSS categories in order to facilitate the comprehension of each identified threat. Table V displays a classification of 15 Wazuh levels into five CVSS categories.

TABLE II. QUALITATIVE SEVERITY RATING SCALE

Category	Score CVSS
None	0.0
Low	0.1 - 3.9
Medium	4.0 - 6.9
High	7.0 - 8.9
Critical	9.0 - 10.0

TABLE III. EXPLOITABILITY METRICS

Metric Name	Metric Value	Description
Attack Vector (AV)	Network (N)	Vulnerabilities that can be exploited remotely over a network
	Adjacent (A)	The vulnerable system is inside the protocol stack, but the attack is limited to the protocol level on logically adjacent topologies.
	Local (L)	Attackers access vulnerable systems through local.
	Physical (P)	This attack requires the attacker to physically touch or manipulate the vulnerable system.
Attack Complexity (AC)	Low (L)	Attackers do not require specific targeting to take advantage of the vulnerability.
	High (H)	Attackers must have additional methods available to bypass existing security systems.
Attack Requirements (AT)	None (N)	Attack success is independent of the deployment and execution conditions of the vulnerable system or there are no attack requirements.
	Present (P)	The success of an attack depends on the conditions that require the preparation of specific targets that must be met in order to achieve vulnerability exploitation.
Privileges Required (PR)	None (N)	No privileges are required for an attacker to successfully exploit the vulnerability.
	Low (L)	Attackers need privileges, but can only access non-sensitive data
	High (H)	Attackers require privileges such as administrator who has full access to vulnerable systems.
User Interaction (UI)	None (N)	Vulnerable systems can be exploited without interaction from the user. remote attackers can send packets to the target system.
	Passive (P)	Successful exploitation of these vulnerabilities requires limited interaction by the targeted user with the vulnerable system. For example: utilizing a modified website to display malicious content when the page is rendered.
	Active (A)	Successful exploitation of these vulnerabilities requires the targeted user to perform specific interactions. For example: importing files into the vulnerable system in a specific way.

TABLE IV. IMPACT METRICS

Metric Name	Metric Value	Description
Confidentiality Impact to the Vulnerable System (VC)	High (H)	Confidential information/data is leaked. the information disclosed has immediate and serious consequences. the attacker has control of the information.
	Low (L)	The attacker has access to some information/data, but the attacker has no control over the information obtained. Leaked information has little impact
	None (N)	The confidentiality of the system is still maintained.
Integrity Impact to the Vulnerable System (VI)	High (H)	Complete loss of integrity. The attacker has full access to the data. Attackers can alter or delete data, leading to immediate and serious consequences on the system.
	Low (L)	Allows modification of data, but the attacker does not have full control over the data. The attacker does not have full control over the data, so it does not have an immediate and serious impact on the Vulnerable System.
	None (N)	System integrity is still maintained.
Availability Impact to the Vulnerable System (VA)	High (H)	Makes the service unavailable
	Low (L)	Performance degrades or interruptions in accessing the system occur. attackers do not have the ability to completely deny service to legitimate users
	None (N)	No impact on availability in Vulnerable Systems.

TABLE V. GROUPING OF 15 WAZUH LEVELS INTO FIVE CVSS CATEGORIES

Wazuh Level	CVSS Score	Description
Level 0	None (Score: 0.0)	It has nothing to do with security.
Level 2	Low (Score: 1.0 - 3.9)	Events that have only a minor impact on security but are not considered a significant threat.
Level 3		
Level 4		
Level 5		
Level 6	Medium (Score: 4.0 - 6.9)	Events that have a moderate impact on security could be a greater threat.
Level 7		
Level 8		
Level 9	High (Score: 7.0 - 8.9)	Events that have a major impact on security and can indicate an active attack or a serious problem requiring immediate action.
Level 10		
Level 11		
Level 12	Critical (Score: 9.0 - 10.0)	The events have had a huge impact on security. Indicates a serious attack to deal with.
Level 13		
Level 14		
Level 15		

A comparative graph of the two tools is created based on the data provided in Table VI. Fig. 12 depicts a comparison between the threat level provided by Wazuh and the CVSS score for different categories of security alerts. The graph displays two values for each sort of alert: the Wazuh level and the CVSS score.

TABLE VI. RECAPITULATION OF WAZUH LEVEL AND CVSS SCORE

No	Alert	Wazuh Level	CVSS Vector		CVSS Severity Level	
			Exploitability	Impact	Score	Descrip
1	URL too long. Higher than allowed on most browsers. Possible attack.	13	AV:N; AC:L; AT:N; PR:N; UI:A.	VC: H; VI: H; VA: H .	8.6	High
2	Multiple web server 400 error codes from same source ip.	10	AV:N; AC:L; AT:N; PR:N; UI:N.	VC:N; VI:N; VA:H.	8.8	High
3	High amount of POST requests in a small period of time (likely bot).	10	AV:N ; AC:L ; AT:N ; PR:N ; UI:N.	VC:N; VI:N; VA:H.	8.7	High
4	Multiple web server 500 error code (Internal Error).	10	AV:N; AC:L; AT:N; PR:N; UI:N.	VC:L; VI:L; VA:H .	8.8	High
5	Multiple Windows logon failures.	10	AV:N; AC:L; AT:N; PR:N; UI:N.	VC:H; VI:H; VA:L.	9.3	Critical
6	SQL injection attempt.	7	AV:N; AC:L; AT:N; PR:N; UI:N.	VC:H; VI:H; VA:L.	9.3	Critical
7	Listened ports status (netstat) changed (new port opened or closed).	7	AV:N; AC:L; AT:N; PR:L ; UI:N .	VC:L; VI:L; VA:L.	6.9	Medium
8	Host-based anomaly detection event (rootcheck).	7	AV:L; AC:L; AT:N; PR:L; UI:N.	VC:L; VI:H; VA:L.	6.9	Medium
9	Integrity checksum changed.	7	AV:L; AC:L; AT:N; PR:H; UI:N.	VC:L; VI:H; VA:H.	6.8	Medium
10	File deleted.	7	AV:L; AC:L; AT:N; PR:H; UI:L .	VC:L ; VI:H; VA:H.	6.8	Medium
11	Common web attack	6	AV:N; AC:L; AT:N; PR:N ; UI:N.	VC:H ; VI:H; VA:L .	9.3	Critical

Description:

- Attack Vector : AV
- Attack Complexity : AC
- Attack Requirements : AT
- Privileges Required : PR
- User Interaction : UI
- Confidentiality Impact : VC
- Integrity Impact : VI
- Availability Impact : VA

- Network : N
- for Attack Vector
- None : N
- for Attack Requirements, Privileges Required, User Interaction
- High : H
- Low : L

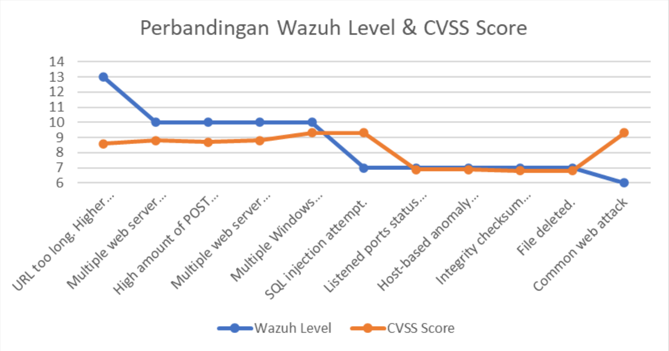


Fig. 13. Comparison chart of wazuh level and CVSS score.

Based on Fig. 13, it can be inferred that there are variations between the Wazuh level and CVSS Score, particularly in relation to the excessive length of the threat URL. Exceeds the maximum limit set by most browsers. There is a potential security breach involving multiple unsuccessful attempts to log in to Windows, an attempt to exploit SQL vulnerabilities, and a common type of attack targeting online applications. The disparity in vulnerability level assessment arises due to the utilisation of CVSS, which assigns a score based on five Exploitability characteristics. This process incorporates the expertise of professionals who determine the assessment category and take into account the potential impact in the event of a successful threat. The outcomes of this evaluation, which involves the expertise of professionals, vary based on the unique attributes of each organisation, although evaluating the same threats.

In general, evaluations utilising Wazuh and CVSS are highly efficient in identifying the severity of identified security risks. By integrating these two instruments, the acquired outcomes become more precise. The purpose of this assessment procedure is to systematically validate the hazards encountered, facilitating the implementation of appropriate mitigation measures. By integrating Wazuh with TheHive and utilising CVSS, organizations can enhance their ability to detect, assess, and address security issues, leading to improved application security.

V. CONCLUSION

Based on the conducted research, it can be inferred that:

1) Wazuh is utilised for the purpose of identifying and recognising potential threats, whilst TheHive is employed to scrutinise and assess the identified hazards. Each of these instruments is customised based on the specific study requirements. This integration also includes supplementary tools, including ZeroTier, which is used to establish a virtual network that links the two entities. Through this combination, Wazuh and TheHive effectively identify and respond to security issues in real-time, promptly notifying administrators via Telegram.

2) All servers should have their date and time settings synchronised to the Indonesian time zone. Synchronisation is crucial as it can impact the formatting of the detection result output in terms of date and time. By adjusting the time parameters appropriately, the resulting information will be more precise.

3) Deploying Wazuh and TheHive on distinct networks offers further advantages to the organisation. By having distinct server locations, the preservation of detection logs is secure, even in the event of one server experiencing a failure. Although integration may pose early challenges, it has the benefit of ensuring the security of log data in case of server issues.

4) Implement Secure Coding Practices, namely employing secure code to mitigate application vulnerabilities. The detection of threats in the form of multiple occurrences of web server 500 error code (Internal Error) twice, and web server 400 error code 2,444 times, suggests that system administrators need to exercise greater vigilance and caution while writing code for the system/application. This highlights a flaw in the code that has to be promptly addressed in order to avoid potential vulnerabilities that could be exploited by attackers.

5) Implement input validation to mitigate the risk of SQL injection, as well as enforce restrictions on the length of URLs received by the server to prevent potential buffer overflow or denial-of-service (DoS) attacks. This step is crucial in order to mitigate potential risks, such as encountering a URL that exceeds the maximum allowable length. Exceeds the maximum limit supported by most browsers. There is a potential attack and an attempt to do SQL injection.

6) Implementing Multi-Factor Authentication (MFA) to monitor user access and ensure authentication and authorisation. The discovered threats involved many instances of failed login attempts, which may suggest unauthorised access attempts or brute force attacks.

7) Assessments combining Wazuh and CVSS 4.0 are highly efficient in identifying the severity of identified threats. By integrating these two instruments, the outcomes achieved become more dependable and precise. The purpose of this assessment technique is to offer a concise and systematic evaluation of the encountered threats.

8) Possibly, TheHive might be integrated with network monitoring tools and intrusion detection systems like Suricata and Zeek (Bro) to enable the comparison of acquired findings for a more thorough study. The speed of technological advancement directly correlates with the increasing magnitude of cyber dangers. In order to enhance their preparedness, security teams should strive to incorporate tools that help streamline their tasks and provide robust security for servers and applications. Suricata and Zeek are mutually synergistic systems, with Suricata functioning as a signature-based detection and prevention system, Zeek providing behavior-based in-depth analysis, and TheHive serving as an incident management platform. The integration of the three components provides organisations with a more comprehensive threat detection system.

9) In the future, it is anticipated that all applications within the Organisational Unit would be able to incorporate wazuh and thehive for the purpose of monitoring servers and apps. This integration will facilitate administrators in monitoring the system, expediting preventive measures, and ensuring the maintenance of system security. This will enhance the efficacy of security management and offer superior safeguarding against prospective risks.

REFERENCES

[1] S. A. Utari, V. Ardia, Jamiati, and D. Fitria, “How an Organization Should Implement Risk Communication in Response to Cyber Attack in Indonesia,” J. Educ., vol. 05, no. 04, pp. 14314–14328, 2023, [Online]. Available: http://jonedu.org/index.php/joe

[2] N. Singh, “Sql i – a w,” vol. 2, no. 6, pp. 42–46, 2012.

[3] A. A. Putra, O. D. Nurhayati, and I. P. Windasari, “Perencanaan dan Implementasi Information Security Management System Menggunakan Framework ISO/IEC 20071,” J. Teknol. dan Sist. Komput., vol. 4, no. 1, p. 60, 2016, doi: 10.14710/jtsiskom.4.1.2016.60-66.

[4] V. Mahendra and B. Soewito, “Penerapan Kerangka Kerja NIST Cybersecurity dan CIS Controls sebagai Manajemen Risiko Keamanan Siber,” Techno.Com, vol. 22, no. 3, pp. 527–538, 2023, doi: 10.33633/tc.v22i3.8491.

[5] Ronal Hadi, Y. Yuliana, and H. A. Mooduto, “Deteksi Ancaman Keamanan Pada Server dan Jaringan Menggunakan OSSEC,” JITSI J. Ilm. Teknol. Sist. Inf., vol. 3, no. 1, pp. 8–15, 2022, doi: 10.30630/jitsi.3.1.58.

[6] M. Ramli et al., “Monitoring dan Evaluasi Keamanan Jaringan dengan Pendekatan Security Information and Security Management (SIEM),” vol. 16, no. 1, pp. 1979–276, 2023, doi: 10.30998/faktorexacta.v16i1.16534.

[7] S. S. Sekharan and K. Kandasamy, “Profiling SIEM tools and correlation engines for security analytics,” Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017, vol. 2018-Janua, pp. 717–721, 2017, doi: 10.1109/WiSPNET.2017.8299855.

[8] M. Sheeraz et al., “Effective Security Monitoring Using Efficient SIEM Architecture,” Human-centric Comput. Inf. Sci., vol. 13, p. 17, 2023.

[9] Wazuh, “Installing the Wazuh central components,” Wazuh. https://documentation.wazuh.com/current/installation-guide/index.html. (accessed May 23, 2024).

[10] N. F. Pratama, “Perancangan Sistem Deteksi Dini Keamanan Informasi DISKOMINFO Kabupaten Bandung,” JATISI (Jurnal Tek. Inform. dan Sist. …, vol. 10, no. 1, pp. 808–820, 2023.

[11] C. Arfanudin, B. Sugiantoro, and Y. Prayudi, “Analisis Serangan Router Dengan Security Information and Event Management Dan Implikasinya Pada Indeks Keamanan Informasi Analysis of Router Attack With Security Information and Event Management and Implications in Information Security Index,” CyberSecurity dan Forensik Digit., vol. 2, no. 1, pp. 2615–8442, 2019.

[12] M. Hafiz and B. Soewito, “Information Security Systems Design Using SIEM, SOAR and Honeypot,” J. Pendidik. Tambusai, vol. 6, no. 2, pp. 15913–15926, 2022.

[13] Wazuh, “Agen Wazuh.” https://documentation.wazuh.com/current/installation-guide/wazuh-agent/index.html#wazuh-agent (accessed May 23, 2024).

[14] M. A. Fahrudi and I. M. Suartana, “Integrasi End-point Security Berbasis Agent dan Bot Messenger untuk Deteksi dan Monitoring Serangan pada

Web Server secara Real-time," J. Informatics Comput. Sci., vol. 04, pp. 275–282, 2023, doi: 10.26740/jinacs.v4n03.p275-282.

[15] Fitri Nova, M. D. Pratama, and D. Prayama, "Wazuh sebagai Log Event Management dan Deteksi Celah Keamanan pada Server dari Serangan Dos," JITSI J. Ilm. Teknol. Sist. Inf., vol. 3, no. 1, pp. 1–7, 2022, doi: 10.30630/jitsi.3.1.59.

[16] Stefan Stanković, Slavko Gajin, and Ranko Petrović, "A Review of Wazuh Tool Capabilities for Detecting Attacks Based on Log Analysis," IX Int. Conf. IcETRAN, no. june, pp. 6–9, 2022.

[17] R. Pradana Aji, Y. Prayudi, and A. Luthfi, "Analysis of Brute Force Attack Logs Toward Nginx Web Server on Dashboard Improved Log Logging System Using Forensic Investigation Method," J. Tek. Inform., vol. 4, no. 1, pp. 39–48, 2023, doi: 10.52436/1.jutif.2023.4.1.644.

[18] H. V. Reddy, "Intrusion Detection System Using Customized Rules for Snort," Int. J. Manag. Inf. Technol., vol. 15, no. 3, pp. 01–14, 2023, doi: 10.5121/ijmit.2023.15301.

[19] N. Shafira Suryawatie Yomo, A. Zafrullah Mardiansyah, and I. Wayan Agus Arimbawa, "Deteksi Serangan SQL Injection Menggunakan Security Information and Event Management (SIEM) Wazuh," pp. 1–9, 2019, [Online]. Available: http://eprints.unram.ac.id/41453/

[20] A. Groenewegen and J. S. Janssen, "TheHive Project: The maturity of an open-source Security Incident Response platform," no. July, 2021.

[21] M. Bharadwaj R.K., J. Yeojin, and G. D. S. Borja, "Implementation of the Common Vulnerability Scoring System to Assess the Cyber Vulnerability in Construction Projects," no. June, pp. 117–124, 2020, doi: 10.3311/ccc2020-030.

Dynamic Priority-Based Round Robin: An Advanced Load Balancing Technique for Cloud Computing

Parupally Venu, Pachipala Yellamma*, Yama Rupesh, Yerrapothu Teja Naga Eswar, Maruboina Mahiddar Reddy
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur 522302, India

***Abstract*—An imbalance of load is an essential problem in cloud computing where the division of work between virtual machines is not well-optimized. Performance bottlenecks result from this unequal resource allocation, which keeps the system from operating at its full capability. Managing this load-balancing issue becomes critical to improving overall efficiency, resource utilization, and responsiveness as cloud infrastructures strive to respond to changing workloads and scale dynamically. Crossing the load-balancing landscape introduced a new strategy to effectively improve the load-balancing factor and ways to improve load-balancing performance by understanding how existing algorithms work, an effective method of load balancing. The "Dynamic Priority Based Round Robin" algorithm is a new approach that combines three different algorithms to improve cloud load balancing. This method improves load balancing by taking the best aspects of previous algorithms and improving them. It works remarkably well and responds quickly to commands, greatly reducing processing time. This DPBRR algorithm also plays an important role in improving cloud load balancing in many ways, including improving resource consumption, inefficiency, scalability, fault tolerance, cost optimization, and other aspects. Since it is a combination of algorithms, it may have its drawbacks, but its cloud computing enhancements are very useful for doing many tasks quickly. Strength and adaptability are quite effective, as is adaptability.**

***Keywords*—Load balancer; traffic distribution; cloud computing; resource utilization; scalability; Dynamic Priority Based Round Robin (DPBRR)**

I. Introduction

A cutting-edge paradigm in information technology, cloud computing marks a radical shift in the direction of the democratization of data access and decentralization of computational capacity. This cutting-edge technical framework breaks through conventional barriers by providing individuals and businesses with access to an infinite resource, from powerful computer power to endless storage. Cloud computing is a fascinating and amazing computing era that has replaced traditional methods. Cloud computing eliminates this necessity and offers services based on user demand and usage, in contrast to traditional computing, which requires users to maintain internal infrastructure. Users are spared from having to buy and maintain processing, storage, and other hardware. Through the internet, data is accessed and stored [1]. These days, cloud computing is essential because it offers pay-as-you-go on-demand services. To provide high-quality services, suppliers are taking advantage of service models like SaaS, PaaS, IaaS. Over the past five years, the public cloud computing markets have grown significantly, expanding by 21.5% [4].

Workload control is essential for load balancing problems since job arrival patterns are unpredictable and cloud node capacity varies. It also helps to preserve system stability. Schemes for load balancing can be either static or dynamic, depending on how essential System dynamics are [11]. Load balancing is a technique used in distributed systems to distribute the workload among multiple resources. As a result, there is an increase in scalability and effective resource usage. Multiple networked computing devices that collaborate are part of distributed systems. The need for efficient resource allocation grows as workload demand rises. This problem is solved by load balancing, which divides up incoming requests or tasks among the servers, virtual machines, or containers that are available. The task load is typically divided among several virtual machines (VM's) when using the load balancing technique. Without it, there could be SLA violations, assert wastage, execution corruption, and uneven burdens during server construction. Thus, employing an appropriate load balancing technique can enhance server utilization and provide more Quality of Service (QoS) assurance [12]. In a cloud environment, load balancing is accomplished in two steps: first the job is divided among the nodes; second, the virtual machine is monitored, and load balancing activities are carried out using task migration or virtual machine migration approach [5]. Load balancing in the cloud occurs in two stages: first, at the level of physical machines, where the load balancer distributes the load among the VMs connected to each physical device and manages the load of physical machines; second, at the level of virtual machines, where the load balancer manages and balances the load across all virtual machines using different LB algorithms [9].

The paper's content is organized as follows: Section II offers a comprehensive review of the literature on several key Load Balancing methods. In Section III, we review the current state of the art for the for DPBRR algorithm to enhance load balancing and introduce our proposed approach, to distribute tasks evenly among resources based on priority, the jobs with higher priorities will become more significant. We discuss the encryption process in detail. Section IV presents the results and a comparative analysis of our recommended methodology. A thorough explanation of the study paper's findings and closing thoughts is provided in Section V.

II. Literature Review

The research paper titled “A Hybrid Algorithm for Scheduling Scientific Workflows in Cloud Computing” by Muhammad Tahir, Muhammad sardaraz in the year 2019 [1] [10]. The authors put forth an algorithm that uses the PSO

*Corresponding Author.

algorithm to schedule scientific operations in two key stages: task preparation and task scheduling [23]. Throughout the scheduling process, the load balance of cloud resources is tracked by this hybrid method.

The research paper titled "Load balancing in cloud computing – A hierarchical taxonomical classification" by G Kavitha, S Afzal in the year 2019 [2]. It primarily categorized the many load balancing algorithm types with clarity and supplied details about the technique employed, the complexity of the algorithm, and its benefits and drawbacks. It primarily demonstrates the functionality and complexity of the load balancing algorithms, with task scheduling, virtual machine scheduling, and resource scheduling serving as the primary criteria. It also compares the algorithms' operational processes and provides percentages to help readers understand the algorithms, covering all the major LB algorithms.

The research paper titled "A Hybrid Bio-Inspired Algorithm for Scheduling and Resource Management in Cloud Environment" by Rajkumar Buyya, Ram Mohana Reddy G, Shridhar G D in the year 2020[3]. By combining modified CSO and PSO algorithms for job scheduling, the authors presented a hybrid with a bio-effect method for asset allocation. Compared to current scheduling and techniques, this algorithm aims to increase response time, reliability and resource utilization.

The research paper titled "A Hybrid Model for Load Balancing in Cloud Using File Type Formatting" by Adnan Sohail, M Junaid, Ahmed, H Alhakami, Imran Ali Khan, Abdullah Baz in the year 2020 [4]. The main goal is to use support vector machine (SVM) technology, which looks at the quantity and quality of files stored in the cloud. It works based on how well it manages processing speed and SLA.

The research paper titled "Dynamic load balancing algorithm for balancing the workload among virtual machines in cloud computing" by S C Sharma, Mohit Kumar in the year 2017 [5]. Cloud computing technologies are used to reduce waiting time and increase resource usage. Demonstrate that the suggested approach outperforms current algorithms regarding work allocation, workload monitoring and dynamic control. Incoming tasks are also assigned to the most appropriate virtual machine.

The research paper titled "A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications" by Azween Abdullah, N Z Jhanjhi, M A Alzain, Dalia Abdulkareem Shafiq in the year 2021[6]. Increase cloud asset efficiency rate 78% relative to current traffic distribution algorithms to achieve decent performance and improve metrics. The proposed method prioritizes virtual machines (VMs) for workload scheduling, underutilization, and quality of service (QoS) balancing. The algorithm mainly focuses on optimizing IaaS cloud models to ensure high performance and minimize underload.

The research paper titled "Cloud Dynamic Load Balancing and Reactive Fault Tolerance Techniques: A Systematic Literature Review (SLR)" by A Yousif, M Bakri Bashir, T M Tawfeeg, Alzubair Hassan, Samar M alqhtani, Awad Ali, Rafik Hamza in the year 2022 [7]. To make cloud load balancing more efficient, the authors of this study focus on implementing fault management techniques along with real time workload management. Adaptive load distribution jobs are dynamically assigned to the virtual machine based on current state of the system. (VMs). In addition, a reactive failover method reacts to system failures to keep cloud services stable.

The research paper titled "Dynamic Resource Allocation Using an Adaptive Multi-Objective Teaching-Learning Based Optimization Algorithm in Cloud" by Mohammadreza Ramezanpour, R Khorsand, Ali Moazeni in the year 2023 [8]. The authors recommend a dynamic resource allocation based on the strategy of AMO-TLBO method. By dynamically allocating resources based on application capacity, this strategy saves costs and optimizes resource consumption. The algorithm has been greatly improved using web-based technology. In addition, the recommended strategy outperforms several popular algorithms such as NSGA-II, MOPSO and TLBO.

The research paper titled "Load Balancing in Cloud Environment: A State-of-the-Art Review" by Durgaprasad G, Yogesh Lohumi, M Zubair Khan, Prakash Srivastava, Abdulrahman in the year 2023 [9]. Efficiency, scalability and performance are negatively affected by the load imbalance that cloud computing constantly faces. This essay examines load balancing and improving service quality in an on-demand computing setting. In addition to explaining load balancing solutions, the taxonomy and classification of load balancing algorithms is discussed. A common well-solved problem was server load imbalance.

Finally, a better answer to cloud computing load balancing issues and resource allocation concerns is offered by load balancing algorithms that integrate dynamic priority-based round robin (DPBRR). Task distribution, resource allocation, and workload modification are increasingly important considerations when working with virtual machines. As a result of the round robin process, which distributes jobs among the available virtual machines in a cyclic fashion to give each VM an equal chance, the priorities are dynamically changed in response to workload patterns. Its fault tolerance and high adaptability allow it to function as a reliable balancing load technique in environments of cloud computing by dynamically reallocating and redistributing resources to VMs with lower priority when any VM becomes overwhelmed.

The table highlights the drawbacks of each methodology and limitations. The proposed load balancing algorithm combines the dynamic nature with a priority based round robin algorithm offers a better solution for scalability, resource allocation, task distribution, fault tolerance and more, this can improve the balancing load in cloud environments.

The Table I provides a literature review on existing methodologies for enhancing balancing of load in virtualized computing environments. It provides a comparison of different algorithms, including Improved WRR, PMHEFT, CMODLB, Dynamic load balancing algorithms, SARM, MOABCQ and more.

TABLE I. LITERATURE REVIEW ON EXISTING METHODOLOGIES

S:NO	AUTHORS	TITLE	APPLIED METHODOLGY	DRAWBACKS
1	A Pravin , N Manikandan	‘An Efficient Improved Weighted Round Robin Load Balancing Algorithm in Cloud Computing’ [13].	Improved WRR	For weight assignment, the updated WRR mostly relies on static server specs. It may not adapt to dynamic changes.
2	S C Jain , M Sohani	‘A Predictive Priority-Based Dynamic Resource Provisioning Scheme With Load Balancing in Heterogeneous Cloud Computing’ [12].	PMHEFT	When the system gets bigger or there are more jobs to complete, its scalability could become an issue.
3	N Panwar, Sarita Negi , M M Singh Rautham , V Kunwar Singh	‘CMODLB: an efficient load balancing approach in cloud computing environment’ [14].	CMODLB	There may be a large processing overhead when several machine learning and optimization techniques are used.
4	Xuqing Ke, Lingjun Zhong, Wei Hou, Limin Meng,	‘Dynamic Load Balancing Algorithm Based on Optimal Matching of Weighted Bipartite Graph’ [15].	KUHN-MUNKRES	It can be more communication-intensive to update the weighted bipartite graph appropriately.
5	Fan Hong, Jiang Zhang, Tom Peterka, H Guo, Xiaoru Yuan	‘Dynamic Load Balancing Based on Constrained K-D tree Decomposition for Parallel Particle Tracing’ [16].	DECOMPOSITION K-D TREE	The method might not be as flexible when it comes to dynamic shifts in the workload allocation or system attributes.
6	Sun-Yuan Hsich, Rajkumar Buyya, Chih-Heng Ke, Albert Y Zomaya, W-K Chung, Yun Li	‘Dynamic Parallel Flow Algorithms With Centralized Scheduling for Load Balancing in Cloud Data Center Networks’ [17].	CDPFSMP & CDPFS	As a result of the centralization, the network may experience a bottleneck or single point of failure.
7	J-P Yang	‘Elastic Load Balancing Using Self-Adaptive Replication Management’ [20].	SARM	It might rely on how well certain thresholds and parameters are adjusted.
8	W Kimpan, B kruekaew	‘Multi-Objective Task Scheduling Optimization for Load Balancing in Cloud Computing Environment Using Hybrid Artificial Bee Colony Algorithm With Reinforcement Learning’[21] [22].	MOABCQ	There may be trade-offs between objectives, and these should be properly weighed. Extended schedule periods due to slower convergence can affect overall system responsiveness and performance.

III. PROPOSED METHODOLOGY

In this section, we will go over how to enhance load balancing to guarantee cloud computing in this section [18] [19]. Improving application performance through faster response times and lower network latency is the primary goal shared by all users. The suggested method is applied to enhance load balancing performance to resolve this issue. This section describes the methods to make load balancing better.

In Fig. 1, the algorithm of load balancing starts by initializing the distributed system and carefully determining the nodes and their capacity as well as the tasks executed on every node. The load balancing algorithm begins by setting up the distributed system and then calculating the nodes and their capacity together with the tasks that are on each node. When the system detects the disparities in workload distribution load balancing is activated before the activation of load balancing it must meet the set of conditions or its threshold value. The process of monitoring the task priorities, workloads and system status continues with the emphasis on gathering information on resource usage and other metrics. Tasks are distributed according to predetermined standards, which can be critical or urgent, which means that priority tasks are processed in a balancing load procedure. An algorithm is then introduced to analyze the workload of each node to account for prioritized tasks and computing load. Considering the priorities of the tasks and the general load of the system, the purpose of this decision process is to compare the current state of the given system with the set thresholds or requirements. Tasks are chosen for migrations considering the current workload distribution and task priorities, aiming to transfer less important tasks from nodes with high load to nodes with low load. Tasks should be selected for migration based on the current workload distribution and task priorities so that less important tasks can be moved from burdened nodes to less burdened nodes. A relocation strategy is formulated, so that the tasks are migrated along with their priority information intact, and that could involve tasks segmenting or transferring them in their entirety. Communication among the nodes involved in the migration process guarantees the process is done effectively so that the data and state information are transferred smoothly. Tasks' execution on target nodes obeys their priorities. There is ongoing monitoring of the system responses to the load balancer. Based on the results, necessary changes are made in the balancing load strategy. The feedback medium is set to continue to ameliorate the cargo balancing strategy by conforming to dynamic system conditions and returning to the regulator when necessary. Eventually, the termination condition, which is either balancing load or reaching a system stability thing, decides when the algorithm stops working.

A. Proposed DPBRR Algorithm

Initialization: Count the number of nodes in the distributed system, then also the capacities of the nodes and the tasks that are running on each node. Establish a threshold or set of requirements to kick off load balancing once imbalances are discovered.

1) Monitoring: Monitor the workload, tasks priorities and system status around the clock. Find out the data on resource utilization and other important indicators.

2) Task prioritization: Tasks should be sorted according to the priority levels that are defined beforehand such as criticality, importance, or urgency. While balancing load, tasks with more priority should be given preference.

3) Load evaluation: Look at the workload on each node at the instant, considering the priorities of the assigned tasks and the load of the computations.

4) Decision making: Evaluate the necessity of load balancing by comparing the current condition of the system with the previously defined boundaries or limitations. Involve yourself in the decision-making process by considering both task priorities and the overall burden.

5) Task selection: Ascertain which jobs need to be transferred in agreement with the current load allocation and their priorities. Firstly, the movement of less significant tasks from overworked to underworked nodes should be given the highest priority.

6) Migration strategy: Create a task migration plan which should be done with the consideration of priority information. Such a division of labor could be achieved by breaking the task into smaller pieces or by shifting the entire task.

7) Communication: Let know the concerned nodes about the migration plan. Ensure that all the required data and state information is transferred without a glitch.

8) Task execution: Ensure the priority of the assigned priorities when performing the migrated tasks on the target nodes. In addition, monitor the system's reaction and adjust as required.

9) Feedback loop: Design a feedback mechanism that allows for the continuous adjustment of the load-balancing strategy as the system adapts. Repeat the method and return to the observation step.

10) Termination condition: Give the algorithm a stop condition. It might be based on attaining a predetermined degree of load balance, system stability or other specific requirements.

Algorithmic steps for the proposed DPBRR Algorithm

```
Step 1: Develop the Task Priority Queue where each task has priority level.
Step 2: Implement the Round Robin server list by yourself.
Step 2.1: Each server receives a unique server ID.
Step 2.2: To start the process, the CSI (Current Server Index) must be set to 0.
Step 3: The Receive Incoming Requests and Tasks step is the most important step.
Step 3.1: If the task is a high priority, then put it into the priority queue and check whether the resources are available or not.
Step 4: Monitor the Priority Queue and the resources availability all the time.
Step 4.1: Work on the server assigned to you to complete the task and check the status of the task from time to time.
Step 5: When the job is done, disconnect from the server or resource.
Step 5.1: If there are any outstanding high-priority tasks in the Priority Queue, direct the next available server to the highest-priority task.
Step 6: After the high-priority tasks are completed, servers get distributed to lower-priority tasks in Round Robin manner.
Step 6.1: Therefore, we must increase the CSI (server index) while considering resource availability and fairness.
Step 7: Keep an eye on the status of the newly arrived tasks, resource availability, and task completions.
Step 8: Create the termination conditions and terminate the algorithm when the termination criteria are fulfilled.
```

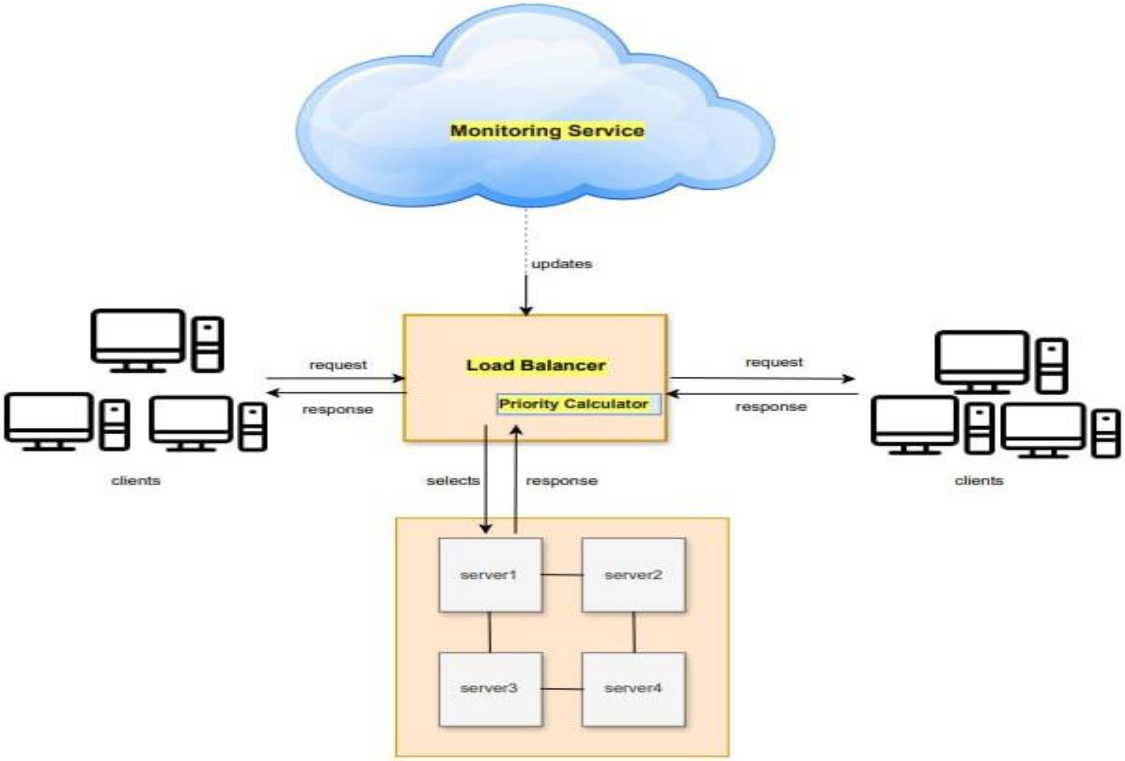


Fig. 1. A Proposed architecture for DPBRR algorithm to enhance load balancing.

- p be the priority
- a be the priority importance
- b be the resource that is available
- c be the load on the system
- q be the index of the resource
- s be the starting index
- j be the particular priority
- T be the total available resources
- Np be the new priority = Np
- Op be the old priority = Op

a) Priority Assignment Equation:

$$p = \alpha * a + \beta * b + \gamma * c$$

b) Task Distribution Equation:

$$q = (s + \Sigma(j)) \% T$$

c) Load Balancing Adjustment Equation:

$$Np = Op + \Delta p$$

These formulas will demonstrate how to compute the available resources, task allocation, and priority factors. When DPBRR applies the equation to distribute tasks evenly among resources based on priority, the jobs with higher priorities will become more significant. Tasks are dynamically altered, and priorities are modified based on changes in priorities as determined by the equations. These suggested equations will aid in effective resource usage and balanced workload distribution within the system.

IV. Result Analysis

This paragraph presents the suggested load-balancing technique as well as experimental results for various load-balancing algorithms. Several parameters are taken into consideration when analyzing it, such as throughput, response time, and resource usage. Existing methods like CMODLB, Kuhn-Munkres, and improved WRR were taken into consideration. As given below.

A. Throughput, Response Time, and Resource Utilization

The amount of material or items passing through a system or process is called throughput, The amount of time taken by a system to respond to a request or input is called Response Time, the efficiency in using the resources is called Resource Utilization, Analysis of CMODLB, Kuhn-Munkres, Improved WRR and proposed DPBRR in Table II, Table III and Table IV.

TABLE II. Computational Throughput Analysis of DPBRR vs. Conventional Algorithms

S.No	Algorithms	Throughput (requests/second)
1	Improved WRR	900 tasks/s
2	Kuhn-Munkres	950 tasks/s
3	CMODLB	980 tasks/s
4	DPBRR	1000 tasks/s

Here, it explores how DPBRR can more efficiently raise the throughput value than other algorithms.

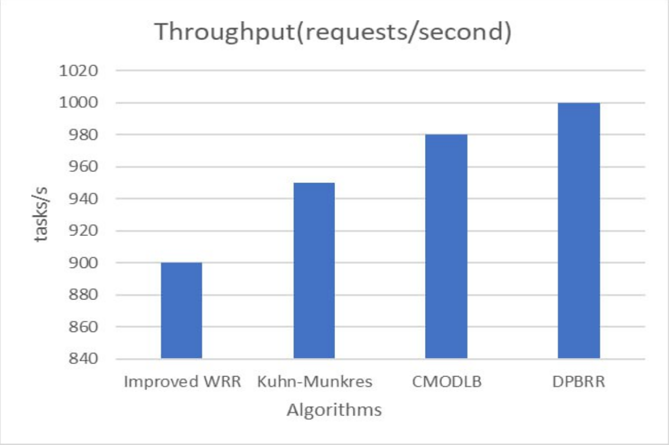


Fig. 2. Computational throughput analysis of DPBRR vs conventional algorithms.

Fig. 2 shows how DPBRR achieves a perfect spectrum since it is faster than other algorithms like Improved WRR, Kuhn-Munkres, CMODLB with a Throughput over 1000 tasks/second. The multi-node adaptive priority-oriented scheduling of this algorithm likely optimizes the resource allocation, with the consequent minimization of latency and improvement in throughput. This benefits DPBRR as businessmen's preferred compute tool in cases where fast processing, well-timed response, and request handling are taken as top priorities, hence, we are distinguishing it as the most efficient algorithm from the others.

Here, it explores how DPBRR can more efficiently decrease the Response time value than other algorithms.

TABLE III. Computational Response Time of DPBRR vs. Conventional Algorithms

S.No	Algorithms	Response Time (milliseconds)
1	Improved WRR	60ms
2	Kuhn-Munkres	55ms
3	CMODLB	52ms
4	DPBRR	50ms

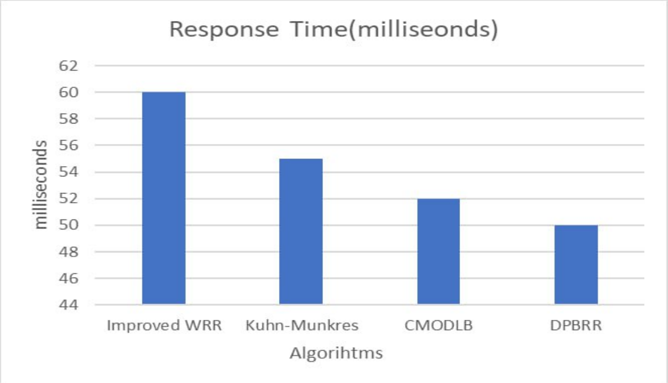


Fig. 3. Computational response time of DPBRR vs. conventional algorithms.

In analyzing the response times of various algorithms from Fig. 3, it is apparent that DPBRR emerges as the most efficient choice. When comparing the reaction times across different algorithms, DPBRR stands out as the most effective option. While Improved WRR exhibits a response time of 60 milliseconds, DPBRR demonstrates a significant improvement with a response time of only 50 milliseconds. Both Kuhn-Munkres and CMODLB show enhancements over Improved WRR, with response times of 55 and 52 milliseconds, respectively. However, DPBRR surpasses them all, showcasing its remarkable ability to handle requests swiftly and minimize system latency. This outstanding performance underscores DPBRR's suitability for scenarios where quick reaction times are crucial, establishing it as the optimal algorithm for maximizing system performance and ensuring prompt request processing.

Here, it explores how DPBRR can more efficiently optimize the Resource utilization value than other algorithms.

TABLE IV. RESOURCE UTILIZATION OF DPBRR VS. CONVENTIONAL ALGORITHMS

S.No	Algorithms	Resource Utilization (%)
1	Improved WRR	75%
2	Kuhn-Munkres	78%
3	CMODLB	72%
4	DPBRR	80%

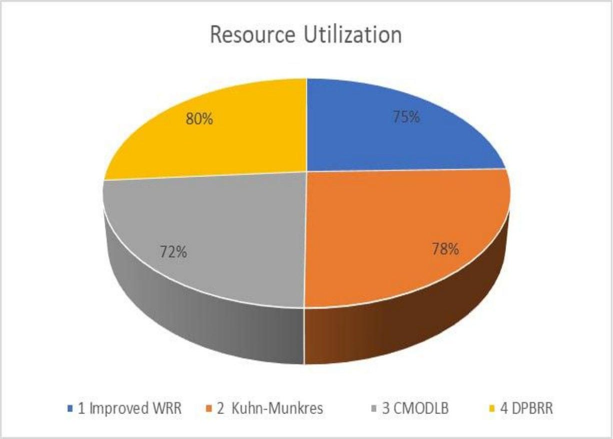


Fig. 4. Resource utilization of DPBRR vs conventional algorithms.

From Fig. 4, the comparisons of resource usage of different algorithms show trends in their effectiveness. The improved WRR consumes resources of 75%, while Kuhn-Munkres raises to 78%. But from the listed algorithms CMODLB has the highest resource consumption of all other algorithms. DPBRR manages to maintain a respectable 80% utilization rate by striking a compromise between resource efficiency and performance. It efficiently optimizes resource allocation and achieves competitive performance. DPBRR is a great option where increasing performance while optimizing resource utilization is crucial because of its balanced resource usage. It manages the system resources even though it does not have the lowest utilization rate.

V. CONCLUSION

In conclusion, the study of Dynamic Priority-Based Round Robin (DPBRR) load balancing is a useful way to solve the important problem of efficient load management in cloud computing settings. DPBRR aims to reduce response time and maximize resource utilization by implementing a dynamic adaptive mechanism that intelligently distributes work across cloud servers. Because DPBRR can dynamically adjust task priorities in response to real-time data, it is particularly useful for workload fluctuations and changing resource demands. Using advanced load balancing technology in virtualized computing, dynamic priority derived from Round Robin is a significant step forward in managing workload imbalances in distributed systems. The recommended approach shows cloud computing load balancing solutions with better performance and economy. By implementing several techniques, DPBRR helps reduce congestion and improve system flexibility and adaptability. The DPBRR study offers a functional and effective way to improve system performance and resource allocation in cloud services, solve important problems, and soon open the door for further developments.

REFERENCES

[1] Pachipala, Y., Dasari, D.B., Rao, V.V.R.M., Bethapudi, P., Srinivasarao, T. Workload prioritization and optimal task scheduling in cloud: introduction to hybrid optimization algorithm (2024) Wireless Networks.

[2] Kavitha .G, Shahbaz Afzal, ‘Load balancing in cloud computing -A hierarchical taxonomical classification.’, Open Access, 2019.

[3] Karimunnisa, S., Pachipala, Y. Deep Learning Approach for Workload Prediction and Balancing in Cloud Computing (2024) International Journal of Advanced Computer Science and Applications, 15 (4), pp. 754-763.

[4] M. Junaid, A. Sohail, A. Ahmed, A. Baz, I. A. Khan, and H. Alhakami, “A Hybrid Model for Load Balancing in Cloud Using File Type Formatting,” IEEE Access, vol. 8, pp. 118135–118155, 2020, doi: 10.1109/access.2020.3003825.

[5] M. Kumar and S. C. Sharma, “Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing,” Procedia Computer Science, vol. 115, pp. 322–329, 2017, doi: 10.1016/j.procs.2017.09.141.

[6] D. A. Shafiq, N. Z. Jhanjhi, A. Abdullah, and M. A. Alzain, “A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications,” IEEE Access, vol. 9, pp. 41731–41744, 2021, doi: 10.1109/access.2021.3065308.

[7] T. M. Tawfeeg et al., “Cloud Dynamic Load Balancing and Reactive Fault Tolerance Techniques: A Systematic Literature Review (SLR),” IEEE Access, vol. 10, pp. 71853–71873, 2022, doi: 10.1109/access.2022.3188645.

[8] A. Moazeni, R. Khorsand, and M. Ramezanpour, “Dynamic Resource Allocation Using an Adaptive Multi-Objective Teaching-Learning Based Optimization Algorithm in Cloud,” IEEE Access, vol. 11, pp. 23407–23419, 2023, doi: 10.1109/access.2023.3247639.

[9] Y. Lohumi, D. Gangodkar, P. Srivastava, M. Z. Khan, A. Alahmadi, and A. H. Alahmadi, “Load Balancing in Cloud Environment: A State-of-the-Art Review,” IEEE Access, vol. 11, pp. 134517–134530, 2023, doi: 10.1109/access.2023.3337146.

[10] L. Yang, Y. Xia, L. Ye, R. Gao, and Y. Zhan, “A Fully Hybrid Algorithm for Deadline Constrained Workflow Scheduling in Clouds,” IEEE Transactions on Cloud Computing, vol. 11, no. 3, pp. 3197–3210, Jul. 2023, doi: 10.1109/tcc.2023.3269144.

[11] G. Xu, J. Pang, and X. Fu, “A load balancing model based on cloud partitioning for the public cloud,” Tsinghua Science and Technology, vol. 18, no. 1, pp. 34–39, Feb. 2013, doi: 10.1109/tst.2013.6449405.

[12] M. Sohani and S. C. Jain, “A Predictive Priority-Based Dynamic Resource Provisioning Scheme With Load Balancing in Heterogeneous Cloud

Computing," IEEE Access, vol. 9, pp. 62653–62664, 2021, doi: 10.1109/access.2021.3074833.

[13] M. N and P. A, "An Efficient Improved Weighted Round Robin Load Balancing Algorithm in Cloud Computing," International Journal of Engineering &amp; Technology, vol. 7, no. 3.1, p. 110, Aug. 2018, doi: 10.14419/ijet.v7i3.1.16810.

[14] S. Negi, M. M. S. Rauthan, K. S. Vaisla, and N. Panwar, "CMODLB: an efficient load balancing approach in cloud computing environment," The Journal of Supercomputing, vol. 77, no. 8, pp. 8787–8839, Jan. 2021, doi: 10.1007/s11227-020-03601-7.

[15] W. Hou, L. Meng, X. Ke, and L. Zhong, "Dynamic Load Balancing Algorithm Based on Optimal Matching of Weighted Bipartite Graph," IEEE Access, vol. 10, pp. 127225–127236, 2022, doi: 10.1109/access.2022.3226885.

[16] J. Zhang, H. Guo, F. Hong, X. Yuan, and T. Peterka, "Dynamic Load Balancing Based on Constrained K-D Tree Decomposition for Parallel Particle Tracing," IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 954–963, Jan. 2018, doi: 10.1109/tvcg.2017.2744059.

[17] W.-K. Chung, Y. Li, C.-H. Ke, S.-Y. Hsieh, A. Y. Zomaya, and R. Buyya, "Dynamic Parallel Flow Algorithms With Centralized Scheduling for Load Balancing in Cloud Data Center Networks," IEEE Transactions on Cloud Computing, vol. 11, no. 1, pp. 1050–1064, Jan. 2023, doi: 10.1109/tcc.2021.3129768.

[18] X. Xu, R. Mo, F. Dai, W. Lin, S. Wan, and W. Dou, "Dynamic Resource Provisioning With Fault Tolerance for Data-Intensive Meteorological Workflows in Cloud," IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 6172–6181, Sep. 2020, doi: 10.1109/tii.2019.2959258.

[19] Karimunnisa, S., Pachipala, Y. Task Classification and Scheduling Using Enhanced Coot Optimization in Cloud Computing (2023) International Journal of Intelligent Engineering and Systems, 16 (5), pp. 501-511.

[20] J.-P. Yang, "Elastic Load Balancing Using Self-Adaptive Replication Management," IEEE Access, vol. 5, pp. 7495–7504, 2017, doi: 10.1109/access.2016.2631490.

[21] Bhargavi, M., Pachipala, Y. Enhancing IoT Security and Privacy with Claims-based Identity Management (2023) International Journal of Advanced Computer Science and Applications, 14 (11), pp. 822-830.

[22] B. Kruekaew and W. Kimpan, "Multi-Objective Task Scheduling Optimization for Load Balancing in Cloud Computing Environment Using Hybrid Artificial Bee Colony Algorithm With Reinforcement Learning," IEEE Access, vol. 10, pp. 17803–17818, 2022, doi: 10.1109/access.2022.3149955.

[23] Aakisetti, R.S.K., Ganta, V., Yellamma, P., Siram, C., Gampa, S.H., Brahma Rao, K.V. Dynamic Priority Scheduling Algorithms for Flexible Task Management in Cloud Computing (2024) International Journal of Intelligent Systems and Applications in Engineering, 12 (13s), pp. 246-256.

Enhancing BLDC Motor Speed Control by Mitigating Bias with a Variation Model Filter

Abdul Rahman Abdul Majid*
Dept. of Electrical & Computer Engineering-College of Engineering, University of Sharjah, Sharjah, UAE

***Abstract*—Brushless DC motors (BLDC) are integral to a wide array of applications, from electric vehicles to industrial machinery, due to their superior efficiency, reliability, and performance. Effective control of BLDC motors is essential to leverage their full potential and ensure optimal operation. Traditional PID controllers often fall short in handling the nonlinear and dynamic characteristics of BLDC systems, while advanced methods like Active Disturbance Rejection Control (ADRC) introduce additional complexity and cost. This research proposes a Variation Model Filter (VMF) based control system that estimates and compensates for the total bias arising from parameter variations and internal uncertainties. This method simplifies the control process, enhances robustness, and boosts performance without requiring extensive parameter tuning or high costs. Additionally, the paper provides a comprehensive mathematical model for the speed dynamics of BLDC motors. Simulation results based on MATLAB/Simulink indicate that the VMF-based PID control system surpasses both linear ADRC and traditional PID controllers in managing speed dynamics and responding to load disturbances. This approach offers an efficient and cost-effective solution for BLDC motor speed control, with significant potential for broader application and further optimization in motor control systems.**

Keywords*—EV's motors; brushless direct current (BLDC) motor; active disturbances rejection control (ADRC); disturbance rejection; bias estimation; Variation Model Filter (VMF)

I. Introduction

Recent advancements in magnet technology have significantly enhanced the performance and efficiency of brushless DC (BLDC) motors, especially those using permanent magnets. These improvements have driven the growing preference for BLDC motors in various applications due to their high power density and energy efficiency [1]. Unlike their brushed counterparts, BLDC motors provide reliable and smooth operation, precise speed control, and reduced electrical noise, making them ideal for dynamic uses such as robotics and automation [2], [3]. In the electric mobility sector, BLDC motors are increasingly chosen for eco-friendly vehicles, including electric cars, scooters, and urban air transport [3]. Additionally, traditional ceiling fans that typically use split-phase induction motors (SPIMs) are now adopting BLDC motors to benefit from better energy efficiency and voltage regulation [4]. BLDC motors are poised to replace traditional induction motors in various industries, such as automotive, pumping, and rolling, by 2030 due to their superior torque, low noise, simplicity, and ease of maintenance [2]. The market for BLDC motors is expected to grow significantly, reaching an estimated value of 15.2 billion USD by 2025 [2].

Despite these advantages, BLDC motors face challenges, such as limited fault tolerance, high electromagnetic interference, acoustic noise, and torque ripple. They operate as complex, multivariable systems with load perturbations and parameter variations, leading to significant current ripple due to factors like armature reaction and phase conversion. To address these issues and enhance BLDC motor performance, researchers have focused on three main areas: motor material and structure, power electronics and drive circuit topologies, and control systems. Recent advancements in motor design include the development of the spherical brushless DC (SBLDC) motor [3], while innovations in power electronics, such as the switched-inductor Zeta active power factor correction converter (SI-ZS-APFC) [4] and phase current overlap time limiting cell (PCOTLC) [5], have significantly reduced torque ripple. Control systems remain critical for BLDC motors, as they manage the motor's operation and ensure the effectiveness of other enhancements.

Two primary control strategies are used for BLDC motor systems: scalar and vector controllers [2]. Scalar controllers include methods like Proportional-Integral-Derivative (PID), Linear Quadratic Regulator (LQR), and Active Disturbance Rejection Control (ADRC). Vector controllers encompass techniques such as Field-Oriented Control (FOC), Direct Torque Control (DTC), and intelligent methods like Particle Swarm Optimization (PSO) and Model Predictive Control (MPC). Among these, PID controllers are popular in industrial applications due to their simplicity and ease of implementation. However, PID controllers struggle with the nonlinear and dynamic nature of BLDC systems, often resulting in suboptimal performance as they are designed for linear systems and cannot handle rapid parameter changes or load disturbances effectively [1], [2], [6].

Vector control methods, including FOC and DTC, and intelligent techniques like PSO and MPC, offer improved dynamic performance and reduced torque and flux ripples. However, they also introduce higher structural and computational complexity. The literature on BLDC motor controllers lacks comprehensive surveys that compare these advanced control schemes, especially in terms of fault tolerance and reducing electromagnetic interference [2]. ADRC, particularly its linear version (LADRC), offers a balance between simplicity and performance. It estimates and compensates for total disturbances using the Extended State Observer (ESO), making it robust against both internal and external perturbations without relying on specific system models [7], [8].

ADRC, introduced by Jingqing Han in the 1990s and further refined into its linear form (LADRC) by Zhiqiang Gao [9], [10] simplifies parameter tuning and implementation for both Single-Input Single-Output (SISO) and Multi-Input Multi-Output (MIMO) systems. While LADRC is easier to implement, it may not perform as well as the nonlinear version (NLADRC) in complex, nonlinear systems. To bridge this gap, researchers have combined LADRC with artificial neural networks for nonlinear state-error feedback control (SEFC) and used intelligent techniques, like genetic algorithms, for optimal parameter tuning [8], [11]. Also, the authors in study [8] proposed to use neural-network as nonlinear SEFC and linear ESO to estimate the total disturbance, they proposed an intelligent version of ADRC (IADRC). Applications of ADRC in BLDC motor control demonstrate its capability in managing speed and current effectively [7], [12].

The study in [13] asserts that PID controllers are adequate for BLDC motor speed control, comparing them with PI and Fuzzy Logic Controllers (FLC). However, their conclusions are limited due to the absence of analysis on load disturbance response, a critical factor in real-world applications. Moreover, the study did not benchmark PID against robust controllers like Active Disturbance Rejection Control (ADRC), focusing only on traditional methods. The PID gains were also unusually high, questioning their cost-effectiveness and practical applicability. In contrast, research in [7] and [12] highlights more resilient approaches using nonlinear and linear ADRC, offering superior disturbance rejection and robustness, thus better addressing the nonlinear and dynamic challenges in BLDC motor systems.

Nevertheless, ADRC, including its linear form, faces challenges such as estimation errors in total disturbances and uncertainty in system parameters, which can impact control gains and lead to noise amplification [14]. In industrial applications, especially with available mathematical models and adequate sensors, simpler solutions like Disturbance Observer Based Control (DOBC) may suffice. Research suggests that DOBC can outperform ADRC in certain conditions due to its simpler design process [15].

This paper aims to enhance the robustness of PID control for BLDC motors by introducing a Variation Model Filter (VMF)-based approach. Unlike traditional PID controllers, which often require high gain settings and struggle with load disturbances and system uncertainties, the VMF method estimates the total bias within the system. This bias accounts for energy variations caused by parameter changes, internal uncertainties, and external disturbances. By compensating for these factors using a robust control law based on output variations from the desired response, the VMF-enhanced PID control demonstrates superior performance. Simulation results indicate that this new approach not only surpasses traditional PID controllers but also outperforms linear Active Disturbance Rejection Control (LADRC) in managing BLDC motor dynamics effectively.

The paper is structured as follows: Section II provides a detailed mathematical model of the BLDC motor. Section III introduces the VMF approach. Section IV presents simulation results, and Section V concludes the paper.

II. Problem Formulation

To model the voltage in a three-phase BLDC motor, we start with the voltage equation for each winding. This relationship between voltage, current, and back EMF is given by:

$$\begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} = \begin{bmatrix} R_a & 0 & 0 \\ 0 & R_b & 0 \\ 0 & 0 & R_c \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} L_a & L_{ab} & L_{ac} \\ L_{ba} & L_b & L_{bc} \\ L_{ca} & L_{cb} & L_c \end{bmatrix} \begin{bmatrix} \frac{di_a}{dt} \\ \frac{di_b}{dt} \\ \frac{i_c}{dt} \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \tag{1}$$

To simplify the modeling of a BLDC motor, several key assumptions are typically made. First, the stator windings are designed as full-pitch windings with a 60-degree phase belt. The air gap magnetic field is assumed to have a trapezoidal distribution with a flat top spanning 120 electrical degrees. We also disregard effects like slot impact, magnetic circuit saturation, magnetic hysteresis, eddy current loss, skin effect, and temperature influence on motor parameters.

Further simplifying, we consider the reluctance between the stator and rotor to be negligible, allowing us to set ; $L_{ab} = L_{ac} = L_{ba} = L_{bc} = L_{ca} = L_{cb} = M$.The three-phase windings are Y-connected and assumed to be symmetric, leading to $R_a = R_b = R_c = R$; $L_a = L_b = L_c = L$; and the condition $i_a + i_b + i_c = 0$; $Mi_b + Mi_c = -Mi_a$. Additionally, mechanical losses and other incidental losses of the motor are not taken into account.

Based on these assumptions, especially the symmetry in windings and negligible reluctance, the voltage equation simplifies to:

$$\begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} = \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} L-M & 0 & 0 \\ 0 & L-M & 0 \\ 0 & 0 & L-M \end{bmatrix} \begin{bmatrix} \frac{di_a}{dt} \\ \frac{di_b}{dt} \\ \frac{i_c}{dt} \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \tag{2}$$

To derive the current dynamics, we apply Kirchhoff's law to the BLDC motor's equivalent circuit, shown in Fig. 1.

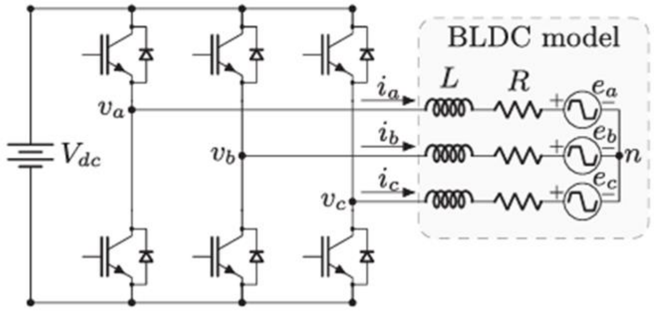


Fig. 1. BLDC motor equivalent drive circuit.

This gives us the following equation for the stator current:

$$\frac{di_s}{dt} = -\frac{R}{L-M} i_s + \frac{1}{2(L-m)} v_L - \frac{1}{2(L-m)} e_L \tag{3}$$

Here, i_s is stator current represents the stator current, while $v_L \in \{v_{ab}, v_{bc}, v_{bc}\}, e_L \in \{e_{ab}, e_{bc}, e_{bc}\}$ are the line-to-line voltage and back EMF, respectively, across different phases.

Given the assumptions about the stator windings and air gap magnetic field, and considering the current dynamic equation, it is feasible to control all three output currents with a single current controller. This leads us to a simplified first-order equation for the current dynamics:

$$\frac{di(t)}{dt} = f_1(i, d_e, t) + b_1 v \qquad (4)$$

In this equation, $f_1(.)$ encapsulates all uncertainties and disturbances in the current dynamics, with d_e representing electrical disturbances in the system given by $d_e = -\frac{1}{2(L-m)} e_L - \frac{R}{L-M} i_s$. The term b_0 is the nominal input gain for current control, equal to $\frac{1}{2(L-m)}$.

Next, we examine the electromagnetic torque, which results from the interaction between the stator winding currents and the rotor's magnetic field. This torque is defined as $T_e = \frac{e_a i_a + e_b i_b + e_c i_c}{\omega_m}$, where ω_m is the rotor speed. Assuming that all the electromagnetic power is converted into rotor kinetic energy, the power equation can be simplified to $P_e = T_e \omega_m$. At steady state, this power can be further expressed as $P_e = \frac{2E_s I_s}{\omega_m} = k_t I_s$. In this context, E_s represents the opposing electromotive force, I_s is the steady-state phase current (max amplitude), andk_t is the motor torque coefficient.

The mechanical movement of the BLDC motor is described by the following equation, which balances the generated torque T_e and the load torque T_L:

$$T_e - T_L = \frac{J d\omega_m}{dt} + B\omega_m \qquad (5)$$

where J denotes the rotary inertia, and B is the damping coefficient. By differentiating this equation and integrating the electromagnetic torque, we derive the second-order speed dynamic equation:

$$\ddot{\omega}_m = \frac{k_t}{J}\frac{di}{dt} - \frac{1}{J}(\dot{T}_L + B\dot{\omega}_m) \qquad (6)$$

Rewriting it more compactly, we get:

$$\ddot{\omega}_m = f_2(\omega_m, d_m, t) + b_2 \frac{di}{dt} \qquad (7)$$

Here, $f_2(.)$ encompasses all uncertainties and disturbances affecting the speed dynamics, with d_m representing mechanical disturbances. The term b_2 is the nominal input gain for speed control, equal to $\frac{k_t}{J}$.

Combining the inner and outer control loops, we propose a comprehensive second-order dynamic model for the BLDC motor speed. By substituting the current dynamics into the speed dynamics, the resulting equation is:

$$\ddot{\omega}_m = f_t(f_1, f_2) + b_0 u(t) \qquad (8)$$

In this final equation, $f_t(.)$ aggregates all the electrical and mechanical uncertainties and disturbances, defined as $f_t = b_2 f_1 + f_2$. The nominal input gain b_0 is given by $b_0 = b_1 b_2 = \frac{k_t}{2J(L-M)}$. The Eq. (4) and Eq. (8) highlight that controlling a BLDC motor involves managing a multi-loop system with an inner loop for current regulation and an outer loop for speed control, as shown in Fig. 2.

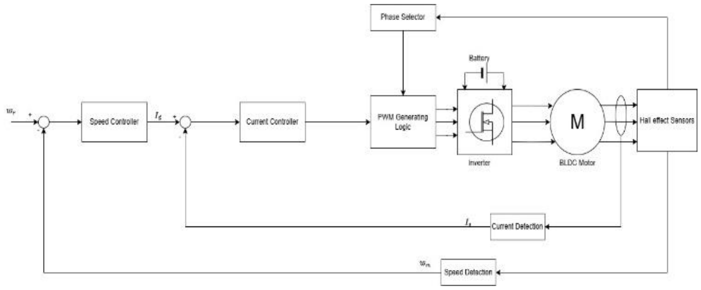


Fig. 2. BLDC motor complete system.

In controlling the speed of BLDC motors, the ADRC framework typically relies on the dynamic equation similar to Eq. (7), where an Extended State Observer (ESO) estimates and compensates for total disturbances primarily in the mechanical domain and some internal uncertainties [7], [12]. However, traditional ADRC designs often overlook electrical disturbances in the current loop, leading researchers to propose a dual-loop ADRC system [12]: one loop for managing electrical disturbances in the current and another for handling mechanical disturbances in the speed. This dual-loop approach provides a more holistic control but adds complexity. Conversely, the bias rejection strategy simplifies the process by focusing only on disturbances and uncertainties that directly impact the dominant state, termed "total bias." This technique, rather than estimating all potential disturbances, concentrates on the most significant ones, making it faster, more cost-effective, and reliable by addressing only the critical issues affecting the system's performance.

III. VARIATION MODEL FILTER (VMF)

To estimate bias accurately in a system, it's essential to develop a mathematical technique that uses a variation function for approximation. This approach involves employing a continuous approximation method based on integral transform to effectively determine and estimate the total bias. The continuous approximation can be expressed through the integral transform:

$$g(s) = \int k(s,t) f(t) dt \qquad (9)$$

Here, $g(s)$ represents known information that might be influenced by noise, while $k(s,t)$ is the kernel function, and s indicates a specific domain. The unknown function $f(t)$ needs to be solved or approximated. In control applications, the system typically operates within defined constraints and spatial domains, making the first kind of Fredholm integral equation suitable:

$$\int_a^b k(s,t) f(t) dt = g(s), \quad c \le s \le d \qquad (10)$$

In this context, $g(s)$ is given within the range $[c, d]$, which may differ from the integral's range $[a, b]$. The kernel function basis technique is often used to smooth and regularize noisy data, minimizing errors over the approximation space. For example, if $f \in L_2[0,1]$, the Riesz representation theorem guarantees a function $\eta_s(t) \in L_2[0,1]$ such that:

$$\int_0^1 \eta_s(t) f(t) dt = \langle \eta_s, f \rangle,\ s = s_1, s_2, \dots, s_n \quad (11)$$

Thus, $\eta_s(t) = k(s,t)$, indicating an appropriate basis can be deduced, and $f(t)$ can be approximated by:

$$f^*(t) = \sum_{i=1}^{n} a_i \eta_s(t) \quad (12)$$

To approximate the total bias in a control system, which represents the unwanted energy within the closed-loop system, leading to output variation from the desired value, we use the general continuous approximation method:

$$N(\delta f, x) = \int k(x,t) \delta f(t) dt \quad (13)$$

Here, $N(\delta f, x)$ is a nonlinear function representing the system's closed-loop output, $k(x,t)$ is the kernel function, and $\delta f(t)$ is the total bias to be estimated. Given engineering applications' spatial and constraint assumptions, the Fredholm integral of the first kind and the Riesz representation theorem provide suitable frameworks. Thus, we propose a linear method for bias estimation:

$$\delta f(t) = \sum_{i=1}^{n} a_i k(x_i, t) = \sum_{i=1}^{k} c v_i(t) = c \sum_{i=1}^{k} v_i(t) (14)$$

This assumes $\delta f(t)$ can be approximated by a finite set of basis functions $v_i(t)$.We can rewrite this method as:

$$\delta f(t) = cV(t) \quad (15)$$

Where $V(t)$ represents the system's total variation, summing all basis functions or a suitably chosen function that provides variation information, and c relates to the regularization gain to enhance robustness. The selection of basis functions is crucial; practical applications often benefit from intuitive and knowledge-based techniques rather than purely mathematical solutions. For dynamic system input smoothing, we use a constant regularization:

$$c = \frac{c_0}{\hat{b}} \quad (16)$$

Where c_0 is a positive constant and $\hat{b}$ is the approximated input gain. The bias estimation method using the operator form in Eq. (15) requires two types of information to find the variation function: global (general) discrepancy and smoothed (processed) discrepancy. These discrepancies reflect how the output deviates from the desired value, providing insights into the variation function. The global discrepancy $n_g(t)$ is calculated as:

$$n_g(t) = y_p(t) - y_m(t) \quad (17)$$

Where $y_p(t)$ is the system's original output and $y_m(t)$ is the reference model or observer output. The smoothed discrepancy $n_s(t)$ is found by filtering or smoothing the actual plant and model outputs:

$$n_s(t) = z_p(t) - z_m(t) \quad (18)$$

Here, $z_p(t)$ and $z_m(t)$ are the filtered outputs of the plant and model, respectively. The variation function $V(t)$ is then approximated as:

$$V(t) = n_g(t) + n_s(t) \quad (19)$$

In summary, variation reflects output deviation due to unwanted energy (bias) in a system. In a closed-loop system, bias arises from varying parameters, uncertainties, noise, and coupling states. The variation model captures this as a function, which serves as a basis to estimate and compensate for the total bias within the system, thereby improving control performance. Fig. 3 shows the Variation Model Filter.

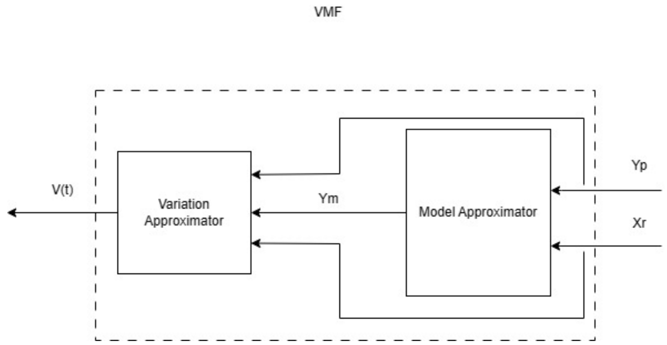


Fig. 3. Variation Model Filter (VMF).

IV. Simulation Results and Discussion

To demonstrate the effectiveness of the proposed control technique, the system was implemented using a MATLAB library example [16], showcasing the capabilities of an ADRC (Active Disturbance Rejection Control) system. This setup allowed for a practical comparison that users could replicate and test independently. The system parameters are listed in Table I, providing the necessary context for the implementation.

TABLE I. System Parameters

Parameter	Value
Number of pole-pair	4
Stator resistance R ($m\Omega$)	1
Self-inductance L (mH)	0.2
Mutual-inductance M (mH)	0.02
Moment of inertia J ($kg.m^2$)	2
Damping coefficient B ($N.\frac{m}{\frac{rad}{s}}$)	0.1
Load when implemented (Nm)	0.1

In Fig. 4, the MATLAB/Simulink model is presented for the speed control of a Brushless DC (BLDC) motor. This comprehensive setup is designed to evaluate and compare the performance of different control strategies. The model includes a block for the BLDC motor and two main controllers: Active Disturbance Rejection Control (ADRC) and Proportional-Integral-Derivative (PID). During testing, either the ADRC or PID controller can be activated to manage the system's response.

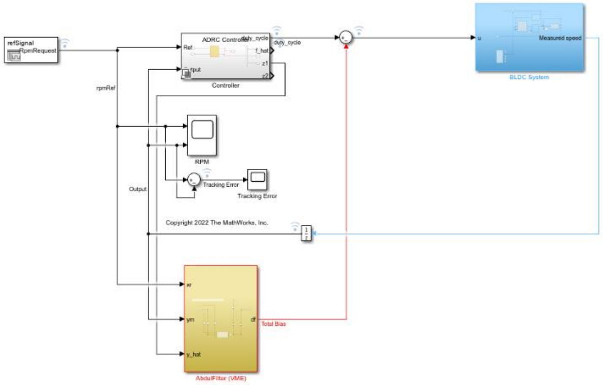


Fig. 4. Simulink Model for speed control of BLDC motor.

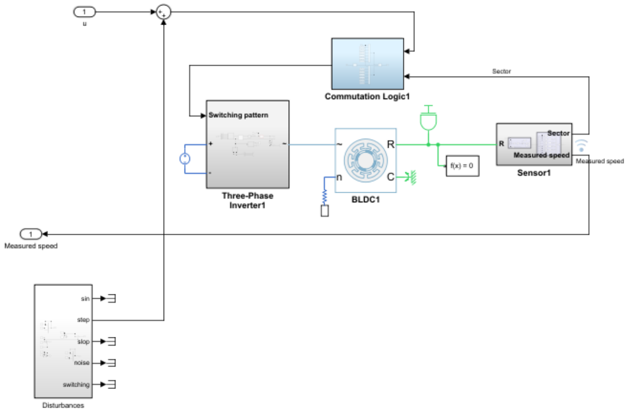


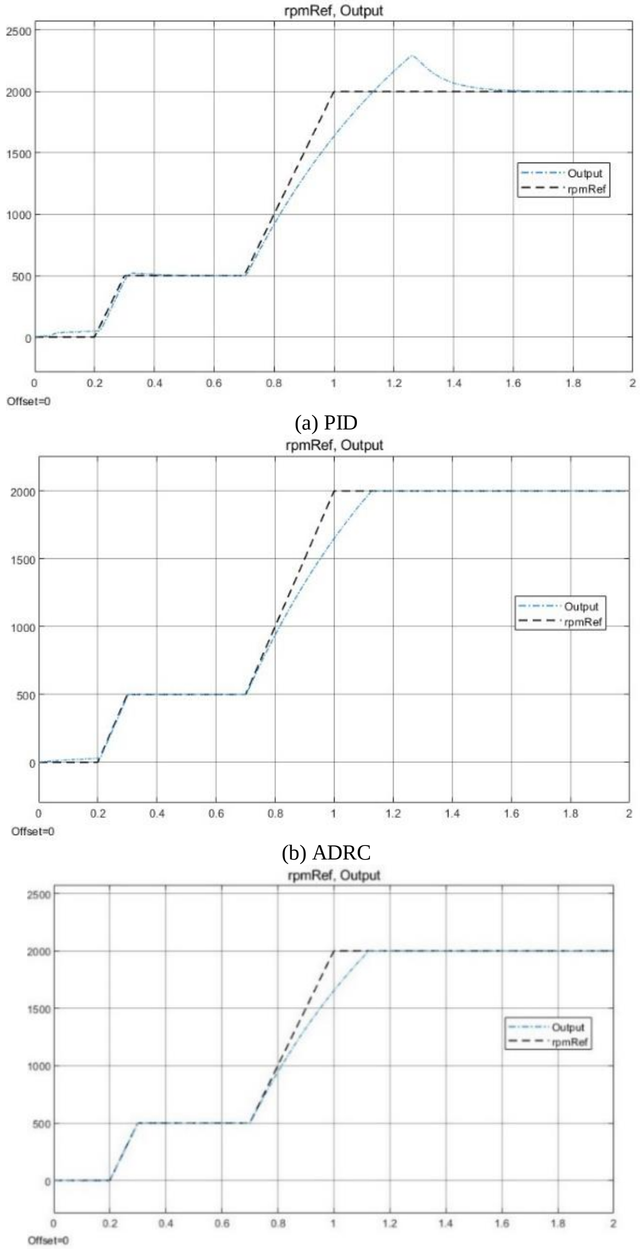
Fig. 5. BLDC motor system.

In Fig. 5, the internal structure of the BLDC motor block is delved within the model. This block encapsulates the dynamics and operational characteristics of the BLDC motor with drive circuit, commutation logic, sensor, and disturbance model, providing a detailed simulation environment for analyzing motor behavior under various control strategies.

The controller block in Fig. 4 is pivotal to the model. It includes both ADRC and PID controllers, where only one is activated at a time to control the motor. The ADRC controller uses an Extended State Observer (ESO) to estimate and counteract disturbances, thereby providing a robust control performance. The output of the ESO, denoted as $\hat{y}$ (in model called y_hat), serves as the reference model output. This output, along with the desired speed signal x_r and the actual motor output y_m feeds into the Variation Model Filter (VMF) block.

The VMF block plays a crucial role by comparing these inputs to estimate the total bias in the system. This estimated bias is then used to enhance the performance of the control system, ensuring accurate and reliable speed control of the BLDC motor.

This setup is essential for analyzing how the different controllers handle the nonlinearities and disturbances inherent in BLDC motor systems, ultimately demonstrating the effectiveness of the VMF-based control approach.



(a) PID

(b) ADRC

(c) VMF based PID

Fig. 6. Speed tracking performance without external load disturbances.

Fig. 6 illustrates the application of three different control strategies to manage the speed of a BLDC (Brushless DC) motor without external disturbances: conventional PID (Proportional-Integral-Derivative) control, ADRC, and PID with the Variable Mode Filter (VMF) technique. As seen in Fig. 6(a) and (b), the ADRC control system significantly outperforms the conventional PID controller. This advantage is due to ADRC's ability to estimate and reject all internal and external disturbances using an Extended State Observer (ESO), which captures these disturbances as "total disturbances”. However, in Fig. 6(c), the PID controller with VMF surpasses ADRC in terms of faster response and eliminating overshoot. Fig. 7 consolidates the performance of all three controllers, providing a single display for direct comparison under the same conditions.

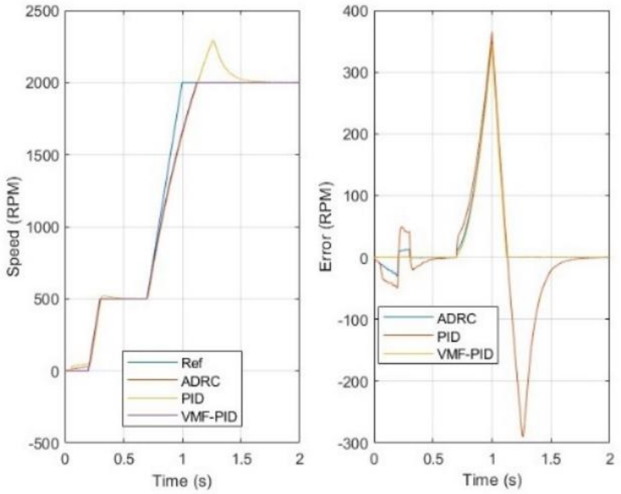


Fig. 7. Performance comparison without external disturbances.

The VMF-based approach, on the other hand, estimates the "total bias" or "discrepancy" in the system. This method simplifies the process by focusing solely on compensating for this total bias, thereby reducing complexity and cost. As illustrated in Fig. 6 and Fig. 7, the PID controller enhanced with VMF not only surpasses the conventional PID controller but also outperforms the ADRC controller in handling system discrepancies.

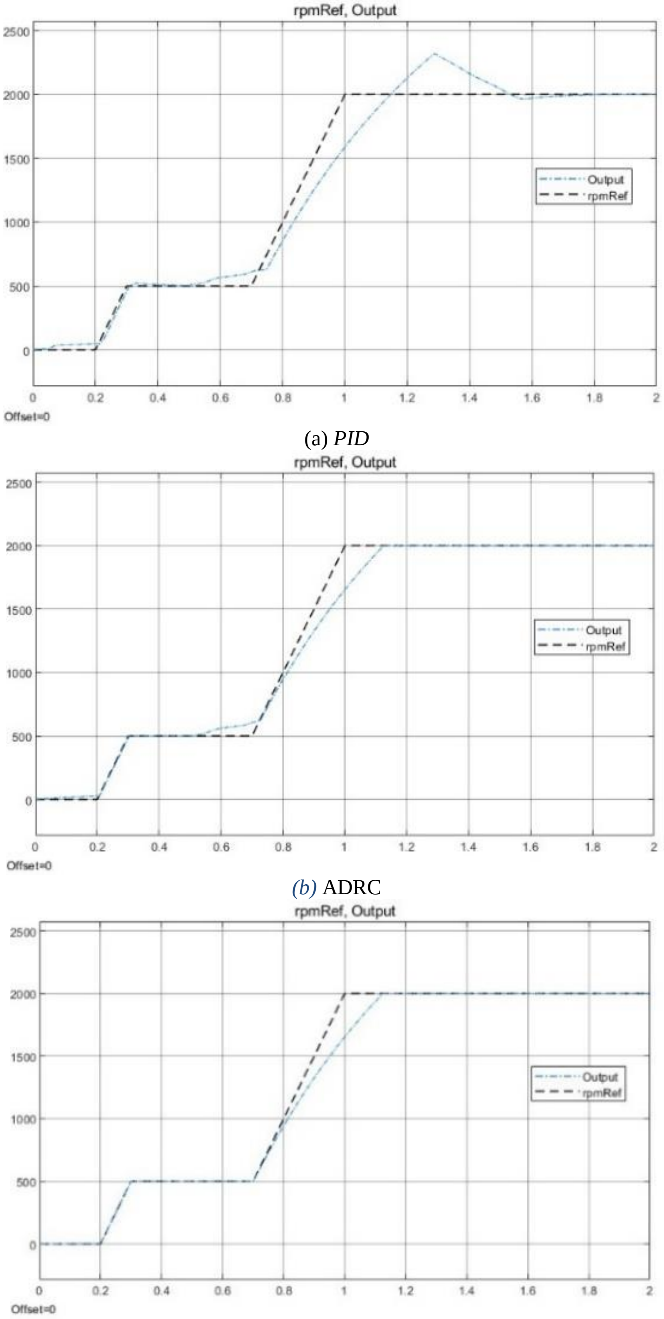
TABLE II. PERFORMANCE COMPARISON

Controller	Rise Time	Settling Time	Overshoot	Normalized MSE
PID	0.1778	0.5011	15.8180	0.0893
ADRC	0.1922	0.1095	0.0570	0.0381
VMF-PID	0.1901	0.1095	0.0058	0.0300

The robustness of these controllers is further evaluated under external disturbances. Table II highlights the performance differences when a load disturbance is applied, specifically observing the transient response as the motor speed changes from 500 rpm to 2000 rpm at 1 second. The comparison includes the normalized mean square error (NMSE), approximated to four decimal places, to quantify the differences among the controllers.

Fig. 8 showcases the speed tracking capabilities of the three control systems when a 0.1 Nm load is applied to the BLDC motor at 0.5 seconds, whereas Fig. 9 shows a performance comparison among them. The VMF-based PID (VM-PID) controller demonstrates superior performance, maintaining more consistent and robust control under load conditions compared to the ADRC controller and the conventional PID controller. This confirms that the VMF-enhanced PID approach is not only effective but also a more reliable solution for controlling BLDC motors, providing better handling of disturbances with less complexity.

Fig. 9 showcases the comparative performance of three different control systems—traditional PID, Active Disturbance Rejection Control (ADRC), and the proposed Variation Model Filter-based PID (VM-PID) controller—under the impact of external disturbances on the BLDC motor. The results indicate that the VM-PID controller significantly outperforms the ADRC and traditional PID controllers when faced with external perturbations. This superior performance is attributed to the VMF's ability to estimate and reject the total bias within the system. By integrating the VMF into the PID controller, it compensates for variations, uncertainties, and disturbances, thus dramatically enhancing the robustness and effectiveness of the traditional PID approach without the complexity associated with ADRC.



(a) *PID*

(b) ADRC

(c) *VMF based PID*

Fig. 8. Speed tracking performance with external load disturbances at 0.5s.

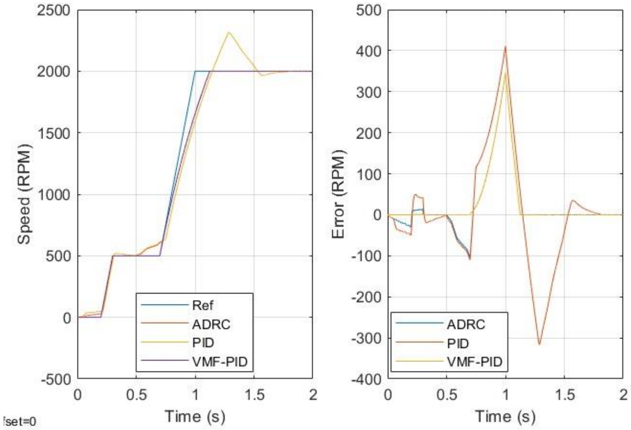


Fig. 9. Performance comparison with external disturbances.

The VM-PID controller's ability to manage and mitigate the effects of external disturbances suggests a promising enhancement over traditional PID control. The VMF allows the system to adapt dynamically to changes and disturbances by continuously estimating and adjusting for the total bias. This makes the VM-PID not only more robust but also simpler and more cost-effective compared to ADRC, which, although effective, requires complex tuning and higher implementation costs. This finding opens up an exciting avenue for further research: the integration of VMF with ADRC. Combining the bias estimation and compensation strengths of VMF with ADRC's adeptness at managing system uncertainties and disturbances could lead to a highly efficient, robust, and adaptive control system for BLDC motors and other applications.

Looking ahead, exploring a hybrid VMF-ADRC approach could provide significant improvements in motor control systems and other robotics systems. Such a combination could leverage the simplicity and robustness of the VMF-based bias compensation with ADRC's sophisticated disturbance rejection capabilities. This integration could result in a control system that not only handles complex dynamics and external disturbances more effectively but also reduces the need for extensive parameter tuning such as nonlinear ADRC cases. Future research could focus on developing and optimizing this hybrid control strategy, potentially setting a new standard for BLDC motor control in applications where both performance and robustness are critical.

It is worth mentioning, that the total bias is related to the dominant states. This means only focusing on the dominant state such as output and estimating anything that affects to those states and this kind of limitation for this technique. Thus, combining this technique with disturbances estimation techniques will dramatically improve the performance of control systems. Thus, it is suggested that in future work combining VMF with ADRC to propose a very robust control system.

In conclusion, the results and comparisons demonstrate the superiority of the VMF technique over traditional methods. The advantages of this approach include:

1) It is a completely model-free method, requiring no prior knowledge of the system model.

2) It enhances both the response and robustness of control systems.

3) The technique is simple to implement and does not require extensive parameter tuning.

V. Conclusion

In conclusion, this paper provides a comprehensive comparative study of control techniques for brushless DC (BLDC) motors, emphasizing a new method based on the Variation Model Filter (VMF). BLDC motors are increasingly popular across various applications due to their high efficiency and performance, necessitating effective control mechanisms to handle their complex dynamics. Traditional PID controllers often struggle with the nonlinear and time-varying characteristics of BLDC systems, resulting in less-than-optimal performance under varying conditions. Although advanced methods like Active Disturbance Rejection Control (ADRC) offer better disturbance management, they introduce additional complexity and cost, which can be a significant drawback in practical applications.

The VMF-based approach proposed in this study provides an innovative solution by focusing on estimating and compensating for the total bias within the system. This total bias encompasses the effects of parameter variations, internal uncertainties, and external disturbances on the dominant states, which are crucial for maintaining robust control. By integrating VMF into the PID control framework, the system can effectively handle these biases, resulting in enhanced robustness and simplified implementation compared to ADRC. Simulation results consistently show that the VMF-enhanced PID controller outperforms both traditional PID and linear ADRC controllers in managing speed dynamics and responding to load disturbances, achieving superior control performance without extensive tuning or added complexity.

This research makes significant contributions by offering a detailed mathematical model for BLDC motor speed dynamics and introducing a new technique for bias estimation. The VMF-based PID control system strikes a balance between performance and simplicity, providing a cost-effective and efficient solution for BLDC motor control. Additionally, this study lays the groundwork for future research, suggesting potential exploration into hybrid approaches that combine the robust bias estimation of VMF with the advanced disturbance rejection capabilities of ADRC. Such combinations could potentially enhance control performance and adaptability further, paving the way for broader applications and optimization of control systems in BLDC motors and beyond.

References

[1] K. Kroičs and A. Būmanis, “BLDC Motor Speed Control with Digital Adaptive PID-Fuzzy Controller and Reduced Harmonic Content,” Energies, vol. 17, no. 6, p. 1311, Mar. 2024, doi: 10.3390/en17061311.

[2] D. Mohanraj et al., “A Review of BLDC Motor: State of Art, Advanced Control Techniques, and Applications,” IEEE Access, vol. 10, pp. 54833–54869, 2022, doi: 10.1109/ACCESS.2022.3175011.

[3] S. Lee and H. Son, “Six Steps Commutation Torque and Dynamic Characteristics of Spherical Brushless Direct Current Motor,” IEEE

Trans. Ind. Electron., vol. 71, no. 5, pp. 5045–5054, May 2024, doi: 10.1109/TIE.2023.3285976.

[4] A. Kumar and B. Singh, “High-Performance Brushless Direct-Current Motor Drive for Ceiling Fan,” IEEE Trans. Ind. Electron., vol. 71, no. 7, pp. 6819–6828, Jul. 2024, doi: 10.1109/TIE.2023.3294649.

[5] U. Soni and R. Tripathi, BLDC motor specific PCOTLC converter with active current wave shaping for torque ripple minimization. 2018, p. 6. doi: 10.1109/ETECHNXT.2018.8385331.

[6] S. Ok, Z. Xu, and D.-H. Lee, “A Sensorless Speed Control of High-Speed BLDC Motor Using Variable Slope SMO,” IEEE Trans. Ind. Appl., vol. 60, no. 2, pp. 3221–3228, Mar. 2024, doi: 10.1109/TIA.2023.3348081.

[7] P. Zhang, Z. Shi, B. Yu, and H. Qi, “Research on the Control Method of a Brushless DC Motor Based on Second-Order Active Disturbance Rejection Control,” Machines, vol. 12, no. 4, p. 244, Apr. 2024, doi: 10.3390/machines12040244.

[8] A. R. A. Majid, R. Fareh, and M. Bettayeb, “Intelligent Active Disturbance Rejection Control for Quadrotor System,” in 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates: IEEE, Nov. 2022, pp. 190–195. doi: 10.1109/ICECTA57148.2022.9990070.

[9] J. Han, “From PID to active disturbance rejection control,” IEEE Trans. Ind. Electron., vol. 56, no. 3, pp. 900–906, 2009.

[10] Z. Gao, “Active disturbance rejection control: From an enduring idea to an emerging technology,” in 2015 10th International Workshop on Robot Motion and Control (RoMoCo), Poznan, Poland: IEEE, Jul. 2015, pp. 269–282. doi: 10.1109/RoMoCo.2015.7219747.

[11] Z. Yan and Y. Zhou, “Application to Optimal Control of Brushless DC Motor with ADRC Based on Genetic Algorithm,” in 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China: IEEE, Aug. 2020, pp. 1032–1035. doi: 10.1109/AEECA49918.2020.9213554.

[12] P. Kumar, A. R. Beig, D. V. Bhaskar, K. A. Jaafari, U. R. Muduli, and R. K. Behera, “An Enhanced Linear Active Disturbance Rejection Controller for High Performance PMBLDCM Drive Considering Iron Loss,” IEEE Trans. Power Electron., vol. 36, no. 12, pp. 14087–14097, Dec. 2021, doi: 10.1109/TPEL.2021.3088418.

[13] M. Mahmud, S. M., A. H., and A. Nurashikin, “Control BLDC Motor Speed using PID Controller,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 3, 2020, doi: 10.14569/IJACSA.2020.0110359.

[14] R. Fareh, M. Al-Shabi, M. Bettayeb, and J. Ghommam, “Robust Active Disturbance Rejection Control For Flexible Link Manipulator,” Robotica, vol. 38, no. 1, pp. 118–135, Jan. 2020, doi: 10.1017/S026357471900050X.

[15] H. V. Nguyen, T. Vo-Duy, and M. C. Ta, “Comparative Study of Disturbance Observer-Based Control and Active Disturbance Rejection Control in Brushless DC Motor Drives,” in 2019 IEEE Vehicle Power and Propulsion Conference (VPPC), Hanoi, Vietnam: IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/VPPC46532.2019.8952367.

[16] “Design Active Disturbance Rejection Control for BLDC Speed Control Using PWM - MATLAB & Simulink.” Accessed: Jun. 02, 2024. [Online]. Available: https://www.mathworks.com/help/slcontrol/ug/design-adrc-for-bldc-motor.html.

Dynamic Monitoring of Bridge Structures via an Integrated Cloud and Edge Computing System

Guoqi Zhang[1], Pengcheng Zhang[2], Xingwang Li[3], Yizhe Yang[4]*
China First Highway Engineering CO., LTD., Beijing 100024, China[1, 2, 3]
College of Civil Engineering, Xuchang University, Xuchang 461000, China[4]

***Abstract*—Traditional bridge monitoring techniques, which predominantly rely on centralized data processing, often exhibit slow and inflexible responses when managing large-scale sensor network data. This study proposes an integrated edge and cloud computing approach to enhance the response time and data processing efficiency of dynamic bridge structure monitoring systems, thereby improving bridge safety and reliability. The proposed monitoring system leverages both edge and cloud computing, incorporating modules such as sensor data management, structural assessment and warning, data processing, monitoring, and data acquisition and transmission. High-performance and cost-effective sensors are utilized to monitor the real-time dynamic responses of the bridge, including displacement, acceleration, tilt, and stress, as well as external loads and environmental effects. The data processing module employs the modal superposition method, frequency response function, and modal analysis for dynamic analysis, while the cloud computing platform facilitates deep learning analysis and long-term data storage. A real case study demonstrates the system's performance across various settings and operational conditions, highlighting the effectiveness of integrating edge and cloud computing. The results indicate that the integration scheme significantly enhances monitoring accuracy, system stability, real-time response capacity, and data processing efficiency.**

Keywords*—Dynamic monitoring of bridge structures; edge computing; cloud computing; data processing; modal analysis

I. INTRODUCTION

As the world's infrastructure ages faster and bridge loads continue to climb, maintaining the durability and safety of bridge structures has become increasingly important for transportation infrastructure managers. Bridges are essential transportation hubs, and public safety and economic prosperity are directly impacted by the state of these structures. Consequently, the importance of monitoring and evaluating bridge health has increased. Recent developments in data collection, processing, and sensing technologies have made dynamic monitoring of bridge structures a hot topic for study. However, a number of challenges face traditional bridge monitoring techniques, including real-time data processing, system adaptability, and data security.

In order to anticipate future failures and safety risks, the goal of health monitoring bridge structures is to gather and evaluate the structural response of bridges in real time. Conventional monitoring techniques typically depend on centralized data processing systems, which frequently experience issues with massive amounts of data, including processing delays, bandwidth bottlenecks, and data loss [1]. Bridge monitoring systems can now collect vast amounts of high-frequency data due to advancements in sensor technology, which place more demands on data processing. Complex sensor data must be processed in real time in modern bridge monitoring in order to assess the structural response of the bridge, identify possible issues, and promptly take corrective action [2]. Thus, it is now a top research priority to investigate new monitoring schemes to guarantee data confidentiality, boost system adaptability, and increase data processing skills.

Many existing structural health monitoring (SHM) systems use traditional wired sensor networks, which are prone to scalability problems. As the number of sensors increases, the complexity of wiring and maintenance grows, leading to higher costs and greater difficulty in system management. Wireless sensor networks (WSNs) have been introduced to address some of these issues, but even WSNs face challenges in terms of signal interference, data loss, and power consumption, especially in large-scale infrastructure like long-span bridges.

While numerous systems claim to offer real-time monitoring, their data processing speeds and transmission methods often lag behind the real-time requirements of critical infrastructures. Most systems are not equipped to handle the massive influx of data generated by high-frequency sampling from multiple sensors, leading to delays in data processing and reporting. Furthermore, interruptions in data transmission due to connectivity issues often result in incomplete or delayed data analysis, making it difficult to monitor structural health accurately in real time.

One major limitation of existing SHM systems is the lack of robust fault tolerance mechanisms. Many systems do not have adequate backup solutions in place to prevent data loss during network outages or hardware failures. The absence of local storage for sensor data during communication interruptions can lead to significant gaps in monitoring, especially during critical events such as natural disasters or severe weather conditions. This undermines the reliability of the data collected and the system's ability to provide timely alerts.

Most traditional SHM systems rely on basic statistical methods for evaluating structural health. While these methods are useful for analyzing deformation, vibration, and load responses, they often fail to provide accurate predictions or insights into long-term structural behavior. The integration of intelligent algorithms such as machine learning, which could predict potential failure points or structural degradation based on historical data, remains underexplored in many existing solutions.

*Corresponding Author

Large volumes of data are produced by bridge monitoring systems, particularly when high frequency sampling is used. Conventional centralized data processing techniques are frequently unable to keep up with the needs of real-time processing [3]. Due to this, data is delayed and any structural anomalies or early warning signals may go unnoticed. Although edge computing, which processes data in real time close to the site of collection, can significantly cut down on transmission delays, it has drawbacks in terms of deep analysis and storage capacity [4, 5]. It is necessary to prevent and address potential security breaches, system malfunctions, and data loss during data transmission and storage with appropriate methods. Temperature, wind speed, traffic volume, and other climatic and operational variables can all have an impact on a bridge's structural response. In order to guarantee the precision and dependability of the monitoring data under diverse circumstances, the monitoring system must be flexible and durable [6, 7].

The three primary data processing methods used in bridge monitoring systems nowadays are distributed data processing, centralized data processing, and edge computing with cloud computing. Among them, centralized data processing techniques mostly depend on a central server for data analysis; however, this approach is less effective at handling large data volumes and is prone to system bottlenecks and data transmission delays [8, 9]. By dividing up the processing work among several sites, distributed data processing techniques boost processing efficiency. However, they also come with high management and maintenance costs and a complex system.

The goal of the new edge-cloud computing solution is to get beyond the drawbacks of the more conventional methods. Cloud computing offers robust storage and deep analytical capabilities, whereas edge computing permits preliminary data processing close to the site of data gathering, reducing the latency of data transmission. This plan can somewhat increase the system’s scalability and real-time data processing [10]. However, there are still difficulties in the process of integrating edge computing with cloud computing, including problems with data security, synchronization, and system complexity.

The contribution of this article is as follows:

- The cloud computing system integrates real-time or near real-time monitoring, which enhances the ability to track bridge deformation and response to external factors like wind loads and traffic.
- The system uses preprocessing to clean sensor data and align it with GPS time, followed by post-processing that produces statistical analyses every 10 minutes. This enables detailed tracking of environmental factors and bridge behavior, ensuring timely identification of structural issues.
- The early warning module, which updates baselines and thresholds iteratively, allows for the proactive identification of abnormal structural behavior, thus enhancing safety management.

The remaining sections of this article are structured as follows:

Related work is given in Section II. Section III presents the bridge diagnostic modeling. Section IV discusses the data processing and monitoring module. It explains the bridge structure evaluation and early warning system. It also describes the software development and sensor system. Section V provides the preliminary analysis of bridge deformation, including statistical evaluations from 2021 and 2022. It covers the initial examination of InSAR image data to monitor ground subsidence. Section VI concludes with the overall findings and implications for future work.

II. Related Work

Several researchers have explored SHM systems for bridges, focusing on real-time data collection, processing, and predictive analysis.

Early SHM systems have primarily relied on wired sensor networks and manual data collection, often requiring substantial human intervention for data processing and analysis. These systems also faced limitations in scalability, real-time monitoring, and the ability to handle large volumes of data. More recent studies have introduced cloud computing platforms for SHM systems, enhancing data storage, processing capabilities, and remote access to monitoring data [3, 9]. For instance, Xie et al. [11] mentions the use of GNSS and acceleration data for bridge vibration analysis, similar to the proposed approach. However, these systems often lack advanced real-time fault tolerance mechanisms and high computational efficiency when dealing with large sensor networks. Bayik has also applied InSAR image processing to detect settlement movements around large-scale infrastructure [12]. However, most studies have treated InSAR data as separate from the real-time SHM system, lacking integration into a unified monitoring platform. This limits their utility for ongoing structural assessment and real-time decision-making.

Existing systems often fail to maintain data integrity during network interruptions or connectivity failures, which can cause significant gaps in data during critical periods. While many studies use basic statistical methods for monitoring, there is limited use of intelligent algorithms that integrate GNSS, acceleration, and environmental data for real-time predictive analysis [13].

The research presented in this paper builds on the work of [14] and others by proposing a **comprehensive cloud-based SHM system** that integrates real-time data processing, intelligent predictive analysis, and fault tolerance mechanisms. Through a backup server capable of storing raw sensor data for up to one month, a feature that existing systems lack.

III. Bridge Diagnostic Modeling

The design of the model and the development of formulas are important components in the dynamic monitoring of bridge structures. A realistic and scientific model should be developed for assessing and forecasting the dynamic response of the bridge in order to achieve an accurate diagnosis of the bridge structure. Based on the integrated edge-cloud computing architecture, a bridge diagnostic model design was developed in this study [15, 16].

A finite element model (FEM), which considers the bridge's geometry, material properties, and boundary conditions, can be used to represent the dynamic response of a bridge structure. The dynamic response of the bridge was simulated using a linear elastic FEM.

The following equation of motion can be used to characterize a bridge's dynamic response:

$$M\ddot{u}(t)+C\dot{u}(t)+Ku(t)=F(t) \tag{1}$$

where, M is the quality matrix, representing the quality distribution of the bridge; C is the damping matrix, representing the damping characteristics of the bridge; K is the stiffness matrix, representing the stiffness characteristics of the bridge; $u(t)$ is a displacement vector, representing the dynamic response of the bridge; and $F(t)$ is an external load vector, representing wind load, traffic load, etc.

Modal analysis was applied to the bridge in order to examine its inherent frequencies and vibration modes. The characteristic equation of the modal analysis is as follows:

$$\left(K-\omega^2 M\right)\phi=0 \tag{2}$$

where, ω is the modal frequency, and ϕ is the modal shape. The bridge's inherent frequency and vibration mode can be determined by resolving the characteristic equation.

A. Evaluation of Structural Health

The structural health of bridges was assessed using the features extracted. The bridge's health index HI can be determined using the following formula:

$$HI=\frac{\bar{d}}{\sigma_b} \tag{3}$$

The presence of structural irregularities in the bridge can be established by comparing the health index of HI with a certain threshold. A warning will be sent out if HI surpasses the cutoff.

B. Wind Speed and Deformation Relationship

The following nonlinear regression model can be used to investigate how wind load affects bridge deformation:

$$\delta_{\text{lateral}}=\alpha\cdot V_{\text{wind}}^2+\beta\cdot V_{\text{wind}}+\gamma \tag{4}$$

where, δ_{lateral} is the amount of lateral deformation; V_{wind} is the wind speed; α, β, and γ are the regression coefficients yielded by the regression analysis.

IV. System Architecture for Cloud Computing

Fig. 1 depicts the general architecture of the Federal Reserve System (FRB) cloud computing system, which is broken down into five subsystems: the data management module, the data processing and monitoring module, the data collection and transmission module, the bridge structure evaluation and early warning module, and the sensor module. The components and interactions between the subsystems are depicted in Fig. 2.

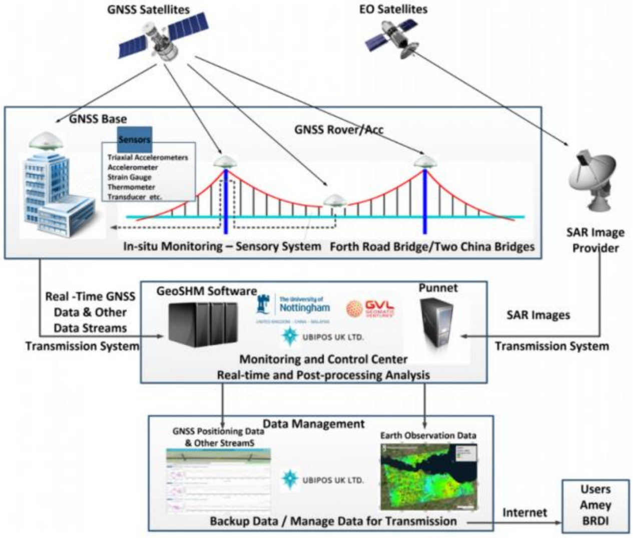


Fig. 1. The overall architecture of the cloud computing system.

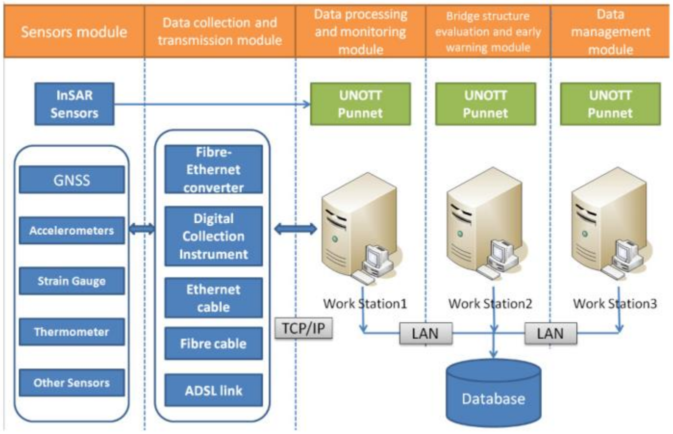


Fig. 2. Data flow and the relationships between the cloud computing subsystems.

C. Sensor Module

The cloud computing's sensor module is made up of various sensor types that can monitor the bridge's displacement, acceleration, inclination, stress, and other structural responses. It can also identify external loads applied to the bridge, such as wind loads and traffic weight, and short- and long-term environmental effects, such as temperature, weather, and ground motion [17, 18]. Fig. 3 illustrates the precise locations of the sensors on the FRB, and Table I lists the different kinds of sensors used in this cloud demonstration project, along with their sampling rates.

Global Navigation Satellite System (GNSS) technology, which makes use of both expensive and low-cost GNSS receivers, forms the basis of the sensor module. A fundamental prerequisite for the creation of an economical sensor module system is the provision of both static profiles and dynamic behavior for both low-cost and high-performance receivers. Three pairs of GNSS receivers were placed across the center and two navigation points in the major region of *B*, in addition to three inexpensive three-axis Sherborne accelerometers. This combination facilitates the integration of acceleration data and

GNSS data for extremely precise measurements of bridge deformation. An accelerometer was also positioned at 1/8 and 3/8 of the major spans, offering more information that could be used to determine the modulation frequency and vibration mode geometries of the FRBs. A triaxial accelerometer was installed atop each of the two major towers since deformation monitoring is critical to their operation. Inclineometers were also erected to show the average deformation at the summits of the towers. In-depth correlation studies of the wind loads on the FRB were made possible by the installation of three anemometers on the structure: two at the top of the two main towers and one at the mid-span. Advanced photovoltaic technology was also used in the cloud computer demonstration project to deliver data on sensors that can negatively impact the main tower foundations and the bridge's overall integrity.

TABLE I. SPECIFICATIONS OF THE SENSORS SET UP FOR THE GEOHAZARD MONITORING REMOTE INITIATIVE

Sensors	Details	Sampling Rates (hz)
GNSS	Leica gro	12
GNSS	Panda DB38	2
Anemometer	Gill windmaster	22
Weather station	Gill Metpak	2
Accelerometer	Sherborne A545-0003-2G	100
Inclinometer	Sherborne LSOP-1	12
InSAR image	EO 1	Image/14 days

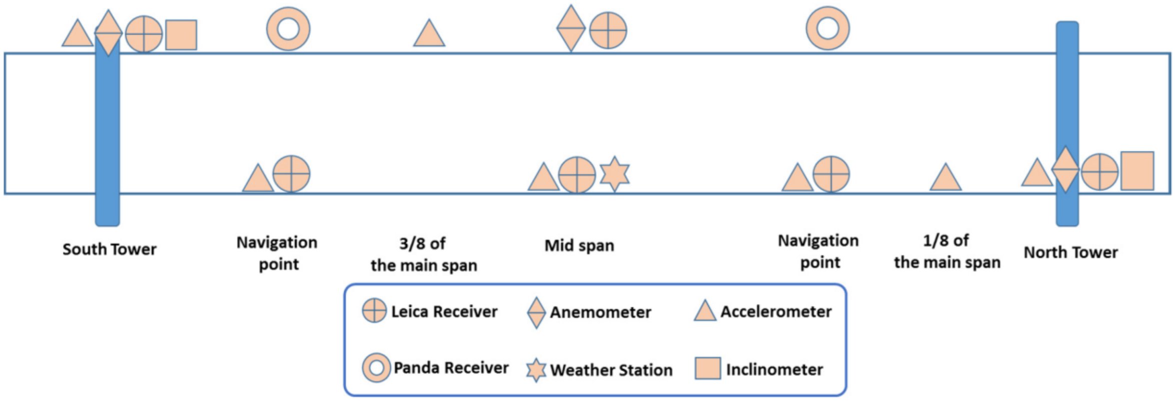


Fig. 3. Installation positions of sensors.

D. Data Collection and Transmission Module

The architecture of the data collection and transmission module installed at the FRB is shown in Fig. 4. The module's primary function is to convey data to a server safely housed in the field control centre using a fiber-optic link for communication with the sensors. Bridge operators can use this server to download data for additional analysis, generate reports on a regular basis, and access monitoring data for real-time monitoring. More crucially, in the event that communication between the bridge site and the main cloud server is lost, this server serves as a backup server, temporarily storing raw sensor data. The backup server is built to hold roughly one month's worth of raw sensor data, which is adequate in the case of a potential connectivity failure given the high sampling rate and numerous sensors.

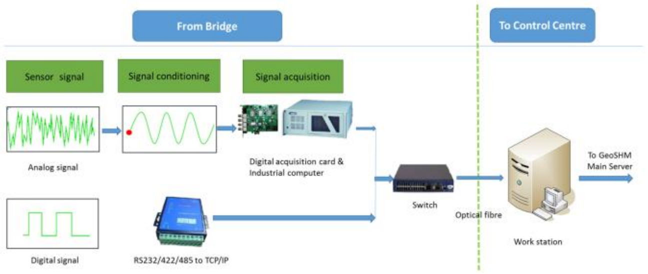


Fig. 4. Schematic diagram of the data collection and transmission module for cloud computing.

E. Data Processing and Monitoring Module

Preprocessing and post-processing units make up the two components of the data processing and monitoring module, which is primarily installed on the primary cloud computing server (Fig. 5). The primary duties of the preprocessing unit are to detect and eliminate anomalies and synchronize all sensor data with GPS time. The preprocessing unit also transforms the bridge deformation data in the bridge coordinate system from the purified GNSS data. The post-processing unit receives the output from the preprocessing unit [19].

The post-processing unit statistically assesses features pertaining to environmental impacts, external loads, and bridge deformation to produce statistical averages that are updated every 10 minutes. These characteristics include the average air temperature, the peak wind coefficient, the average inclination of the main tower, and the mean and standard deviation of the bridge deformation in the span. The cloud computing data strategy provides a precise definition for these low-level aspects.

The cloud computing system has an automatic and advanced system identification algorithm that uses GNSS and acceleration data to predict the modal frequencies and shapes of vibration patterns in order to perform intelligent data analysis. Furthermore, the device possesses real-time or almost real-time monitoring capabilities that leverage the previously mentioned attributes to track bridge deformation in response to wind loads and additional operational and environmental variables [20, 21].

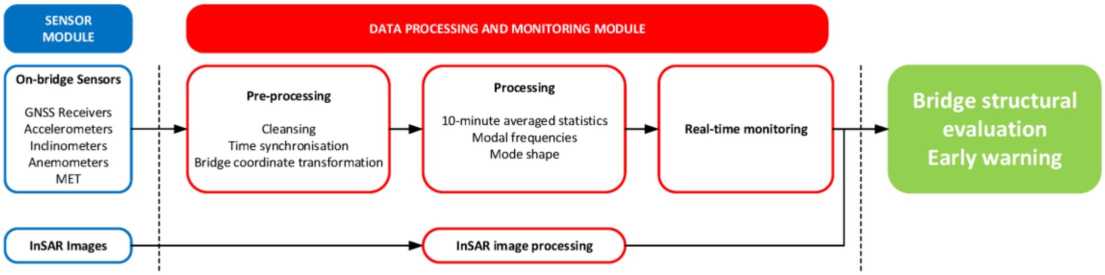


Fig. 5. Main features of the data processing and monitoring module for cloud computing.

It is noteworthy that, because of its low cost and high computational needs, Interferometric Synthetic Aperture Radar (InSAR) image processing is carried out on a monthly basis. This makes sense because, in contrast, settlement takes place over a longer time frame.

F. Bridge Structure Evaluation and Early Warning Module

The design and execution of an alert system based on the requirements of the cloud computing data policy are shown in Fig. 6. This module's performance depends on baselines and thresholds that were initially established using previous monitoring data and bridge operator expertise. These baselines and thresholds were then continuously refined using an iterative update mechanism to accurately reflect the bridge's current condition. When the measured bridge reaction goes over these limits and baselines, an alarm will sound, signaling that the bridge's structural behavior is aberrant.

A more sophisticated structural evaluation process can be utilized to investigate an alarm further after it has been set off. This helps identify whether the alarm is caused by alterations in the operating and environmental circumstances, modifications to the structural system, or the failure of a member. Additionally, the method uses modal parameters taken out of the deformation data to update the structural model in a rolling fashion. By facilitating simulation, the updated model helps bridge operators make better management decisions [22].

When the findings of InSAR image processing are obtained, they are applied to a structural model in order to evaluate how long-term ground motion affects structural stiffness. Bridge stability and long-term safety can be improved with the use of this procedure.

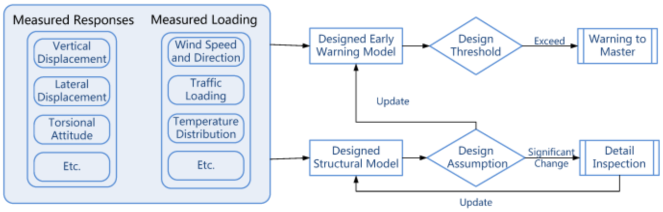


Fig. 6. Module flowchart for cloud computing.

V. Experimentation

A. Software Development and Sensor System Status

A cloud computing web application was developed in terms of software, as seen in Fig. 7. User engagement with the cloud computing system is facilitated by the platform. Some of this web application's functions are available to users, such as real-time monitoring, immediate alarms, and historical data queries. The following section goes into additional detail about a few of the outcomes this web application produces.

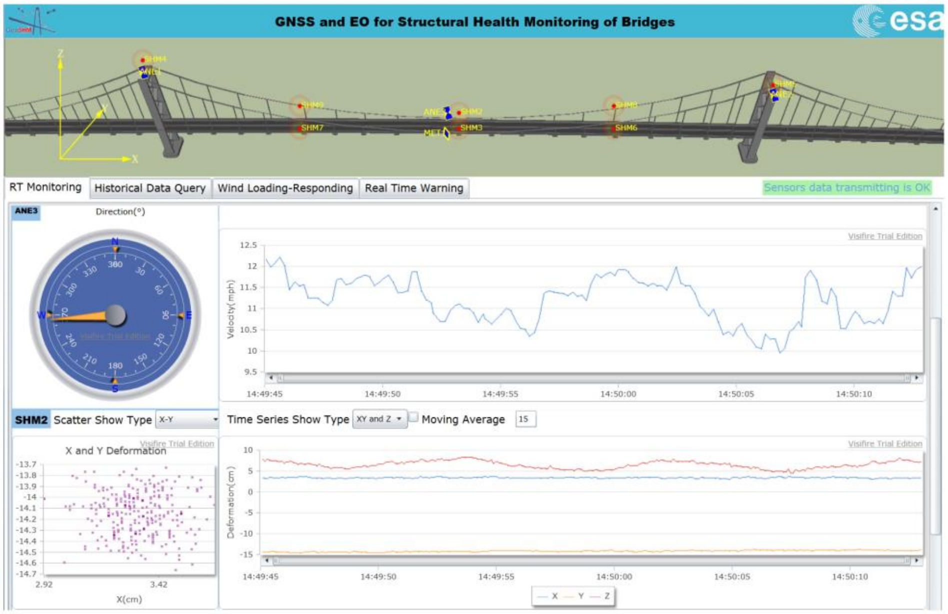
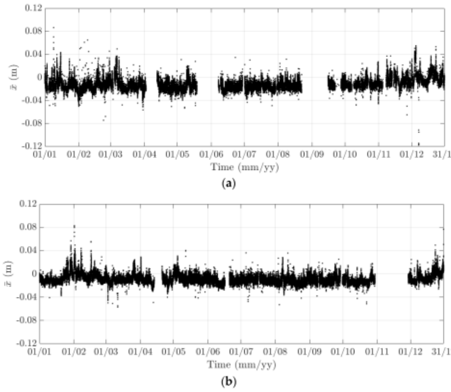


Fig. 7. Web applications for cloud computing.

B. Preliminary Analysis of Bridge Deformation

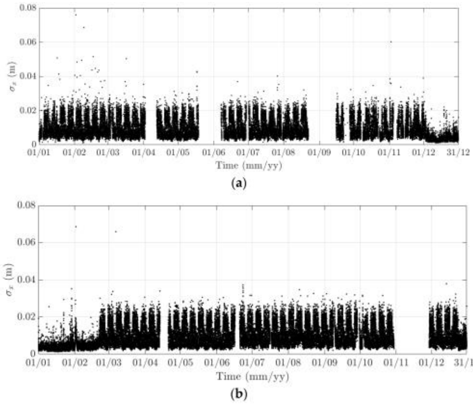
The cloud computing web application gives customers access to 10-minute average feature statistics from the cloud computing database in addition to real-time monitoring. With the help of this capability, the user may comprehend the bridge's response patterns and history as well as further examine how the bridge reacts to system changes or structural part failures. Furthermore, the 10-minute average statistics for temperature, wind speed, intrinsic frequency, and bridge response evaluation are crucial for establishing baselines, thresholds, and short- and long-term trends. These data are important for developing bridge structure evaluation and warning algorithms, as well as for evaluating the typical structural behavior of FRBs under various operational and environmental situations. This section presents and discusses a few of the evaluations' findings.

The cloud computing web application has the ability to automatically create several kinds of statistical graphs on demand. MATLAB was utilized for their high-resolution presentations. The main focus is on the 10-minute mean and standard deviation responses of the FRBs, along with their relationship to wind speed, air temperature, and traffic. The bridge's response involves deformation in four directions: vertical (along the *z*-axis), torsional (around the *x*-axis), transverse (along the *y*-axis), and longitudinal (along the *x*-axis). The bridge's steady state is represented by the 10-minute mean of its long-term deformation, and its dynamic reaction is shown by the 10-minute standard deviation.



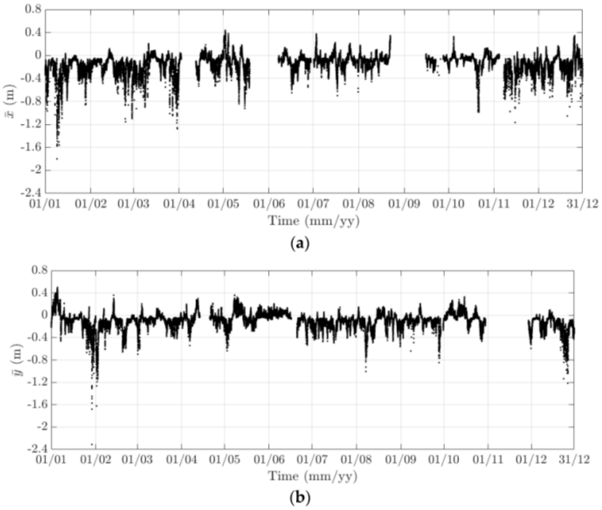
(a)
(b)

Fig. 8. The 10-minute mean changes in the longitudinal response in (a) 2021 and (b) 2022 throughout the FRB.



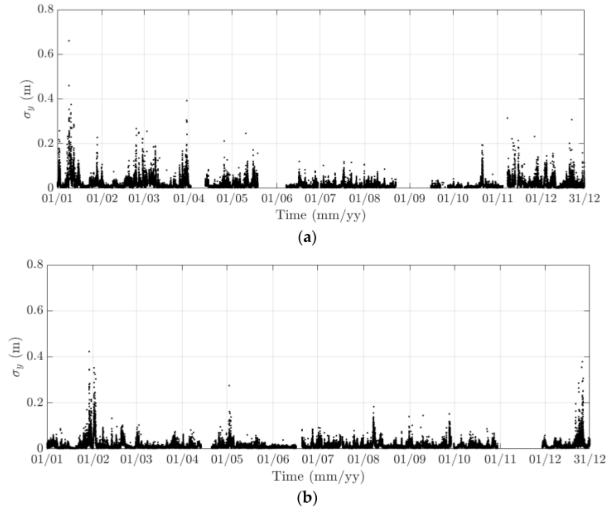
(a)
(b)

Fig. 9. The 10-minute standard deviation changes in longitudinal response across the FRB in (a) 2021 and (b) 2022.



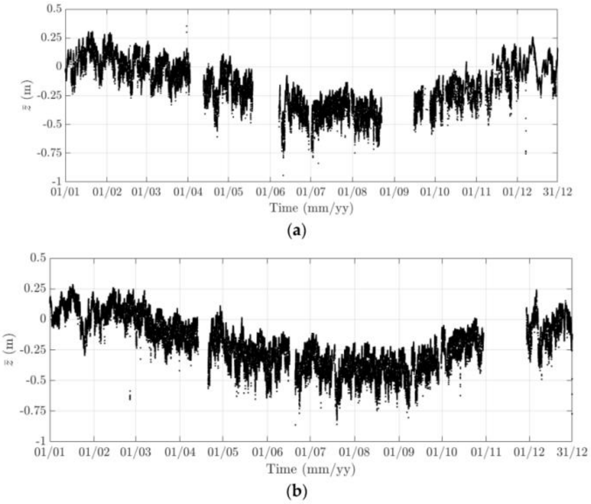
(a)
(b)

Fig. 10. The 10-minute mean changes in the lateral response across the FRB in (a) 2021 and (b) 2022.



(a)
(b)

Fig. 11. Variation of lateral response in FRB in span (10-minute standard deviation) for years 2021 and 2022.



(a)
(b)

Fig. 12. The average 10-minute heave response changed in 2021 and 2022 throughout the FRB.

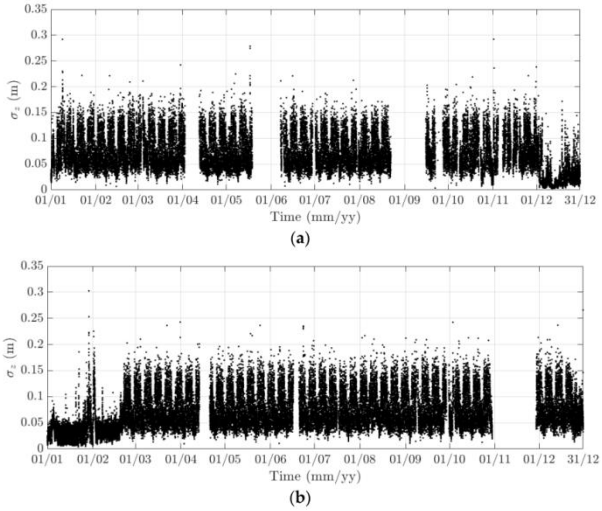


Fig. 13. Variations in the 10-minute standard deviation of the FRB oscillating respiration in (a) 2021 and (b) 2022.

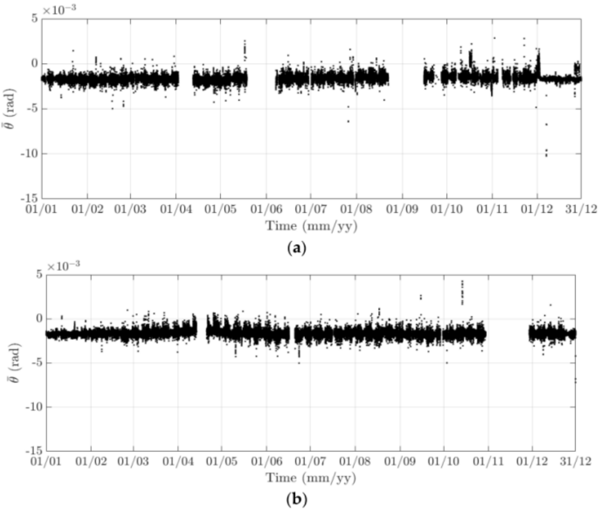


Fig. 14. Torsional response changes throughout a 10-minute period in the FRB in (a) 2021 and (b) 2022.

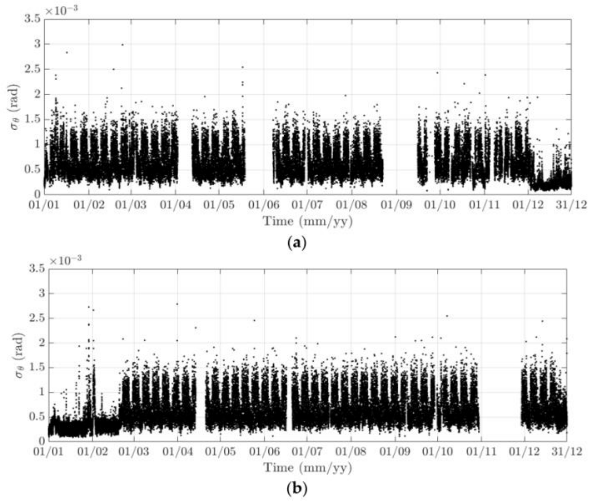


Fig. 15. Torsional response over the FRB in (a) 2021 and (b) 2022: 10-minute standard deviation change.

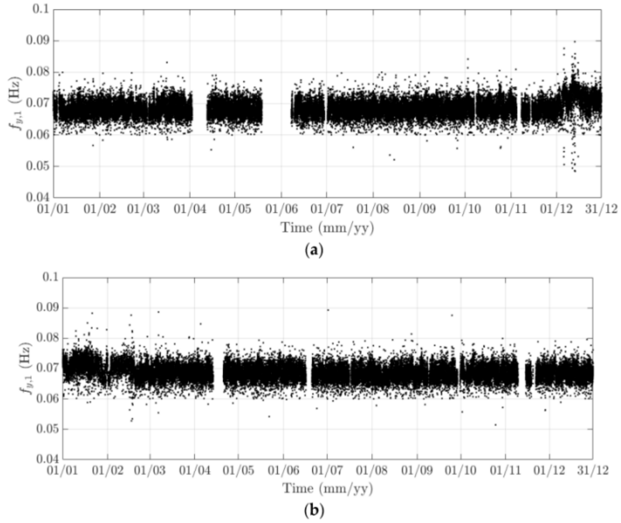


Fig. 16. Intrinsic frequency variations during a 10-minute period for the initial transverse model in (a) 2021 and (b) 2022.

Periodic characteristics of the bridge response and intrinsic frequency were identified through the analysis of the monitoring data for the years 2021 and 2022. These characteristics can be categorized into daily and weekly cycles. Certain data collected by the SHM system on the 560-meter Chinese bridge in Hong Kong, Zhuhai, and Macao bears a striking resemblance to some of these observations [23]. As further discussed in this section, the bridge response and the intrinsic frequency, however, frequently do not follow these patterns. Fig. 8 to Fig. 9 display the 10-minute averaged features for the years 2021 and 2022. Fig. 10 to Fig. 11 offer a thorough examination of a few chosen features for a brief period of time (August 1 to August 14, 2022).

A pattern of diurnal cycles is evident when examining the 10-minute standard deviations of the longitudinal response (Fig. 12), undulation response (Fig. 13), and torsion response (Fig. 14). These 10-minute average figures for the period August 1-14, 2022, demonstrate notable variations between day and night. The variance in traffic flow is closely related to the fact that the standard deviation values are substantially higher during the day than at night. The dynamic reaction is most intense when traffic peaks between 03:00 and 04:00, and it tapers off after 15:00 when traffic starts to decline.

Weekday traffic volume is higher than weekend traffic volume, which also results in a larger standard deviation fluctuation. Furthermore, the analysis of the 10-minute averages of the undulation deformation (Fig. 14), which represents the amount of sag in the mid-span of the FRB, can provide additional insight into the diurnal periodicity. Fig. 18 demonstrates a distinct sag volume difference between day and night; however, this cyclical pattern is less evident because of natural temperature swings. The lower weekend traffic results in a drop in the sag volume at the midspan. The 10-minute natural frequencies of the initial transverse and undulation patterns (Fig. 15, 16, and 17) also depict these diurnal cycles. These frequencies decreased by 7% and 2%, respectively, as a result of warmer daytime temperatures and more mass brought on by traffic; on weekends, this pattern was less noticeable.

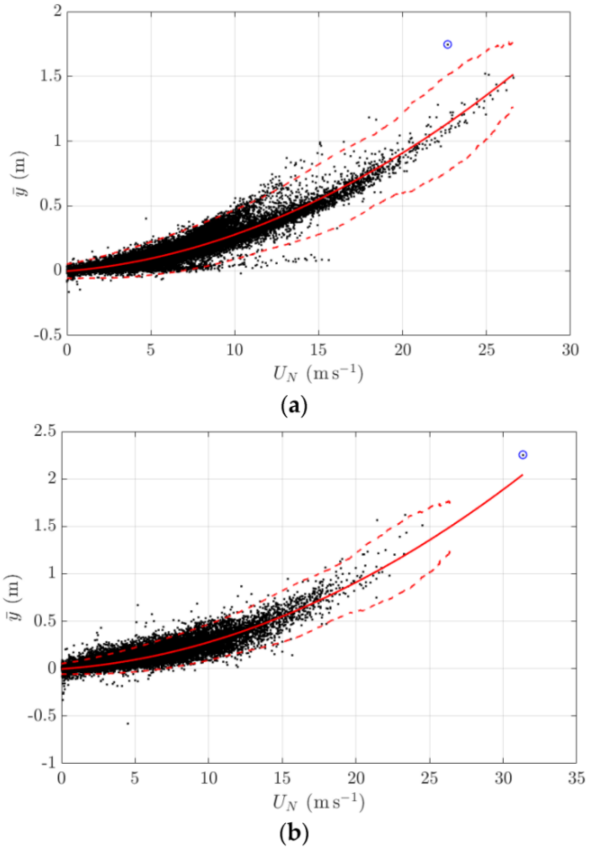
There are multiple instances in 2021 and 2022 with bridge response and intrinsic frequency deviating from the average trend: (i) early January 2021; (ii) December 2021 through February 2022; and (iii) late December 2022. During these periods, the intrinsic frequency rises dramatically, particularly during (ii), but the change in standard deviation within a day is negligible. It was discovered that the primary cause of the temperature and traffic fluctuations during these occurrences was the decreased volume of traffic on the bridges. To be more precise, (i) and (iii) are both public holidays (such as Christmas and New Year's), whereas (ii) is connected to the break in the northeast end-connection, which led to traffic restrictions and bridge closures. The 10-min averages of torsional and longitudinal responses were significantly altered by these events, but the 10-min averages of sag deformation were only slightly affected (Fig. 8 and Fig. 14).

Furthermore, an obvious annual cycle can be seen in the 10-minute average heave deformation (Fig. 14). The temperature increase caused the sag in the middle of the FRB span to progressively climb from January through August, eventually reaching a mean value of about 0.4 m. After August, as the outside temperature dropped, the FRB progressively moved back to its former location. There was no discernible annual cycle in the other bridge response or intrinsic frequency components. The data showed more random behavior with no discernible short- or long-term trends, as evidenced by the 10-minute averages and standard deviations of lateral deformations that were mostly impacted by wind speed (Fig. 12 and Fig. 13). On the other hand, as shown in Fig. 12, the analysis of the wind load response demonstrates a quadratic relationship between the mean lateral deformation and the positive component of the mean wind speed. The lower and upper thresholds, denoted as Un, are shown in Fig. 17, where circle is the standard deviation of the data samples for the positive component of mean wind speed within a window of three seconds. Though some data points considerably depart from the specified quadratic curve or exceed the upper threshold, most data points fall between these two criteria.

Upon examining the monitoring data from 2021 and 2022, certain distinctive characteristics of the FRB's structural response under external excitation were discovered. Certain recurrent patterns of bridge response and modal frequencies were affected by temperature. Based on their cause and duration, these patterns were divided into daily, weekly, and annual cycles. Certain departures from these cyclic patterns were found to be triggered by changes in operational conditions brought on by public holidays or FRB closures. Furthermore, in order to guarantee a normal structural reaction of the FRB in the event of severe wind, lower and upper bounds on the wind-induced response were established. It is critical to establish these characteristic behaviors of the FRB for subsequent years of monitoring data analysis, thereby helping identify systematic changes in the structure and their causes, which is an important part of the development of the cloud data strategy.

C. Initial Examination of InSAR Pictures

In this part, some initial findings from the InSAR image processing encompassing the Sanqi Bridge (Shanghai, China) and the surrounding area of the FRB are presented. As depicted in Fig. 18, the movement of subsidence in the vicinity of the FRB is minimal, mostly within a radius of approximately 2 km from the FRB, where its influence is minimal. In certain regions, the movement is at a rate of around 5 mm/year. On the other hand, Fig. 19 for the Sanqi Bridge (Shanghai, China) demonstrates that there are notable settlement movements occurring in the vicinity of the bridge, up to a maximum of 20 mm annually, in the area 1 km away. These settlement patterns move in the direction of the bridge, endangering its structural stability.



(a)

(b)

Fig. 17. Comparison of the produced quadratic curves and thresholds in (a) 2021 and (b) 2022, with blue circles representing extreme events.

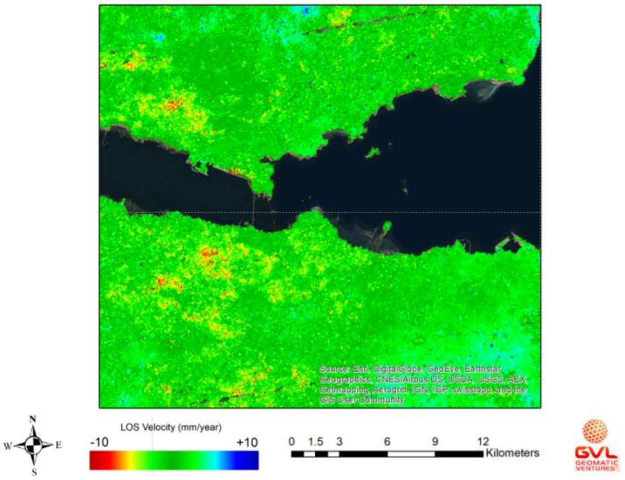

Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community
N
W
E
S
LOS Velocity (mm/year)
-10
+10
0
1.5
3
6
9
12
Kilometers
GVL
GEOMATIC VENTURES

Fig. 18. InSAR image processing of the Sanqi Bridge (October 2017).

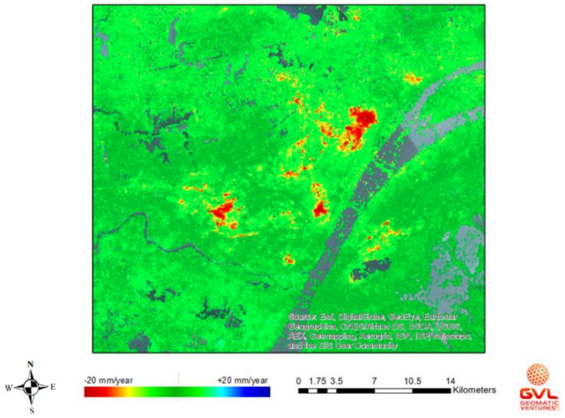

Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community
N
W
E
S
-20 mm/year
+20 mm/year
0 1.75 3.5 7 10.5 14
Kilometers
GVL
GEOMATIC VENTURES

Fig. 19. InSAR image processing of the surrounding area of the Sanqi Bridge (December 2023).

VI. Conclusion

This paper presents a cloud-based SHM system for long-span bridges, integrating GNSS, accelerometers, and InSAR technologies for real-time data collection and analysis. The system ensures continuous monitoring through a fiber-optic communication link and a backup server, while the cloud processing module refines data accuracy and evaluates structural responses to environmental factors. We observed cyclical patterns in bridge behavior influenced by traffic and temperature, and detected anomalies linked to specific events. While the system shows promise in providing real-time insights, limitations include the infrequent processing of InSAR data and challenges in scaling to larger infrastructures. Future work will address these challenges and further optimize the system for broader applications in infrastructure monitoring.

Building on our paper, future work could focus on several key areas to enhance and expand the capabilities of the proposed SHM system: Address the infrequent processing of InSAR images by developing more efficient algorithms or increasing processing frequency. This will provide more timely insights into long-term ground movement and its impact on structural health. Explore the incorporation of other sensor types, such as acoustic emission sensors or fiber optic sensors, to capture a broader range of structural responses and potential failure mechanisms. Improve the data processing and analysis algorithms to better handle large volumes of data and detect subtle anomalies. This could involve advanced machine learning techniques or AI-based predictive models.

References

[1] Al-Ali AR, Beheiry S, Alnabulsi A, Obaid S, Mansoor N, Odeh N, Mostafa A. An IoT-based road bridge health monitoring and warning system. Sensors, 2024;24(2):469.

[2] Wang T, Liang Y, Shen X, Zheng X, Mahmood A, Sheng QZ. Edge computing and sensor-cloud: Overview, solutions, and directions. ACM Computing Surveys, 2023, 55(13s):1-37.

[3] Peng Z, Li J, Hao H. Development and experimental verification of an IoT sensing system for drive-by bridge health monitoring. Engineering Structures, 2023;293:116705.

[4] Yuan J, Xiao H, Shen Z, Zhang T, Jin J. ELECT: Energy-efficient intelligent edge–cloud collaboration for remote IoT services. Future Generation Computer Systems, 2023;147:179-194.

[5] Kyriou A, Mpelogianni V, Nikolakopoulos K, Groumpos PP. Review of Remote Sensing Approaches and Soft Computing for Infrastructure Monitoring. Geomatics, 2023;3(3):367-392.

[6] Tan Y, Yi W, Chen P, Zou Y. An adaptive crack inspection method for building surface based on BIM, UAV and edge computing. Automation in Construction, 2024;157:105161.

[7] Su W, Xu G, He Z, Machica IK., Quimno V, Du Y, Kong Y. Cloud-edge computing-based ICICOS framework for industrial automation and artificial intelligence: a survey. Journal of Circuits, Systems and Computers, 2023:32(10):2350168.

[8] Chen Q, Cao J, Zhu S. Data-driven monitoring and predictive maintenance for engineering structures: Technologies, implementation challenges, and future directions. IEEE Internet of Things Journal, 2023;10(16):14527-14551.

[9] Liang W, Xiao J, Chen Y, Yang C, Xie K, Li KC, Di Martino B. TMHD: Twin-bridge scheduling of multi-heterogeneous dependent tasks for edge computing. Future Generation Computer Systems, 2024;158:60-72.

[10] Costin A, Adibfar A, Bridge J. Digital twin framework for bridge structural health monitoring utilizing existing technologies: New paradigm for enhanced management, operation, and maintenance. Transportation Research Record, 2024;2678(6):1095-1106.

[11] Xie Y, Zhang S, Meng X, Nguyen DT, Ye G, Li H. An innovative sensor integrated with GNSS and accelerometer for bridge health monitoring. Remote Sens. 2024;16:607. https://doi.org/10.3390/rs16040607

[12] Bayik C, Abdikan S, Ozdemir A, Arıkan M, Sanli FB, Dogan U. Investigation of the landslides in Beylikdüzü-Esenyurt Districts of Istanbul from InSAR and GNSS observations. Nat Hazards. 2021;109:1201–1220. https://doi.org/10.1007/s11069-021-04875-7

[13] Zhuang C, Zhao H, Hu S, Sun C, Feng W. Integrity Monitoring Algorithm for GNSS-based Cooperative Positioning Applications. Proceedings of the 32nd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2019). Miami, Florida, September 2019; pp. 2008-2022. https://doi.org/10.33012/2019.16881

[14] Mishra M, Lourenço PB, Ramana GV. Structural health monitoring of civil engineering structures by using the internet of things: A review. J Build Eng. 2022;48:103954. https://doi.org/10.1016/j.jobe.2021.103954

[15] Zhang C, Zhou G, Li J, Chang F, Ding K, Ma D. A multi-access edge computing enabled framework for the construction of a knowledge-sharing intelligent machine tool swarm in Industry 4.0. Journal of Manufacturing Systems, 2023;66:56-70.

[16] Raeisi-Varzaneh M, Dakkak O, Habbal A, Kim BS. Resource scheduling in edge computing: Architecture, taxonomy, open issues and future research directions. IEEE Access, 2023, 11: 25329-25350.

[17] Negi P, Singh R, Gehlot A, Kathuria S, Thakur AK, Gupta LR, Abbas M. Specific soft computing strategies for the digitalization of infrastructure and its sustainability: A comprehensive analysis. Archives of Computational Methods in Engineering, 2024;31(3):1341-1362.

[18] Sadhu A, Peplinski JE, Mohammadkhorasani A, Moreu F. A review of data management and visualization techniques for structural health monitoring using BIM and virtual or augmented reality. Journal of Structural Engineering, 2023;149(1):03122006.

[19] Kumar R, Sangwan KS, Herrmann C, Thakur S. A cyber physical production system framework for online monitoring, visualization and control by using cloud, fog, and edge computing technologies. International Journal of Computer Integrated Manufacturing, 2023;36(10):1507-1525.

[20] Guo Z, Yu K, Kumar N, Wei W, Mumtaz S, Guizani M. Deep-distributed-learning-based POI recommendation under mobile-edge networks. IEEE Internet of Things Journal, 2022;10(1):303-317.

[21] Zhang C, Roh BH, Shan G. (2023). Poster: Dynamic clustered federated framework for multi-domain network anomaly detection. In Companion of the 19th International Conference on emerging Networking EXperiments and Technologies NY, USA, 2023;pp.71-72.

[22] Gong T, Zhu L, Yu FR, Tang T. Edge intelligence in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 2023;24(9):8919-8944.

[23] Dai Z, Zhang Q, Zhao L, Zhu X, Zhou D. Cloud-Edge computing technology-based internet of things system for smart classroom environment. International Journal of Emerging Technologies in Learning, 2023;18(8):79-96.

Optimized Fertilizer Dispensing for Sustainable Agriculture Through Secured IoT-Blockchain Framework

IoT-Blockchain Framework for Sustainable Agriculture

B. C. Preethi[1]*, G. Sugitha[2], T. B. Sivakumar[3]

Department of Electronics and Communication Engineering, St.Xavier's Catholic College of Engineering, Nagercoil, India[1]
Department of Computer Science and Engineering, Muthayammal Engineering College (Autonomous), Rasipuram, India[2]
Department of Computer Science and Engineering-Vel Tech Rangarajan,
Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India[3]

Abstract—**Precision farming is essential for optimizing resource use and improving crop yields to attain sustainable agriculture. However, challenges like data insecurity, fertilizer costs, and inadequate consideration of soil health pose a hindrance to achieving these goals. To overcome these issues, the proposed work presents a novel approach for optimizing fertilizer dispensing by developing a framework connecting IoT and blockchain with a community of greenhouses. The system consists of IoT sensors installed inside the greenhouses to measure soil pH and nutrient values. This collected sensor data is compressed and stored securely and in an off-chain manner by the IPFS (Inter-Planetary File System) hash using the Keccak-256. MetaMask transfers the data for blockchain registration and authentication. The data is then preprocessed using Z-score normalization, Label Encoding, and One-Hot Encoding to obtain a precise analysis. A Deep Learning-based Convolutional Neural Network (DL-CNN) is used to classify soil conditions and determine the appropriate fertilizer requirements. The results of the DL-CNN model are viewed in a dashboard through a Decentralized Application (D-App) that we developed to provide real-time information to consumers, field analysts, and agricultural organizations. Field analysts use the information to establish a control center for precisely applying fertilizers. The proposed method achieves a classification accuracy rate of 98.86%, thus increasing soil health and providing a solution for effectively managing fertilizers.**

Keywords—***Fertilizer dispensing; IoT sensors; blockchain; deep learning; convolutional neural network; greenhouse management; and decentralized application***

I. Introduction

Smart agriculture, also known as smart farming, focuses on using advanced technologies and data-driven operations to improve sustainability in agricultural production [1]. In this area, IoT-enabled sensors are used to monitor some parameters in real-time and provide cutting-edge management of farming data [2]. The use of blockchain technology in the field of smart agriculture guarantees security, transparency, and tamper-proof data storage, as well as the maintenance of trust and provenance along the agricultural chain [3]. Moreover, blockchain in smart agriculture securely transfers and stores data to decide soil health and fertilizer applications [4, 5]. This method optimizes resources, increases crop yields, minimizes environmental impact, and enhances food safety by offering verifiable data to consumers [6]. Blockchain and IoT agriculture improve decision-making and the ability to implement decentralized solutions that help farmers and provide better outcomes [7].

Previous research in smart agriculture has mostly addressed the application of IoT sensors that monitor environmental conditions to improve crop management [8, 9]. The techniques of machine learning (ML), deep learning (DL), and cloud computing have been applied for processing and analyzing the data collected from these sensors [10, 11]. Yet, several of these methods still depend on centralized systems for data storage and processing, which introduces problems regarding data security, flexibility, cost, and accessibility [12, 13]. Although some studies have explored the use of blockchain to secure data transfers, they mostly do not implement DL models for the real-time decision-making process based on soil health parameters [14, 15]. This work addresses these gaps by integrating IoT, blockchain, and DL models into the timely fertilizer distribution and constructing a secure system for monitoring greenhouse conditions.

The main contributions of this work are as follows:

- A novel way of optimizing fertilizer dispensing in the greenhouse environment using IoT sensors, blockchain technology, and DL models to monitor soil health is proposed. It provides accurate soil monitoring and secure data collection and storage to improve fertilizer dispensing accuracy.
- The proposed method uses the Lempel Ziv Welch (LZW) compression method for efficient data storage and transmission. In addition, the use of the SHA-3 (Keccak) hashing algorithm with a chaotic key for data encryption contributes to the improvement of security and makes the system robust against cryptographic attacks.
- This work introduces a DL model defined as a DL-CNN used to predict the soil pH values and NPK levels. Hence, the impact of the application of fertilizers can be determined.

- We have developed a Decentralized Application (D-App) dashboard that provides real-time data regarding soil health and greenhouse parameters. The application allows farmers and other stakeholders to view the key information, make decisions about fertilizer application, and consequently care for the best growing conditions.

The structure of this paper is organized as follows: Section II provides a literature review covering advancements in smart agriculture and the application of blockchain and DL. Section III details the proposed methodology for optimizing fertilizer dispensing using a blockchain method for secure smart greenhouse management. Section IV presents the results and analysis. Finally, Section V concludes the paper.

II. Literature Review

This section examines and explores different modern techniques, technologies, and weaknesses in smart agriculture to improve soil management, crop yield, and overall farm efficiency.

A. Emerging Techniques in Sustainable Agriculture

Wei et al. [16] demonstrated the utility of Capacitive Coupled Contactless Conductivity Detection (C4D) in soil nutrient detection. This method encompasses C4D data and cluster analysis to improve nutrient monitoring and management. Despite that, a very low level of nutrients can limit the method's performance, and the system's efficiency may vary due to different soils.

Thorat et al. [17] proposed the Transition Probability Function (TPF) and Convolutional Neural Network (CNN) for precision farming. In this work, a system identifies pests, prescribes insecticides, and examines the soil nutrients to indicate the most suitable fertilizers. However, the model lacks the integration of sensors like pH, temperature, humidity, and soil moisture sensors to capture a range of environmental data.

Senapaty et al. [18] suggested an IoT-enabled soil nutrient classification and crop recommendation (IoTSNA-CR) system. This model encompassed IoT sensors, cloud storage, and a multi-class support vector machine (MSVM) to classify soil nutrient levels. However, drawbacks include the need for significant investment in sensors and the need to clean data, which is a big challenge.

Pechlivani et al. [19] have coined soil and environmental monitoring using an IoT-enabled device with sensors to collect data on key soil and environmental parameters. The summation of sensors paired with an ESP32 microcontroller, data visualization, an Android app, and 3D printing for control housing. However, the study has inaccuracies in the sensor measurements and more validation in the different agricultural fields.

Indira et al. [20] proposed using Artificial Intelligence (AI), machine learning (ML), and Long-Range (LoRa) technology to transform farming operations. They analyzed these data using AI to provide actionable insights for improving agricultural practices. Nevertheless, challenges stem from issues such as data security and the difficulty of deploying AI in remote areas with limited internet connectivity.

Vincent et al. [21] introduced IoT and Al in agriculture to evaluate land suitability for cultivation. This technique is used for sensor networks to collect data on soil and environmental variables. It also uses the Multi-Layer Perceptron (MLP) neural networks to classify data into four suitability classes. This study has limitations due to incomplete sensor data and overfitting in the MLP model due to its complexity.

Singh et al. [22] focused on developing a portable device for soil nutrient analysis. This method detects soil nitrogen (N) and phosphorus (P) levels to optimize crop yield and minimize fertilizer use integrated with IoT and LED sensors. Yet, its limitations include potential inaccuracies at higher nutrient concentrations and the need for further refinement to improve sensitivity.

Morchid et al. [23] introduced the applications of IoT and sensor technology in agriculture and food security. The paper also enumerates IoT's advantages in agriculture, including efficiency and resource reduction. Nevertheless, it still cannot offer any real-world examples and only speculations about solutions to the challenges.

TABLE I. Analysis and Applications of Blockchain in Greenhouse Agriculture

References	Technique Used	Objectives	Results	Limitation
[26]	Blockchain and IoT	Enhance transparency and traceability in the agricultural process	Improved trust and reliability in food certification	High initial setup cost and complexity
[27]	DL-based Image Processing	Measurement and monitoring of greenhouse environmental parameters	98% success rate in image processing	Limited dataset size for training
[28]	ML (Random Forest)	Classify greenhouse gas (GHG) emissions during groundnut harvesting	Accuracy: 86.46%	Limited by data variability and missing values
[29]	Blockchain and IoT	Ensure anonymity and security in e-voting	High voter privacy and verifiability	Requires computational resources
[30]	RNN and · Blockchain	Secure agricultural data in IoT network	Accuracy: 97.7%,	Design complexity
[31]	Blockchain- consensus algorithm	Ensure a secure and transparent supply chain in smart agriculture	Reduced fraud and counterfeiting	High computational cost
[32]	Blockchain and DL	Ensure food safety and transparency in the supply chain	RMSE values of 872.56	Increased implementation costs
[33]	ML: Multi-regression analysis	Identify the next hop for agricultural data transferring	Improved energy usage efficiency by 13% and 16%	High packet congestion

Nayagam et al. [24] addressed the issue of controlling disease in smart agriculture with IoT technology. It suggested using multiple GPUs in a Parallel and Distributed Simulation Framework (PDSF) with IoT to oversee crop surveillance and pest management. However, multiple GPUs could reduce system performance, and more efficient data processing algorithms are also needed.

Rajak et al. [25] suggested the integration of IoT and smart sensors in agriculture for crop production and minimizing economic losses. Techniques like AI algorithms and ML are used for real-time data processing and precision farming. Yet, high implementation costs and data security concerns impact farmers by making the technology less accessible.

B. Blockchain Integration in Sustainable Agriculture

The literature shows advanced agricultural techniques for improving agricultural practices with several limitations (as shown in Table I). This work proposes an integrated IoT, blockchain, and DL approach to solve this issue to improve soil management and optimize greenhouse agriculture.

III. Proposed Methodology for Optimal and Appropriate Fertilizer Dispensing

The proposed system optimizes fertilizer dispensing by accurately assessing soil conditions. This model employs IoT sensors at different greenhouses to collectively monitor soil nutrition and pH values. The collected data from these sensors is compressed and assigned to the IPFS hash code for secure transfer to the blockchain, specifically in an off-chain setup. This data is transferred to the blockchain using MetaMask for registration and login authentication. Once transferred to the blockchain, the data is input into a Convolutional Neural Network (DL-CNN) model to classify pH value and NPK (Nitrogen, Phosphorous, Potassium) amount in the soil to apply the required fertilizer. The DL model's output can be accessed through a dashboard in a Decentralized Application, thereby providing insights to consumers, field analysts, and agricultural organizations. The Field analysts can use this data for activating the control center which further implements the necessary solutions within the greenhouse environment to provide necessary solutions for using appropriate fertilizer for optimal growing conditions. Fig. 1 explains the overall workflow of this system. It integrates IoT sensors, blockchain technology, and DL models to accurately assess soil conditions and optimize fertilizer application in greenhouse environments.

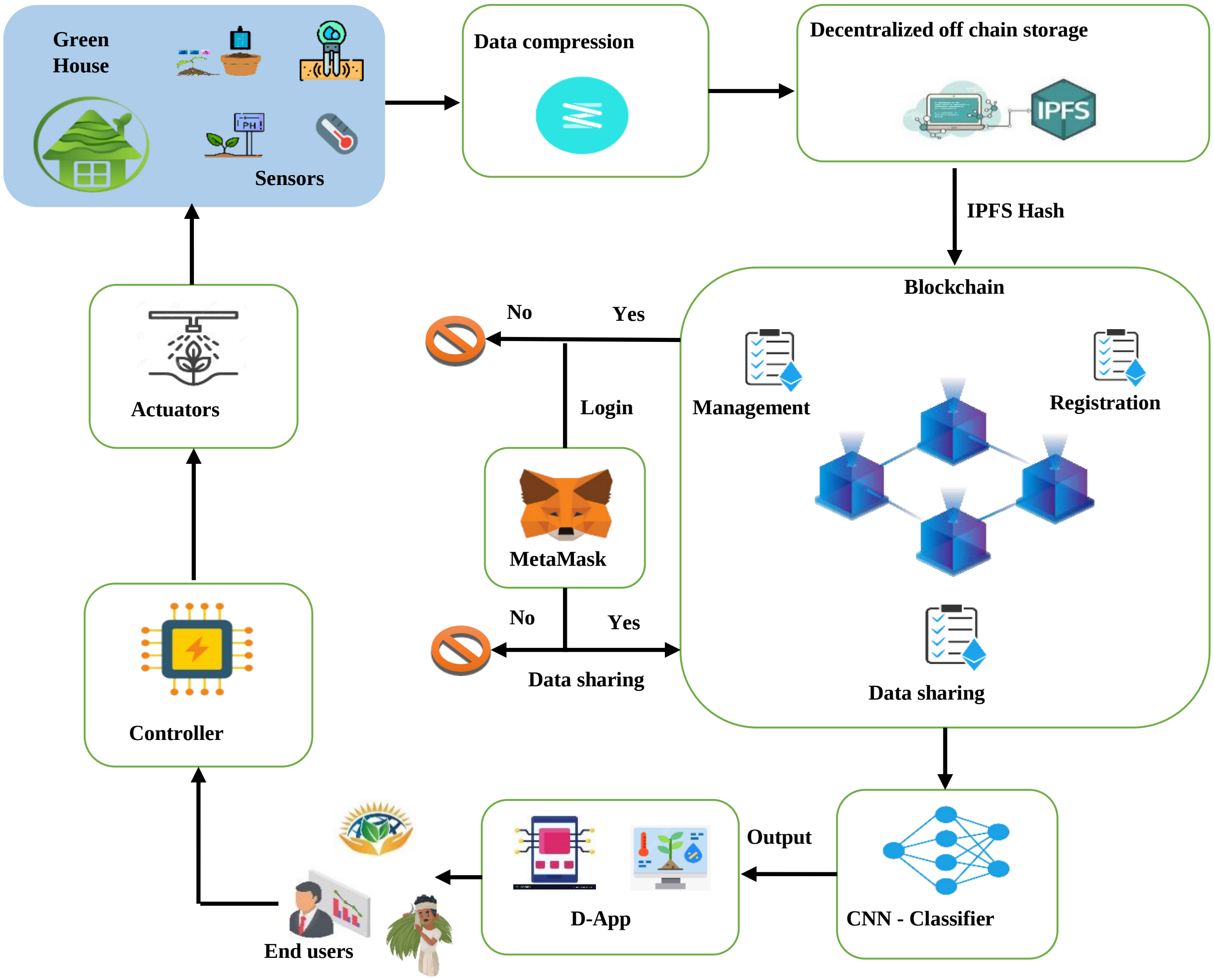


Fig. 1. Workflow of the proposed system.

A. Data Collection

In this framework, IoT sensors, including pH, moisture, temperature, and soil nutrients, are placed across different greenhouses to continuously monitor and collect data on soil conditions. These sensors capture real-time data that reflects the dynamic conditions of the greenhouse environment to provide a dataset for assessing soil health and making informed decisions on resource management.

B. Data Processing

The collected sensor data is compressed using the Lempel-Ziv-Welch (LZW) technique [34] to store large data volumes efficiently. LZW compresses data by replacing repeated sequences of characters with shorter codes. It takes a dictionary filled with all single-character strings. Meanwhile, it checks each sequence against this dictionary. If the sequence is found, it is expanded with the next character and continues. Otherwise, the sequence is added to the dictionary, and the code for the previous sequence is output.

C. Blockchain Configuration

1) Data Collection and IPFS Storage: After the compressed data is collected, it is securely assigned an IPFS (Inter-Planetary File System) hash code for securely storing off-chain. IPFS enables distributed storage, in which data is divided up into small parts, and each is signed with a cryptographic hash using the Keccak-256 algorithm [35]. Such chunks are then shared among several IPFS networks.

2) MetaMask with blockchain integration: Here, MetaMask provides secure data to be transferred to the blockchain through registration and log-in authentication. It makes use of Keccak hashing. If data chunks are stored in IPFS, their hash codes are recorded on the blockchain. Hence, it provides a secure reference to the off-chain data, whereas the blockchain storage requirements are minimized. To improve security, the SHA-3 (Keccak-256) hashing algorithm is used for the encryption process and is combined with the chaotic key for encryption and decryption. A chaotic key, generated with the help of logistic maps, is integrated with the SHA-3 hash to create a secure and unpredictable hash resistant to cryptanalysis attacks. Moreover, this algorithm is used to encrypt IoT devices with low computing resources, which are decrypted upon reaching the blockchain and fed into the DL-CNN model.

Fig. 2 provides the data flow and security mechanisms in this setup. The encrypted data is stored off-chain using IPFS and securely transferred through MetaMask. The data undergoes blockchain validation to improve security for real-time decision-making.

D. Data Classification

1) Data preprocessing: Before the classification step, the data goes through initial processing steps to ensure the optimal format for analysis. Eventually, the Z-score normalization method [36] eliminates the effect of outliers by standardizing the data to have a mean of zero and a standard deviation of one to reduce the influence of extreme values. After that, a Label Encoder [37] is applied to change the strings of the categorical values into integer numbers, thereby enabling numerical analysis. Following this, the One-Hot Encoder transforms these categorical integers into binary format to create a sparse matrix where each category is represented as a separate binary column.

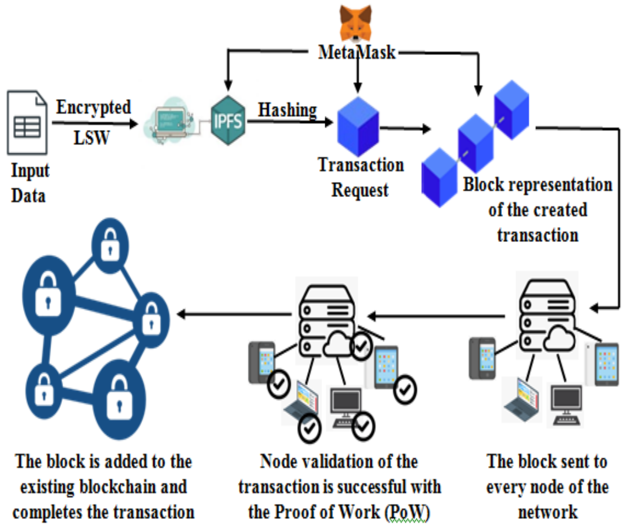


Fig. 2. Blockchain-based data transfer workflow.

2) Classification: The binary numerical data processed through preprocessing is loaded into the DL-CNN [37] model, which predicts the pH (soil acidity or alkalinity) and NPK content (determining the nutrition value of the soil). The CNN architecture features several levels structured for capturing and classifying important data. It starts with a 55*55*3 input layer which are the dimensions of the feed data. This is followed by multiple convolutional layers with different sizes of filters as (4*4) or (3*3). It identifies various features caught in the data. The batch normalization layer is placed after each convolutional layer to improve the stability of learning, and the activation layer employs the P-ReLU function to inject non-linearity. Pooling layers are inserted to reduce the spatial dimensions, decrease the computational cost, and improve the quality of feature extraction. As the network progresses through these layers and captures more intricate patterns in the data. A dropout layer is implemented to avoid overfitting wherein neurons are randomly disrupted with a fraction of them deactivated during training. Finally, the model contains a fully connected layer and a regression layer using Error Regression that classifies the final labeling (soil pH and nutrient content) to recommend the optimal fertilizers.

3) Visualization: The outputs from the DL-CNN model are visualized through the dashboard within a Decentralized Application (D-App). This dashboard offers a user interface for the customers, the field researchers, and the agricultural organizations to access real-time information on soil health, such as pH values and NPK content. By displaying these results in an accessible format, stakeholders understand the soil's condition and make informed decisions regarding fertilizer application and plan crop management.

4) Control center activation: Based on the dashboard's analysis results, field analysts can activate the control center to require modifications within the greenhouse environment. This can involve applying the appropriate amount of fertilizer to improve the quality of the soil and crop growth so that plants receive the nutrients they need for optimal development.

IV. Results and Discussion

The proposed system was implemented on a blockchain platform integrated with Python, utilizing smart contracts and secured data for accurate data handling. The results show the system can accurately classify soil pH and NPK levels, thus enabling precise fertilizer application while maintaining data integrity and security through blockchain technology.

A. Dataset Description

This dataset consists of one million soil samples that have been simulated from different places across the world. Each sample comes with information about soil texture, pH, organic matter content, moisture content, bulk density, nutrient levels (N, P, K), cation exchange capacity, electrical conductivity, color, porosity, and water holding capacity. This dataset, which has been developed for environmental scientists, agronomists, and data scientists, is excellent for research, ML, DL models, and educational purposes.

Dataset link: Global Soil Characteristics Dataset (1 Million) (kaggle.com).

B. Performance Analysis

Table II shows that compression techniques are effective in storing and transferring large volumes of sensor data to the blockchain in a compressed format. LZW ranks a balanced performance, compared to Huffman and Arithmetic. In its case, the LZW may show the best harmony with the compression efficiency and the speed of processing. Thus, it is a perfect choice for real-time applications, where a quick response time and a small storage footprint are crucial.

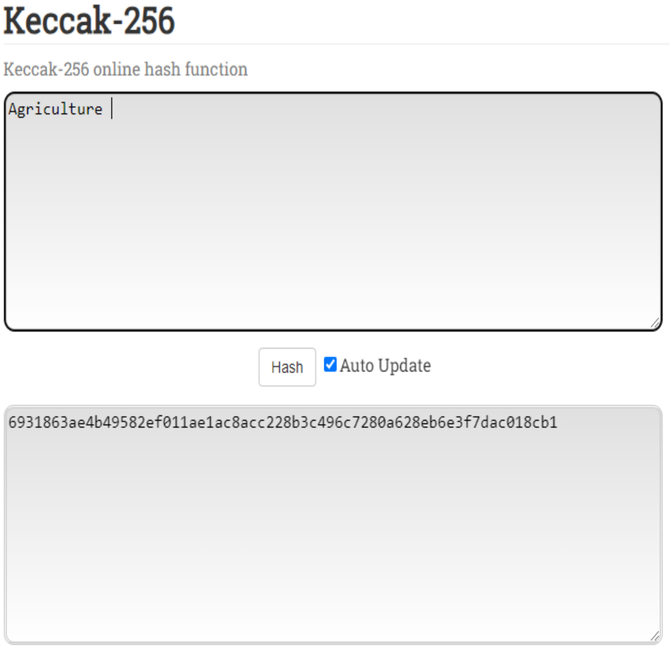
TABLE II. Comparative Analysis of Data Compression Techniques

Technique	Compression Ratio	Compression Time (s)	Compressed Size (bytes)	Memory Usage (bytes)
LZW	**2.108725**	**17.37507**	**1.31E+08**	**1.31E+08**
Huffman	2.398807	21.33783	1.16E+08	1.16E+08
Arithmetic	2.279311	226.2766	1.22E+08	1.22E+08

Fig. 3 demonstrates the Keccak-256 hash tool interface used in the system. It includes a text field with "Agriculture" which is the default text to be hashed. Below the input fields, "Hash" and "Auto Update" give the users the ability to generate and update the hash of the corresponding text. The resulting hash depicted as an alphanumeric string in the output field proves the capability of the Keccak-256 hashing function in securely changing and verifying data.

Fig. 4 shows a transaction detail from a blockchain explorer. It shows the transaction hash, block number, block hash, contract address, sender and recipient addresses (partially blocked for privacy) the gas used, and input data details. This detailed view checks and observes the transactions carried on the blockchain to ensure security in the management of digital assets within the system.

Fig. 5 illustrates a command line interface that exhibits a program execution output with data storage and verification. The screen displays, "Files are stored in local server" and "Data is already stored for this file name," then "wait - compiling/applying successfully in 65ms (51 modules)" and "true," which means that the "Hash is Stored in a Smart Contract". This output expresses the process of collecting and whether the data is correctly stored. Thus, by proving that the system records the data and verifies it in a blockchain setting.

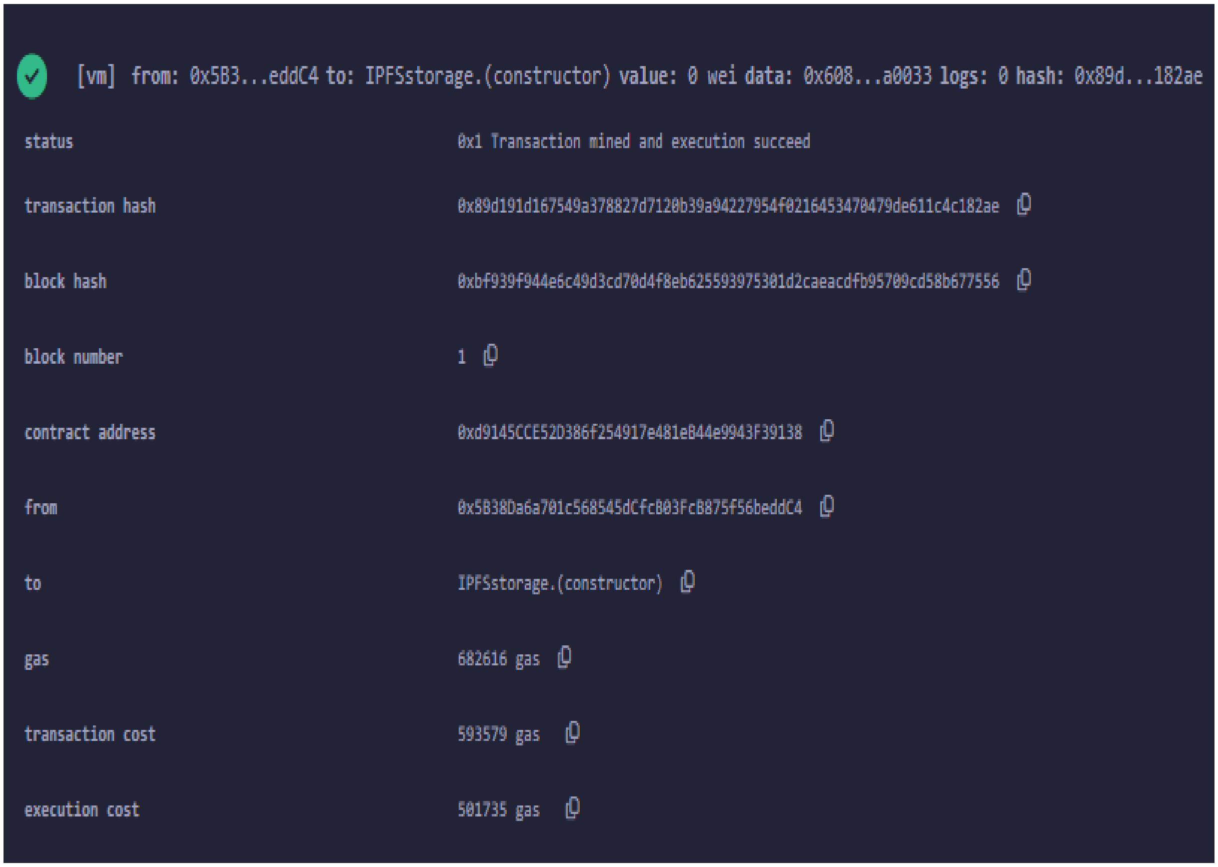

Keccak-256

Keccak-256 online hash function

Agriculture

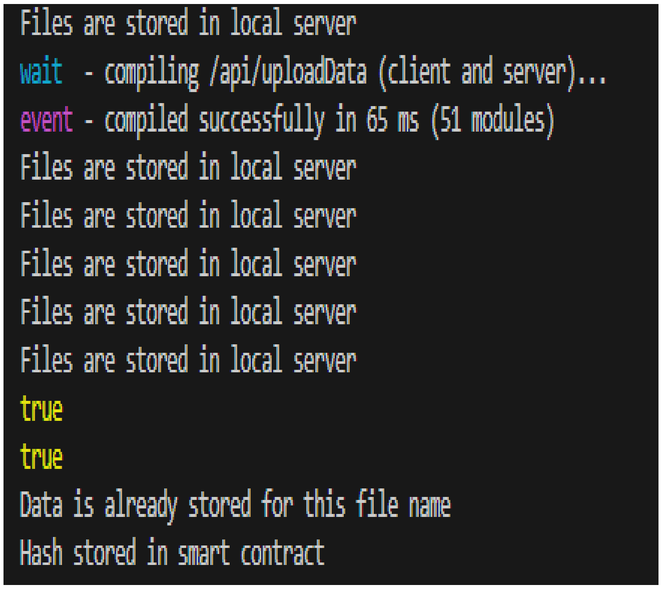
6931863ae4b49582ef011ae1ac8acc228b3c496c7280a628eb6e3f7dac018cb1

Fig. 3. Keccak-256 hash tool interface.



```
✓  [vm]  from: 0x5B3...eddC4 to: IPFSstorage.(constructor) value: 0 wei data: 0x608...a0033 logs: 0 hash: 0x89d...182ae

status                         0x1 Transaction mined and execution succeed

transaction hash               0x89d191d167549a378827d7120b39a94227954f0216453470479de611c4c182ae

block hash                     0xbf939f944e6c49d3cd70d4f8eb625593975301d2caeacdfb95709cd58b677556

block number                   1

contract address               0xd9145CCE52D386f254917e481eB44e9943F39138

from                           0x5B38Da6a701c568545dCfcB03FcB875f56beddC4

to                             IPFSstorage.(constructor)

gas                            682616 gas

transaction cost               593579 gas

execution cost                 501735 gas
```

Fig. 4. Blockchain transaction details.



```
Files are stored in local server
wait  - compiling /api/uploadData (client and server)...
event - compiled successfully in 65 ms (51 modules)
Files are stored in local server
Files are stored in local server
Files are stored in local server
Files are stored in local server
Files are stored in local server
true
true
Data is already stored for this file name
Hash stored in smart contract
```

Fig. 5. Execution output for data storage and verification.

Fig. 6 shows the performance of a DL-CNN model employed for predicting soil states as well as determining the need for fertilizers. Specifically, the metrics of accuracy (98.86%), precision (98.3%), sensitivity (98.3%), specificity (99.15%), F-measure (98.3%), Matthews-correlation coefficient (97.45%), and the negative predictive value (99.15%) highlight the model performance in soil condition classification. This in-depth evaluation establishes the model's high dependability and precision, which are necessary for ensuring accurate fertilizer recommendations.

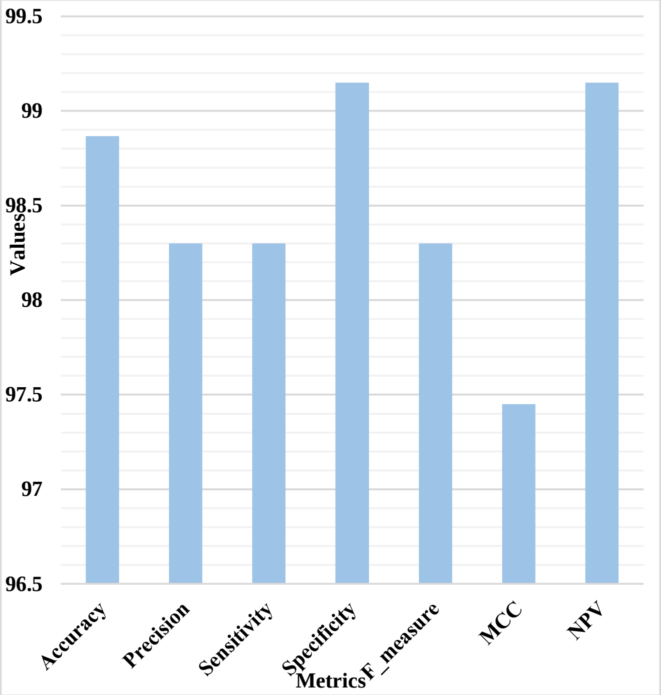


Fig. 6. Performance metrics of the DL-CNN classification model.

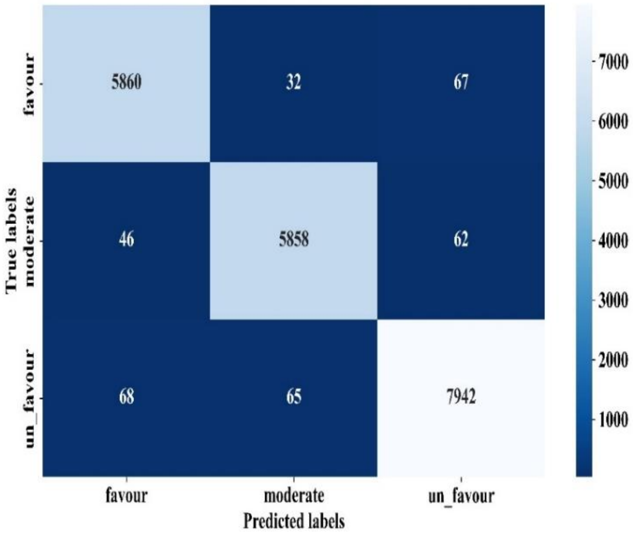
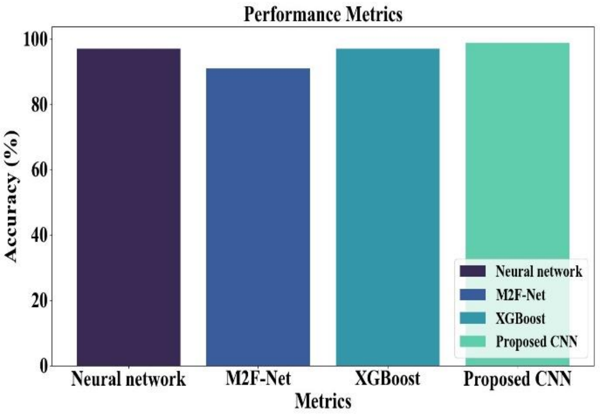


Fig. 7. Confusion matrix for DL-CNN classification model.

Fig. 7 presents a confusion matrix in evaluating the classification model's performance in predicting soil conditions. This matrix compares actual target values with the model's predictions across three classes: 'favor,' 'moderate,' and 'un favor.' The diagonal cells represent correctly classified instances. It indicates the model’s accuracy for each class, while the off-diagonal cells reveal misclassifications, indicating areas where the model's predictions deviate from actual values. This evaluation gauges the applicability of DL- CNN in the classification of soil pH and NPK by facilitating improvement for accurate fertilizer recommendations.



A suitable name for this figure could be:

Fig. 8. Accuracy comparison for soil condition analysis.

Fig. 8 presents a comparison of the performance of various state-of-the-art methods in terms of accuracy for soil condition analysis. Neural Network [38] with a result of 97%, while Multimodal Fusion Network (M2F-Net) [39] with a score of 91%, and eventually, XGBoost [40] with 97%. At last, the proposed CNN tops the list with an accuracy of 98.86% among all the methods being compared. This outstanding performance of the model is traced back to its architecture, which, in developing CNN, allows it to better identify soil pH and nutrient content classifications.

C. Discussion

This part is intended to demonstrate the proposed system's excellent abilities and working. Fig. 3 uses the Keccak-256 hash tool interface that models the secure hash function of data for trust in blockchain. Fig. 4 shows a detailed view of blockchain transaction information, thereby proving the secure administration and verification of transactions. Fig. 5 shows the output of data storage and verification, which firmly says that the storage of data on the blockchain and the fact that it was validated were successful. Fig. 6 shows the platform with the best performance, with the DL-CNN model reaching the highest accuracy of 98.86%. Thus, it is the most effective method when doing soil condition analysis and fertilizer recommendations. The confusion matrix of Fig. 7 is lastly used to provide the model's accuracy and identify the weak points in the model. Finally, various state-of-the-art methods can be compared with the proposed CNN, which has the highest accuracy and ability to classify soil pH and nutrient content. These results, in totality, assure the productivity and the benefits of relating IoT, blockchain, and DL technologies in smart agriculture to precision and security.

V. Conclusion

This research presents a revolutionizing system for making fertilizer use efficient in sustainable agriculture through the integration of IoT sensors, blockchain technology, and a DL-CNN. The suggested technique effectively addresses key challenges such as lack of data security and precise allocation of resources. The system uses IoT sensors to keep track of soil conditions and the IPFS and blockchain for secure data storage and transfer which guarantees the data conforms data integrity and confidentiality. The DL-CNN model delivers 98.86% classification accuracy, thus highlighting its capability for soil pH and nutrient levels evaluation. The real-time information given by a DE App to make the right decisions about the amount of fertilizer required for soil fertility improvement and, consequently, higher crop yields. Overall, this framework offers a solution for modernizing fertilizer management in agriculture, combining advanced technology with practical applications, guaranteeing increased agricultural output and sustainability. Nonetheless, unsuitable sensors can mislead outcomes, which can significantly impede the system's performance. Further research is likely to cover the remote sensing tools for pest detection and the possibility of the use of drone type of technology for precise water application.

References

[1] Kaur, A., Bhatt, D.P. and Raja, L., 2024. Developing a Hybrid Irrigation System for Smart Agriculture Using IoT Sensors and Machine Learning in Sri Ganganagar, Rajasthan. *Journal of Sensors*, *2024*(1), p.6676907.

[2] Rahaman, M., Lin, C.Y., Pappachan, P., Gupta, B.B. and Hsu, C.H., 2024. Privacy-Centric AI and IoT Solutions for Smart Rural Farm Monitoring and Control. *Sensors*, *24*(13), p.4157.

[3] Md. Mamun Hossain, Md. Ashiqur Rahman, Sudipto Chaki, Humayra Ahmed, Ahsanul Haque, Iffat Tamanna, Sweety Lima, Most. Jannatul Ferdous and Md. Saifur Rahman, “Smart-Agri: A Smart Agricultural Management with IoT-ML-Blockchain Integrated Framework” International Journal of Advanced Computer Science and Applications (IJACSA), 14(7), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01407107.

[4] Aliyu, A.A. and Liu, J., 2023. Blockchain-Based Smart Farm Security Framework for the Internet of Things. *Sensors*, *23*(18), p.7992.

[5] Akella, G.K., Wibowo, S., Grandhi, S. and Mubarak, S., 2023. A systematic review of blockchain technology adoption barriers and enablers for smart and sustainable agriculture. *Big Data and Cognitive Computing*, *7*(2), p.86.

[6] Lv, G., Song, C., Xu, P., Qi, Z., Song, H. and Liu, Y., 2023. Blockchain-based traceability for agricultural products: a systematic literature review. *Agriculture*, *13*(9), p.1757.

[7] Chen, H.Y., Sharma, K., Sharma, C. and Sharma, S., 2023. Integrating explainable artificial intelligence and blockchain to smart agriculture: Research prospects for decision making and improved security. *Smart Agricultural Technology*, *6*, p.100350.

[8] T C Jermin Jeaunita, Sarasvathi V, Fault Tolerant Sensor Node Placement for IoT based Large Scale Automated Greenhouse System, International Journal of Computing and Digital Systems, 8-2, 2019. http://dx.doi.org/10.12785/ijcds/080210

[9] Sabir Hussain Awan, Sheeraz Ahmed, Asif Nawaz, Sozan Sulaiman Maghdid, Khalid Zaman, M.Yousaf Ali Khan, Zeeshan Najam and Sohail Imran, "BlockChain with IoT, an Emergent Routing Scheme for Smart Agriculture" International Journal of Advanced Computer Science and Applications(IJACSA), 11(4), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110457

[10] Araújo, S.O., Peres, R.S., Ramalho, J.C., Lidon, F. and Barata, J., 2023. Machine learning applications in agriculture: current trends, challenges, and future perspectives. *Agronomy*, *13*(12), p.2976.

[11] Syed, L., 2024. Smart Agriculture using Ensemble Machine Learning Techniques in IoT Environment. *Procedia Computer Science*, *235*, pp.2269-2278.

[12] Adli, H.K., Remli, M.A., Wan Salihin Wong, K.N.S., Ismail, N.A., González-Briones, A., Corchado, J.M. and Mohamad, M.S., 2023. Recent advancements and challenges of AIoT application in smart agriculture: A review. *Sensors*, *23*(7), p.3752.

[13] Alahmad, T., Neményi, M. and Nyéki, A., 2023. Applying IoT sensors and big data to improve precision crop production: a review. *Agronomy*, *13*(10), p.2603.

[14] Naseer, A., Shmoon, M., Shakeel, T., Ur Rehman, S., Ahmad, A. and Gruhn, V., 2024. A Systematic Literature Review of the IoT in Agriculture-Global Adoption, Innovations, Security Privacy Challenges. *IEEE Access*.

[15] Taji, K. and Ghanimi, F., 2024. Enhancing security and privacy in smart agriculture: A novel homomorphic signcryption system. *Results in Engineering*, *22*, p.102310.

[16] Wei, Y., Wang, R., Zhang, J., Guo, H. and Chen, X., 2023. Partition management of soil nutrients based on capacitive coupled contactless conductivity detection. *Agriculture*, *13*(2), p.313.

[17] Thorat, T., Patle, B.K. and Kashyap, S.K., 2023. Intelligent insecticide and fertilizer recommendation system based on TPF-CNN for smart farming. *Smart Agricultural Technology*, *3*, p.100114.

[18] Senapaty, M.K., Ray, A. and Padhy, N., 2023. IoT-enabled soil nutrient analysis and crop recommendation model for precision agriculture. *Computers*, *12*(3), p.61.

[19] Pechlivani, E.M., Papadimitriou, A., Pemas, S., Ntinas, G. and Tzovaras, D., 2023. IoT-based agro-toolbox for soil analysis and environmental monitoring. *Micromachines*, *14*(9), p.1698.

[20] Indira, P., Arafat, I.S., Karthikeyan, R., Selvarajan, S. and Balachandran, P.K., 2023. Fabrication and investigation of agricultural monitoring system with IoT & AI. *SN Applied Sciences*, *5*(12), p.322.

[21] Vincent, D.R., Deepa, N., Elavarasan, D., Srinivasan, K., Chauhdary, S.H. and Iwendi, C., 2019. Sensors driven AI-based agriculture recommendation model for assessing land suitability. *Sensors*, *19*(17), p.3667.

[22] Singh, H., Halder, N., Singh, B., Singh, J., Sharma, S. and Shacham-Diamand, Y., 2023. Smart farming revolution: portable and real-time soil nitrogen and phosphorus monitoring for sustainable agriculture. *Sensors*, *23*(13), p.5914.

[23] Morchid, A., El Alami, R., Raezah, A.A. and Sabbar, Y., 2023. Applications of internet of things (IoT) and sensors technology to increase food security and agricultural Sustainability: Benefits and challenges. *Ain Shams Engineering Journal*, p.102509.

[24] Nayagam, M.G., Vijayalakshmi, B., Somasundaram, K., Mukunthan, M.A., Yogaraja, C.A. and Partheeban, P., 2023. Control of pests and diseases in plants using iot technology. *Measurement: Sensors*, *26*, p.100713.

[25] Rajak, P., Ganguly, A., Adhikary, S. and Bhattacharya, S., 2023. Internet of Things and smart sensors in agriculture: Scopes and challenges. *Journal of Agriculture and Food Research*, *14*, p.100776.

[26] Hasan, H.R., Musamih, A., Salah, K., Jayaraman, R., Omar, M., Arshad, J. and Boscovic, D., 2024. Smart agriculture assurance: IoT and blockchain for trusted sustainable produce. *Computers and Electronics in Agriculture*, *224*, p.109184.

[27] Frikha, T., Ktari, J., Zalila, B., Ghorbel, O. and Amor, N.B., 2023. Integrating blockchain and deep learning for intelligent greenhouse control and traceability. *Alexandria Engineering Journal*, *79*, pp.259-273.

[28] El Hathat, Z., Venkatesh, V.G., Sreedharan, V.R., Zouadi, T., Manimuthu, A., Shi, Y. and Srinivas, S.S., 2024. Leveraging Greenhouse Gas Emissions Traceability in the Groundnut Supply Chain: Blockchain-Enabled Off-Chain Machine Learning as a Driver of Sustainability. *Information Systems Frontiers*, pp.1-18.

[29] Toma, C., Popa, M., Boja, C., Ciurea, C. and Doinea, M., 2022. Secure and anonymous voting D-app with IoT embedded device using blockchain technology. *Electronics*, *11*(12), p.1895.

[30] Mahalingam, N. and Sharma, P., 2024. An intelligent blockchain technology for securing an IoT-based agriculture monitoring system. *Multimedia tools and applications*, *83*(4), pp.10297-10320.

[31] Srikanth, M., Mohan, R.J. and Naik, M.C., 2023. Blockchain-based consensus for a secure smart agriculture supply chain. *European Chemical Bulletin*, *12*(4), pp.8669-8678.

[32] Khan, Prince Waqas, Yung-Cheol Byun, and Namje Park. "IoT-blockchain enabled optimized provenance system for food industry 4.0 using advanced deep learning." Sensors 20, no. 10 (2020): 2990.

[33] Saba, T., Rehman, A., Haseeb, K., Bahaj, S.A. and Lloret, J., 2023. Trust-based decentralized blockchain system with machine learning using Internet of agriculture things. *Computers and Electrical Engineering*, *108*, p.108674.

[34] Hassan, A., Javed, S., Hussain, S., Ahmad, R. and Qazi, S., 2024. Arithmetic N-gram: an efficient data compression technique. *Discover Computing*, *27*(1), p.1.

[35] Ali, Alaa Abid Muslam Abid, Manar Joundy Hazar, Mohamed Mabrouk, and Mounir Zrigui. "Proposal of a Modified Hash Algorithm to Increase Blockchain Security." *Procedia Computer Science* 225 (2023): 3265-3275.

[36] Kosuru, V.S.R. and Kavasseri Venkitaraman, A., 2023. A smart battery management system for electric vehicles using deep learning-based sensor fault detection. *World Electric Vehicle Journal*, *14*(4), p.101.

[37] Abbas, S., Sampedro, G.A., Abisado, M., Almadhor, A., Kim, T.H. and Zaidi, M.M., 2023. A Novel Drug-Drug Indicator Dataset and Ensemble Stacking Model for Detection and Classification of Drug-Drug Interaction Indicators. *IEEE Access*.

[38] Musanase, C., Vodacek, A., Hanyurwimfura, D., Uwitonze, A. and Kabandana, I., 2023. Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices. Agriculture, 13(11), p.2141.

[39] Dhakshayani, J. and Surendiran, B., 2023. M2F-Net: A deep learning-based multimodal classification with high-throughput phenotyping for identification of overabundance of fertilizers. Agriculture, 13(6), p.1238.

[40] Senapaty, Murali Krishna, Abhishek Ray, and Neelamadhab Padhy. "A decision support system for crop recommendation using machine learning classification algorithms." Agriculture 14, no. 8 (2024): 1256.

A Capacity-Influenced Approach to Find Better Initial Solution in Transportation Problems

Md. Toufiqur Rahman[1], A R M Jalal Uddin Jamali[2],
Momta Hena[3], Mohammad Mehedi Hassan[4], Md Rafiul Hassan[5]
Department of Mathematics, Khulna University of Engineering and Technology, Bangladesh[1, 2, 3]
College of Computer and Information Science, King Saud University, Riyadh 11543, Saudi Arabia[4]
Department of Computer Science, University of Maine - Presque Isle, USA[5]

***Abstract*—Finding an Initial Basic Feasible Solution (IBFS) is the first and essential step in obtaining the optimal solution for any Transportation Problem. Numerous approaches are available in the literature to determine the IBFS; however, many of these methods are modifications of Vogel's Approximate Method (VAM) and/or the Least Cost Method (LCM). None of the existing methods directly consider the capacity of distributions among the nodes when selecting the allocation steps. While researchers have proposed various approaches and demonstrated improved solutions with numerical instances, they have not thoroughly investigated the underlying causes of these results. In this article, we explore the impact of capacity distributions among the nodes on the VAM and LCM in an experimental domain. The study introduces a novel and unique Capacity-Influenced Distribution Indicator (CI-DI) designed to control the flow of allocation. Ultimately, we propose a novel Capacity-Influenced approach that embeds both LCM and VAM to determine the IBFS for Transportation Problems (TPs). The novelty of the proposed approach lies in its direct consideration of capacity distribution among the nodes in the flow of allocations, this feature is lacking in LCM, VAM, and other established approaches. The proposed method develops a novel distribution indicator and a novel cost entry embedded capacity-based matrix to control the flow of allocations and thereby finds the IBFS for the Transportation Problem. We have conducted extensive numerical experiments to assess the effectiveness of the proposed approach. Experimental analysis demonstrates that the proposed method is more efficient in finding the IBFS than existing approaches. Moreover, as it uses a one-time generated Distribution Indicator (DI) for all steps of allocation, it is computationally cheaper than VAM, which generates a DI for each step of allocation.**

Keywords*—Transportation problem; least cost method; Vogel's approximate method; cost matrix; transportation tableau; node; capacity; route; capacity-influenced; weighted opportunity cost

I. Introduction

In Transportation Problem (TP), commodities are transported from a set of sources (called source nodes) to destinations (called destination nodes) subject to capacity (supply and demand) constraints in such a way that the total cost of transportation is minimized. TP is a multi-disciplinary field of study [1-3]. It is directly involved with real-life problems [1- 4]. The application of the TP extends beyond its traditional domain and finds relevance in various other fields [4], [8]. These fields include personnel assignment, inventory control, employee scheduling, and more [1-8]. It is known that the general Linear Programming (LP) methods are so tedious and time-consuming [6-8]. Researchers have developed several alternative methods for finding the IBFS by leveraging the special (unique) characteristics of TPs [2-6]. The two well-known classical methods, the LCM and VAM are very simple and can yield better IBFS for TPs [2], [5], [6], [14], etc.

In LCM, the flow of allocation is directly controlled by the cost matrix, with a preference for the least cost. Vogel introduced VAM in 1958 as a modification of LCM. In VAM, the flow of allocation is controlled differently, but it still considers the cost matrix. It begins by developing a control vector called the DI, formed through the manipulation of the cost matrix. Subsequently, the flow of allocation is guided by both the DI and the cost matrix. Based on the LCM and VAM, many researchers proposed several approaches for finding IBFS of TPs [2], [5], [6-13], [14], [21], [22], [23], [29-30], [38-39], [41], [42] etc. For the importance of TPs in real life, researchers are continuously devoted to finding better approaches for solving TPs. It is observed in the literature that most of the approaches are variations of VAM [6-10], [24], etc. A few of the research works related to TPs are pointed out below.

In the article [24], the author introduced the Total Opportunity Matrix (TOM) by manipulating cost entries rather than DI used in VAM to determine the flow of allocations. Authors in the article [1] developed Total Opportunity Cost (TOC) matrix and after that they formed DI tableau for allocation by considering TOC. In the article [43], authors proposed an enhanced version of [10] modified VAM for the unbalanced transportation problem. Both approaches utilize the VAM method, with the modification focused on transforming an unbalanced TP into a balanced one. In [44] author considered balanced transportation problems, with the modification applied solely to the manipulation of the cost matrix. In [45] authors proposed another embedded modified VAM method called Logical Development of Vogel's Approximation Method (LD-VAM) for finding the IBFS in TP. In [39] author introduced a modified VAM in which the modification is directed toward finding the DI, considering the cost matrix as well. Some other modified approaches based on VAM approach are found in the recent publications of [25], [29] [33], [36].

Besides modification of VAM, some other approaches are available in the literature. Recently, in [46] authors, at first, developed TOC then they formed DI tableau for allocation by considering the average of TOC of cells along each row identified as Row Average Total Opportunity Cost (RATOC) and the average of TOC of cells along each column identified as Column Average Total Opportunity Cost (CATOC). Allocations of costs are started in the cell along the row or column which has the highest RATOCs or CATOCs. This approach is also developed by considering only cost matrix.

In the article [35], author proposed a modified method to north-west corner method for finding IBFS. Very recently, some modified approaches based on LCM are found in the publications of [7], [26], etc. Some statistical methods are found in [36], [40]. In [35] authors have presented an alternative method of NWC method by using Statistical tool called Coefficient of Range (CoR) by statistically analyzing the cost matrix. Author, in [47], introduced a new algorithm for solving TPs. They proposed the Gauss Jordan pivoting method to solve the TPs. They consider only cost matrix and iteratively it finds out the solution. This algorithm is faster than Simplex method.

Some heuristic methods of TPs to find out IBFS are found in the articles [3], [9], [11], [12], [20], [21], [37], [48-56]. In fuzzy environment, many research publications are found in the literature of which some recent publications are included in the articles [13], [23], [27], [32]. Researchers also dedicated to find better IBFS for unbalanced TPs [10], [17], [32], [34]. A good survey of TP for finding IBFS is observed in [31], [57-58]. On the other hand, [15], [16] and [17] proposed a new technique for controlling the flow of allocation named Weighted Opportunity Cost (WOC) matrix. The WOC matrix is formed by demand and/or supply as a weight factor corresponding to each transportation cost. Authors in [15], [16] and [17] also considered some numerical instances to test the efficiency of the proposed algorithms. In [59] authors considered a fractional objective function rather than a linear objective to solve TPs. In [60], presented a modified VAM specifically designed for maximizing profit, with the flow of allocation controlled by cost entries as well. Moreover, in recently published articles [61-64], authors have proposed various methods to find the IBFS of TPs in which the flow of allocations is controlled by manipulating only cost entries.

It is observed that many approaches are available in the literature, and researchers are continuously working to develop more efficient methods to solve the TPs. But as far as it is known, none of the approaches is the best for solving all TPs. In our earlier work [15] Jamali also noticed this pitfall and proposed a newer approach to LCM. Though it [15] performed better compared to LCM but it frequently obtains worse IBFS compared to VAM.

It is known that classical methods like the LCM and VAM are relatively less computationally expensive compared to other approaches for finding the Initial Basic Feasible Solution [28] (IBFS) of Transportation Problems (TPs) [18]. Although VAM generally performs better than LCM, there are cases where LCM outperforms VAM. This raises the question: Is the performance variation due to the distribution of node capacities?

Additionally, it is observed that in the transportation sector, increasing the amount of goods often reduces transportation costs. Based on this observation, we investigate how node capacities influence transportation costs. Notably, no researchers have yet developed methods that leverage the effect of node capacity on the allocation and flow of commodities.

In this paper, we extensively examined the impact of capacity distribution among the nodes on classical well-known approaches, namely LCM and VAM, considering the problem mentioned in the existing literature. Next, we developed a capacity-blended flow-controlling matrix based on the cost matrix. Finally, we proposed a novel capacity-influenced approach to find the IBFS of TPs.

The novelties of the proposed approach are explained in the following:

- We propose a novel approach to find better IBFS for the TPs by developing a new, unique distribution indicator (DI) that combines the capacity vectors and DI vector to control the flow of allocation.
- Our approach proposes a more effective solution for the TPs that can overcome the limitations of existing methods, namely VAM and LCM, by addressing the impact of capacity distribution among the nodes.
- Proposed approach introduces a capacity-influenced weighted factor and a capacity-influenced Weighted Opportunity Cost (WOC) Matrix to find a more suitable IBFS in a computationally efficient way for VAM and LCM.
- We extensively verified the impact of capacity distribution and proposed methods have been experimentally verified to evaluate its performance over other approaches using real examples.

The remainder of the paper is organized as follows. Section II explains the mathematical model of TPs. Section III extensively examines the impact of capacity distribution among nodes for both LCM and VAM for TPs. Section IV explains the proposed method of this paper. This section develops and explains proposed mathematical model named as “Capacity influenced weighted factor” to control the flow of allocations. It also develops the formulation of control matrix of the proposed approach. Detailed results of numerical experiments to validate the efficiency of the proposed method is explained in Section V and conclusion of this study is presented in the last section.

II. Mathematical Model of Transportation Problem

By considering the equality characteristics of TPs, it can be represented using a specialized tableau known as the Transportation Tableau (TT). The typical view of TT is shown in Table I. In the TT, O_i indicates the ith source with the amount of availability is ai which is shown in the far-right column. On the other hand, D_j denotes the jth destination with demand b_j, which is shown in the bottom row of TT. In this table, there is an m×n matrix containing cost entries. The cell in the *i*th row and *j*th column is called the C_{ij} cell and the transportation cost is denoted as cij, which represents the unit shipping cost from the *i*th source to the *j*th destination. So, a TT can be viewed as a (m+1) × (n+1) matrix shown in Table I.

TABLE I. A TT OF A TP WITH M SOURCES AND N DESTINATIONS

	Sinks/Destinations						
Origins/Sources		D_1	D_2	$\cdots$	D_{n-1}	D_n	Supply
	O_1	c_{11}	c_{12}	$\cdots$	c_{1n-1}	c_{1n}	a_1
	O_2	c_{21}	c_{22}	$\cdots$	c_{2n-1}	c_{2n}	a_2
	O_3	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	a_3
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
	O_m	c_{1m}	c_{2m}	$\cdots$	c_{mn-1}	c_{mn}	a_m
	Demand	b_1	b_2	$\cdots$	b_{n-1}	b_n	

III. IMPACT OF THE CAPACITY DISTRIBUTIONS AMONG NODES ON LEAST COST AND VOGEL'S APPROXIMATION METHODS

First three typical balanced TPs are considered which are presented in the Examples 1(a) – 1(c) and their corresponding comparative analyses are presented in Table II to Table X. It should be noted that in all three TPs, the cost matrices are identical, and the total capacity remains the same. The only difference lies in the distribution of capacity among the nodes. Numerical experiments have been conducted to investigate the effect of capacity on both the LCM and VAM in terms of finding the IBFS and total cost [19].

Example 1(a):

TABLE II. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 1

	D_1	D_2	D_3	Supply
O_1	2	5	8	20
O_2	6	4	14	20
O_3	15	12	13	20
Demand	20	20	20	

Example 1(b):

TABLE III. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 2

	D_1	D_2	D_3	Supply
O_1	2	5	8	30
O_2	6	4	14	20
O_3	15	12	13	10
Demand	10	20	30	

Example 1(c):

TABLE IV. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 3

	D_1	D_2	D_3	Supply
O_1	2	5	8	30
O_2	6	4	14	20
O_3	15	12	13	10
Demand	20	35	5	

TABLE V. (EX. 1(A) LCM) STEP-BY-STEP FLOW OF ALLOCATIONS OF LCM AND VAM FOR EX. 1(A) –1(C)

LCM: IBFS of Ex. 1 (a)				
	D_1	D_2	D_3	S
O_1	**2** **20**	5 0	8 0	20
O_2	6 0	**4** **20**	14 0	20
O_3	15 0	12 0	**13** **20**	20
D	20	20	20	

TABLE VI. (EX 1(A) VAM) STEP-BY-STEP FLOW OF ALLOCATIONS OF LCM AND VAM FOR EX. 1(A) –1(C)

VAM: IBFS of Ex. 1 (a)				
	D_1	D_2	D_3	S
O_1	2 **0**	5 0	**8** **20**	20
O_2	**6** **20**	4 **0**	14 0	20
O_3	15 0	**12** **20**	13 **0**	20
D	20	20	20	

TABLE VII. (EX. 1(B) LCM) STEP-BY-STEP FLOW OF ALLOCATIONS OF LCM AND VAM FOR EX. 1(A) –1(C)

LCM: IBFS of Ex. 1 (b)				
	D_1	D_2	D_3	S
O_1	**2** **10**	5 0	**8** 20	30
O_2	6 0	**4** 20	14 0	20
O_3	15 0	12 0	**13** 10	10
D	10	20	30	

TABLE VIII. (Ex. 1(B) VAM) Step-by-Step Flow of Allocations of LCM and VAM for Ex. 1(a) –1(c)

VAM: IBFS of Ex. 1 (b)				
	D_1	D_2	D_3	S
O_1	2 0	5 0	**8** 30	30
O_2	**6** 10	**4** 10	14 0	20
O_3	15 0	**12** 10	13 0	10
D	10	20	30	

TABLE IX. (Ex. 1(C) LCM) Step-by-Step Flow of Allocations of LCM and VAM for Ex. 1(a) –1(c)

LCM: IBFS of Ex. 1 (c)				
	D_1	D_2	D_3	S
O_1	**2** 20	**5** 10	8 0	30
O_2	6 0	**4** 20	14 0	20
O_3	15 0	**12** 5	**13** 5	10
D	20	35	5	

TABLE X. (Ex. 1(C) VAM) Step-by-Step Flow of Allocations of LCM and VAM for Ex. 1(a) –1(c)

VAM: IBFS of Ex. 1 (c)				
	D_1	D_2	D_3	S
O_1	**2** 20	**5** 5	**8** 5	30
O_2	6 0	**4** 20	14 0	20
O_3	15 0	**12** 10	13 0	10
D	20	35	5	

To examine the effect of capacity on LCM and VAM, we have first compared the step-by-step flow of allocations. The intensive comparison of the step-by-step flow of the allocations procedure of the two approaches is concisely shown in Table XI. It is observed in Table XI that due to the change of capacity distribution among the nodes, the pattern of flow of allocation is changed significantly for both approaches. Furthermore, it is observed that the IBFS obtained using both methods undergo significant changes for each instance.

Now we have compared IBFSs and Total Transportation Costs (TTC) for each instance. The effect of the capacity distribution of nodes for each instance is shown in Table XI. It is observed that only by a change of capacity distribution among the nodes, the total transportation cost for each instance produced by LCM is changed significantly. Similarly, only by a change of capacity distribution among the nodes, the total transportation cost for each instance produced by VAM is changed significantly. It is also observed that the IBFSs produced by LCM of the three instances are changed significantly. Similarly, it is also observed that the IBFSs produced by VAM of the three instances are changed significantly. It has been observed that in Example 1(a), the LCM produced a superior IBFS compared to VAM, with a notable and significant difference between the two solutions. But only the change of capacity distribution, in Example 1(b), VAM produced a better solution compared to LCM. Moreover, in Example 1(c), though LCM produced a better solution, but the difference between the two solutions is not large.

TABLE XI. Comparison Between Three Examples Regarding the Effect of Capacity

	Ex. 1(a) Equal capacity Demand: 20, 20, 20 Supply: 20, 20, 20		Ex. 1(b) Unequal Capacity Demand: 30, 20,120 Supply: 10, 20, 30		Ex1 (c) Unequal Capacity Demand: 30, 20, 10 Supply: 20, 35, 5	
	T. Cast	IBFS	T. Cost	IBFS	T. Cost	IBFS
LCM	**380**	$x_{11}\rightarrow x_{22}\rightarrow x_{33}$ 20→ 20 → 20	520	$x_{11}\rightarrow x_{22}\rightarrow x_{13}\rightarrow x_{33}$ 10 → 20 → 20→ 10	295	$x_{11}\rightarrow x_{22}\rightarrow x_{12}\rightarrow x_{32}\rightarrow x_{33}$ 20 → 20→10 → 5 → 5
VAM	520	$x_{13}\rightarrow x_{21}\rightarrow x_{32}$ 20 →20 →20	**470**	$x_{13}\rightarrow x_{21}\rightarrow x_{22}\rightarrow x_{32}$ 30 →10→ 10 →10	**305**	$x_{13}\rightarrow x_{11}\rightarrow x_{22}\rightarrow x_{12}\rightarrow x_{32}$ 5 → 20 → 20 → 5 → 10

We have performed further experiments to examine the effect of capacity distribution on the flow of allocations of both LCM and VAM. For this numerical experiment, we have considered some more examples shown in Table XII.

We have performed both approaches, namely LCM and VAM to find out IBFS. The experimental result is shown in Table XII. It is observed in Table XII that instances 2(a) –2(c) has the same cost matrix but the capacity distributions are different. Though instances 2(a) – 2(c) have different distributions of capacity, the total supply/demand is equal for each instance. Similarly, instances 3(a) to 3(c) have identical cost matrices, but the capacity distributions vary. Though Examples 2(a) – 2(c) have different capacity distributions but total supply/demand is equal for each instance. It is observed for Example 2(a), that VAM produced a better solution which is also an optimal solution, but only a change of capacity distribution, for Example 2(b) and 2(c), LCM obtained a better solution compared to VAM. Similarly, it is observed that for Example 3(a) LCM produced a better solution which is also an optimal solution, but only a change of capacity distribution, for Example 3(c), LCM obtained a better solution compared to VAM, but for Example 3(b) VAM obtained a better solution compared to LCM. The experimental results are examined intensively. We have found out significant effects of the distribution of capacities on the IBFS of each approach.

TABLE XII. SOME MORE EXAMPLES REGARDING LCM AND VAM

Instances	Cost matrix	Capacity	Total Capacity	LCM	VAM	Optimal

2 (a)	{7,8,7};
	{18,8,12};
	{8,12,12}
		S:12, 12, 12
		D:12, 12, 12
			36	324	**276**	**276**

2 (b)	{7,8,7};
	{18,8,12};
	{8,12,12}
		S; 14,14,8
		D: 15,8,13
			36	**326**	334	**298**

2 (c)	{7,8,7};
	{18,8,12};
	{8,12,12}
		S:25,30, 5
		D:35,20, 5
			60	**525**	555	**525**

3 (a)	{2,5,8};{6,4,14}; {15,12,18}	S:20, 20, 20 D:20, 20, 20	**60**	**480**	520	**480**
3 (b)	{2,5,8};{6,4,14}; {15,12,18}	S: 43,15,2 D:2,53,5	60	314	**308**	**308**
3 (c)	{2,5,8};{6,4,14}; {15,12,18}	S:10, 10, 10 D:10, 10, 10	30	**240**	260	**240**

IV. OUR PROPOSED METHOD: A NOVEL DISTRIBUTION INDICATOR AND CAPACITY-INFLUENCED APPROACH TO IBFS FOR TRANSPORTATION PROBLEMS

In classical approaches, like North-West Corner Rule [35], LCM, the flow of allocation is controlled directly by the cost entries only. Once again, in some other classical transportation approaches, such as Vogel's VAM method and all its variants, the flow of allocations is controlled solely by the manipulation of cost entries. Moreover, based on our literature review, it is observed that almost all approaches have been developed by manipulating cost entries exclusively. None of these approaches considers the node's capacity when formulating the control matrix for the flow of allocation. However, as observed in the previous section of this article, the distribution of commodities plays a crucial role in controlling the flow of allocations. In the previous investigation section, it is observed that the distribution of commodities significantly alters the performance of the approaches. In this article, we first developed a novel capacity-influenced weighted factor and then proposed a capacity-influenced WOC Matrix-based algorithm to find the IBFS of TPs. These are explained in the following sections.

A. *The Proposed Capacity-Influenced Weighted Factor to Control the Flow of Allocation*

Vogel's method formulates the DI by calculating the difference between the smallest and the next-to-smallest cost entries for each node (supply/destination). In this formulation, DI serves as a weight factor corresponding to the cost entries, controlling the flow of allocations. While existing literature emphasizes DI as a controlling tool, it is observed in the previous section that node capacity also significantly influences allocation flow, alongside cost entries. Therefore, a novel approach is needed to control the flow of allocations in TPs by combining the weight factor DI with the weight factor of the corresponding node's capacity. The primary challenge involves developing a suitable weight factor for the capacity of each node. Subsequently, a new flow of allocation matrix needs to be formulated, incorporating both the capacity of nodes and DI as a combined weight factor for the corresponding cost entries. The first challenge is to find out the weight factor to the cost matrix from the capacity of nodes.

The WOC matrix is a new concept to control the flow of allocations of TP to find out IBFS. In WOC, amount of supply/demand of each route is a weighted factor corresponding to the cost entry. The procedures to find out capacity influenced approach is discussed step by step below:

Step 1: Finding cell weight: At first, we have found out the maximum possible allocation of the cell $\mathrm{C_{ij}}$, which is $(\mathrm{S_i}, \mathrm{D_j})$, where $\mathrm{S_i}$ denotes total supply at node i and $\mathrm{D_j}$ indicates total demand at node j. Therefore, sum over of all possible allocations is as follow:

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\min\left(\mathrm{S_i}, \mathrm{D_j}\right)$$

Therefore, for each cell C_{ij} its weight will be as follows

$$\min\left(\mathrm{S_i}, D_j\right) \Big/ \sum_{i=1}^{p}\sum_{j=1}^{q}\min\left(\mathrm{S_i}, D_j\right)$$

So that total weight becomes one. i.e.,

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\left\{ \min\left(\mathrm{S_i}, D_j\right) \Big/ \sum_{i=1}^{p}\sum_{j=1}^{q}\min\left(\mathrm{S_i}, D_j\right) \right\} = 1$$

But since every cell of cost matrix will contain the factor $1 \Big/ \sum_{i=1}^{p}\sum_{j=1}^{q}\min\left(\mathrm{S_i}, \mathrm{D_j}\right)$, so we have ignored this factor to reduce computational cost. Therefore, for each cell C_{ij} its weight will be just $min\ (S_i\ , D_j)$.

Step 2: The second challenge is to find out the combined weight factor formulated by the weight factor DI and the weight factor WOC to the cost matrix for each route (cell). The proposed modified weight factor, $\mathrm{W_{ij}^{m}}$, for the cell Cij is formulated as follows in equation (1):

$$W_{ij}^{m} = \min\{a_i, b_j\} \cdot \max\{\{D_i^{Ir}, D_j^{Ic}\}, \} \qquad (1)$$

Where a_i is the amount capacity of source node O_i, b_i is the amount of the capacity of destination node D_i. Moreover, D_i^{Ir} is the DI corresponding to the source node i and D_j^{Ic} is the DI corresponding to the sink node j. Then total weight corresponding to all cells will be as in Eq. (2).

$$TW = \sum_{i}^{m}\sum_{j}^{n}\min\{a_i, b_j\} \cdot \max\{\{D_i^{Ir}, D_j^{Ic}\} \forall\ i = 1,2,\dots,m\ ; j = 1,2,\dots,n\} \qquad (2)$$

Then, the actual weight factor corresponding to each route (cell) C_{ij} is

$$AW_{ij}^{m} = \frac{\min\{a_i,b_j\}\cdot\max\left\{\{D_i^{Ir},D_j^{Ic}\right\}}{\sum_i^m\sum_j^n \min\{a_i,b_j\}\cdot\max\left\{\{D_i^{Ir},D_j^{Ic}\right\}}, \forall\, i = 1,2,\dots,m\,; j = 1,2,\dots,n \qquad (3)$$

Here in Eq. (3), $\max\{\{D_k^{Ir}, D_l^{Ic}\}\ \forall\, i,j$, is fixed and constant and also $\sum_i^m\sum_j^n AW_{ij}^{m} = 1$.

But since the term $\frac{1}{\sum_i^m\sum_i^n \min\{a_i,b_j\}\cdot\max\left\{\{D_i^{Ir},D_i^{Ic}\right\}}$ is common to all AW_{ij}^{m} $\forall i,j$ and as AW_{ij}^{m} act as a controller of the flow of allocation and it has no any real effect to measure the total transportation cost, so without loss of generality we can ignore this factor in flow of allocation matrix.

Therefore, the weight factor corresponding to the cell C$_{ij}$ can be expressed as follows in Eq. (4):

$$W_{ij}^{m} = \min\{a_i,b_j\} \times \max\left\{\{D_i^{Ir},D_j^{Ic}\right\} \qquad (4)$$

It is noted that this reduces a significant amount of computational cost. The significance of this weight is that larger weight poses larger possibility to flow of allocation.

Step 3: After the successful formulation of the weight factor corresponding to each cell (route), our next task is to formulate the capacity-influenced WOC Matrix. But now we have to face a problem regarding the accumulation of this weight factor to any cost entries. Since the cell with a lower cost has a preference for allocation first, on the other hand, the cell with a larger weight factor has a preference for allocation first. So, it is not directly possible to formulate a WOC matrix just by simply multiplying the weight factor by the cost entry to find out meaningful elements of the WOC matrix. To overcome this difficulty and for the formulation of a meaningful WOC matrix, we should transform one of the two so that the multiplication of the two will be meaningful. This can be done by inversing the cost elements. Therefore, the weighted opportunity cost corresponding to the cell cost C$_{\text{ij}}$ as stated in Eq. (5) below:

- $W_{c_{ij}}^{m} = \frac{1}{c_{ij}}\min\{a_i,b_j\} \times \max\left\{D_i^{Ir},D_j^{Ic}\right\}, c_{ij} \neq 0$ (5)

Here $W_{c_{ij}}^{m}$ and c_{ij} denote the modified weighted cost factor and actual cost entry corresponding to the cell C_{ij} respectively.

But another problem arises if the cost at any cell is zero. Since when $c_{ij} = 0$ then $\frac{1}{c_{ii}}$ becomes undefined. So, to overcome this difficulty it is needed some more special attention. We can overcome this shortcoming by replacing zero with a significantly large value. So, if there exists any cell whose cost entry is zero, then we can formulate the virtual weighted cost to the cell C_{pq} as follows:

- If $c_{pq} = 0$ and $\{c_{ij}: 0 < c_{ij} < 1, \forall i,j\} = \varphi$, (i.e. null set) then set $W_{c_{ij}}^{m} = \max\left\{a_i, b_j; \forall i,j\right\} \times \min\{a_i,b_j\} \times \max\left\{\{D_i^{Ir}, D_j^{Ic}\right\}$
- Else if $c_{pq} = 0$ and $\{c_{ij}: 0 < c_{ij} < 1, \forall i,j\} \neq \varphi$ (i.e. not null set), then set $W_{c_{ij}}^{m} = \frac{\max\left\{a_i,b_j;\forall i,j\right\}}{[\min\left\{c_{ij}:0<c_{ij}<1,\forall i,j\right\}]} \times \min\{a_i,b_j\} \times \max\left\{\{D_i^{Ir},D_j^{Ic}\right\}$

B. *Proposed Formulation of the Capacity-Influenced Control Matrix*

After development of modified weighted cost factor, we can easily formulate the Modified Weighted Cost (MWOC) Matrix$\left[W_{c_{ij}}^{m}\right]$ which is as follows:

- If $c_{ij} \neq 0$; $W_{c_{ij}}^{m} = \frac{1}{c_{ij}}\min\{a_i,b_j\} \times \max\left\{D_i^{Ir},D_j^{Ic}\right\}$
- If $c_{ij} = 0$ and $\{c_{pq}: 0 < c_{ij} < 1, \forall p,q\} = \varphi$, (i.e. null set) then set $W_{c_{ij}}^{m} = \max\left\{a_i, b_j; \forall i,j\right\} \times \min\{a_i,b_j\} \times \max\left\{\{D_i^{Ir}, D_j^{Ic}\right\}$
- Else if $c_{pq} = 0$ and $\{c_{pq}: 0 < c_{ij} < 1, \forall p,q\} \neq \varphi$ (i.e. not null set), then set $W_{c_{ij}}^{m} = \frac{\max\left\{a_i,b_j;\forall i,j\right\}}{[\min\left\{c_{ij}:0<c_{ij}<1,\forall i,j\right\}]} \times \min\{a_i,b_j\} \times \max\left\{\{D_i^{Ir},D_j^{Ic}\right\}$

C. *Algorithm of Proposed Capacity-Influenced Approach*

Step 1: Form the MWOC weight factor matrix.

Step 2: Allocate (as much as possible), i.e., min {Si, Di}, to the cell (route) which has the largest weight factor.

Step 3: Update the cost matrix by crossing out exhausted cells and corresponding weight factors.

Step 4: Terminate if all demand requirements are satisfied; otherwise, go back to step 2.

Once all cells are allocated, calculate the total transportation cost by multiplying the allocated units with their respective costs and summing up all these products.

V. NUMERICAL EXPERIMENTATION

Now we will implement the proposed method and will compare its performance with existing approaches namely LCM and VAM. For the experimental study we have considered another TP 4 (see Example 4) whose TT is given in the Table XIII.

Example 4:

TABLE XIII. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 4

	D$_1$	D$_2$	D$_3$	Supply
O$_1$	4	3	5	90
O$_2$	6	5	4	80
O$_3$	8	10	7	100
Demand	70	120	80	

TABLE XIV. MODIFIED WEIGHTED OPPORTUNITY COST MATRIX OF THE PROBLEM 4

	D$_1$	D$_2$	D$_3$	Supply	DI
O$_1$	$\frac{140}{4}$	$\frac{180}{3}$	$\frac{80}{5}$	90	1
O$_2$	$\frac{140}{6}$	$\frac{160}{5}$	$\frac{80}{4}$	80	1
O$_3$	$\frac{140}{8}$	$\frac{200}{10}$	$\frac{80}{7}$	100	1
Demand	70	120	80		
DI	2	2	1		

To find the IBFS of the proposed method, we first need to find out the flow of the control matrix called the Modified Weighted Opportunity Cost (MWOC) matrix shown in Table XIV. To explain how to form the MWOC matrix, let us consider the cell C_{12}. It is observed that the unit cost $c_{12} = 4$ which is not zero. Therefore, the algorithm executes case (a) of the proposed method, formally:

$$W^{m}_{c_{12}} = \frac{1}{c_{12}} \min\{a_1, b_2\} \times \max\{ D_1^{Ir}, D_2^{Ic}\}$$

$$= \frac{1}{3} \times \min\{90,120\} \times \max\{ 1, 2\}$$

$$= \frac{1}{3} \times 90 \times 2 = \frac{180}{3}$$

TABLE XV. MODIFIED WEIGHTED OPPORTUNITY COST INCLUDED TRANSPORTATION TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply	DI
O_1	$\frac{140}{4}$ 4	$\frac{180}{3}$ 3	$\frac{80}{5}$ 5	90	1
O_2	$\frac{140}{6}$ 6	$\frac{160}{5}$ 5	$\frac{80}{4}$ 4	80	1
O_3	$\frac{140}{8}$ 8	$\frac{200}{10}$ 10	$\frac{80}{7}$ 7	100	1
Demand	70	120	80		
DI	2	2	1		

We can represent the cost matrix and the MWOC matrix in a single tableau as shown in Table XV. In Table XV, the entry in the top-left corner of each cell represents the weighted opportunity cost factor associated with that cell, while the entry in the top-right corner represents its transportation cost. The step by step of each allocation's procedure of the proposed MWOC-based approach is shown in Tables XVI to XXI. The IBFS of the problem obtained by the proposed method is shown in Table XVI. It is observed that the TTC for finding the IBFS by the proposed method is 1440.

TABLE XVI. THE FIRST ALLOCATION OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply	DI
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 **90**	$\frac{80}{5}$ 5 ×	~~**90**~~	1
O_2	$\frac{140}{6}$ 6	$\frac{160}{5}$ 5	$\frac{80}{4}$ 4	80	1
O_3	$\frac{140}{8}$ 8	$\frac{200}{10}$ 10	$\frac{80}{7}$ 7	100	1
Demand	70	~~**120**~~ , 30	80		
DI	2	2	1		

TABLE XVII. THE SECOND ALLOCATION OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 90	$\frac{80}{5}$ 5 ×	~~90~~
O_2	$\frac{140}{6}$ 6	$\frac{160}{5}$ 5 **30**	$\frac{80}{4}$ 4	~~**80**~~,50
O_3	$\frac{140}{8}$ 8	$\frac{200}{10}$ 10 ×	$\frac{80}{7}$ 7	100
Demand	70	~~120~~ , ~~**30**~~	80	

TABLE XVIII. THE THIRD ALLOCATION OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 90	$\frac{80}{5}$ 5 ×	~~90~~
O_2	$\frac{140}{6}$ 6 **50**	$\frac{160}{5}$ 5 30	$\frac{80}{4}$ 4 ×	~~80~~,~~**50**~~
O_3	$\frac{140}{8}$ 8	$\frac{200}{10}$ 10 ×	$\frac{80}{7}$ 7	100
Demand	~~**70**~~, 20	~~120~~ , ~~**30**~~	80	

TABLE XIX. THE FOURTH ALLOCATION OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 90	$\frac{80}{5}$ 5 ×	~~90~~
O_2	$\frac{140}{6}$ 6 50	$\frac{160}{5}$ 5 30	$\frac{80}{4}$ 4 ×	~~80,50~~
O_3	$\frac{140}{8}$ 8 **20**	$\frac{200}{10}$ 10 ×	$\frac{80}{7}$ 7	~~**100**~~, 80
Demand	~~70~~, ~~**20**~~	~~120~~ , ~~30~~	80	

TABLE XX. THE FIFTH ALLOCATION OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 90	$\frac{80}{5}$ 5 ×	~~90~~
O_2	$\frac{140}{6}$ 6 50	$\frac{160}{5}$ 5 30	$\frac{80}{4}$ 4 ×	~~80,50~~
O_3	$\frac{140}{8}$ 8 20	$\frac{200}{10}$ 10 ×	$\frac{80}{7}$ 7 **80**	~~100, **80**~~
Demand	~~70~~, ~~20~~	~~120~~ , ~~30~~	~~**80**~~	

TTC = 3×90+6×50+5×30+8×20+7×80 =1440

TABLE XXI. THE IBFS OF THE PROPOSED METHOD FOR THE TABLEAU PROBLEM 4

	D_1	D_2	D_3	supply
O_1	$\frac{140}{4}$ 4 ×	$\frac{180}{3}$ 3 90	$\frac{80}{5}$ 5 ×	~~90~~
O_2	$\frac{140}{6}$ 6 50	$\frac{160}{5}$ 5 30	$\frac{80}{4}$ 4 ×	~~80,50~~
O_3	$\frac{140}{8}$ 8 20	$\frac{200}{10}$ 10 ×	$\frac{80}{7}$ 7 **80**	~~100, **80**~~
Demand	~~70~~, ~~20~~	~~120~~ , ~~30~~	~~**80**~~	

Now have solved the problem with the existing LCM, VAM, and WOC-LCM and compared it with the proposed method named MWOC-VAM. The comparison is shown in Table XXII. It is observed in Table XXII that the proposed MWOC-VAM needs the least amount of transportation cost to obtain the IBFS compared to all other approaches namely LCM, VAM, and WOC-LCM. It is also observed that the starting allocation of LCM, WOC-LCM, and MWOC-VAM are the same but differ from VAM. It is also observed in the second column of Table XXII that the pattern of the flow of allocation for each approach is different.

TABLE XXII. COMPARISON REGARDING THE FLOW OF ALLOCATIONS AND TOTAL TRANSPORTATION COST TO FIND OUT THE IBFS OF THE PROBLEM 4

Method	Flow of allocations and IBFS	Total Cost
LCM	$x_{12} \rightarrow x_{23} \rightarrow x_{31} \rightarrow x_{32}$ 90→ 80 →70 → 30	1450
VAM	$x_{11} \rightarrow x_{33} \rightarrow x_{12} \rightarrow x_{22} \rightarrow x_{32}$ 70→ 80 →20 → 80 → 20	1500
WOC-LCM	$x_{12} \rightarrow x_{23} \rightarrow x_{32} \rightarrow x_{31}$ 90 → 80 → 30→ 70	1450
MWOC-VAM (proposed)	$x_{12} \rightarrow x_{22} \rightarrow x_{21} \rightarrow x_{31} \rightarrow x_{33}$ 90→ 30 → 50 → 20 → 80	**1440**

Now we have considered another problem 5 (Example 5), whose TT is given in the Table XXIII in which one route has zero transportation cost.

Example 5:

TABLE XXIII. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 5

	D_1	D_2	D_3	Supply
O_1	1	16	17	10
O_2	**0**	6	8	2
O_3	3	3	7	3
Demand	10	3	2	

We have again represented the cost matrix and the MWOC matrix in a single tableau in Table XXIV. It is observed in Table XXIV that c_{21} = 0 and its corresponding weight cost factor is 120. Once again, let's illustrate how to calculate the weight factor for that unit cost entry. According to the proposed method, since c_{21} = 0, the algorithm executes case (b). Formally:

(b) As $c_{21} = 0$ and $\{c_{pq}: 0 < c_{ij} < 1, \forall p, q\} = \varphi$, (i.e. null set), so

$$W_{c_{21}}^{m} = \max\{a_i, b_j; \forall i, j\} \times \min\{a_2, b_1\} \times \max\{D_2^{Ir}, D_1^{Ic}\}$$

$$=\max\{10,2,3; 10,3,2\} \times \min\{2,10\} \times Max\{1,6\}$$

$$= 10 \times 2 \times 6 = 120$$

It is observed in the Table XXIV that the weight opportunity cost of the cell C_{21} is 120 corresponding to the minimal cost i.e., c_{21} = 0 obtained by the case (b) of the proposed approach. On the other hand, the weight opportunity cost of the cell C_{11} is 150 corresponding to the cost 1 i.e., c_{11} = 1. It is worthwhile to mention here that the weight factor corresponding to the route (cell C_{11}) is largest though it's cost entry is not minimum. Now we have solved the problem with the proposed MWOC-VAM as well as LCM, VAM, and WOC-LCM. The experimental result is shown in Table XXV.

It is observed in Table XVI that the proposed MWOC-VAM and VAM need the least amount of transportation cost to obtain the IBFS compared to all other approaches, LCM and WOC-LCM. It is observed that the starting allocation of VAM and proposed MWOC-VAM are the same but different from both LCM and WOC-LCM. On the other hand, the starting allocation of LCM and proposed WOC-LCM are the same. Moreover, VAM and MWOC-VAM need fewer steps to get IBFS. It is also observed in the second column of Table XXVI that the pattern of the flow of allocation for each approach is different. Now we have considered another problem 6 (Example 6) in which one route has zero transportation cost and some route's transportation cost is less than 1 but getter than zero.

TABLE XXIV. MODIFIED WEIGHTED OPPORTUNITY COST INCLUDED TRANSPORTATION TABLEAU PROBLEM 5

	D_1	D_2	D_3	supply	DI
O_1	$\frac{150}{1}$ 1	$\frac{45}{16}$ 16	$\frac{30}{17}$ 17	10	15
O_2	120 0	$\frac{12}{6}$ 6	$\frac{12}{8}$ 8	2	6
O_3	$\frac{3}{3}$ 3	$\frac{9}{3}$ 3	$\frac{2}{7}$ 7	3	0
Demand	10	3	2		
DI	1	3	1		

TABLE XXV. COMPARISON REGARDING THE FLOW OF ALLOCATIONS AND TOTAL TRANSPORTATION COST TO FIND OUT THE IBFS OF THE PROBLEM 5

Method	**Flow of allocations and IBFS**	**Total Cost**
LCM	$x_{21} \rightarrow x_{11} \rightarrow x_{32} \rightarrow x_{13}$ 2 → 8 → 3 → 2	51
VAM	$x_{11} \rightarrow x_{32} \rightarrow x_{23}$ 10 → 3 → 2	**35**
WOC-LCM	$x_{21} \rightarrow x_{11} \rightarrow x_{32} \rightarrow x_{13}$ 2 → 8 → 3 → 2	51
MWOC-VAM (proposed)	$x_{11} \rightarrow x_{23} \rightarrow x_{32}$ 10 → 2 → 3	**35**

Example 6:

TABLE XXVI. TRANSPORTATION TABLEAU OF TRANSPORTATION PROBLEM 6

	D_1	D_2	D_3	Supply
O_1	0	3	0.5	8
O_2	3	7	10	3
O_3	1	0.7	11	9
Demand	6	6	8	

It is observed in the Table XXVI that c_{11} = 0, c_{13} = 0.5 and c_{32} = 0.7. So, to find out the weight cost factor corresponding to the cost entry 0, the algorithm executes the case (c) of the proposed method. Formally:

(c) As $c_{11} = 0$ and $\{c_{pq}: 0 < c_{ij} < 1, \forall p, q\} = \{0.5, 0.7\} \neq \varphi$, so

$$W_{c_{11}}^{m} = \frac{\max\{a_i, b_j; \forall i, j\}}{[\min\{c_{ij}: 0 < c_{ij} < 1, \forall i, j\}]} \times \min\{a_1, b_1\} \times \max\{D_1^{Ir}, D_1^{Ic}\}$$

$$= \frac{\max\{8,3,9, \quad 6,6,8\}}{[\min\{0.5,0.7\}]} \times \min\{8,6\} \times \max\{1, 0.5\}$$

$$= \frac{9}{0.5} \times 6 \times 1 = 108$$

We have again represented the cost matrix and the MWOC matrix in a single tableau as Table XXVII. It is observed in the Table XXVII that the weight opportunity cost of the cell C_{11} is 108 corresponding to the minimal cost i.e., c_{21} = 0 obtained by the case (c) of the proposed algorithm. Moreover, the weight opportunity cost of the cell C_{13} and C_{32} are 152 and 19.71 respectively which are calculated according to the case (a) as well. Now we have solved the problem by the proposed MWOC-VAM as well as LCM, VAM and WOC-LCM. The experimental result is shown in the Table XXVIII.

It is observed in Table XXVIII that the proposed MWOC-VAM and VAM need the least transportation cost to obtain the IBFS compared to the other two approaches, LCM and WOC-LCM. It is observed that the starting allocation of VAM and proposed MWOC-VAM are the same but different from both LCM and WOC-LCM. On the other hand, the starting allocation of LCM and proposed WOC-LCM are the same. It is also observed in the second column of Table XXVIII that the pattern of the flow of allocation for each approach is different.

To analyze the performance and effectiveness of the proposed method, we considered an additional 10 randomly generated numerical instances. The experimental results are displayed in Table XXIX. It is evident from Table XXIX that the proposed method consistently outperforms both the existing LCM and WOC-LCM approaches. Furthermore, in two instances, the proposed method surpasses VAM, while in other cases, it yields equivalent total costs compared to VAM. It is also observed that the IBFSs obtained by the proposed method are optimal or near optimal.

We collected an additional 8 numerical instances from published international journals/conferences to evaluate the efficiency and effectiveness of the proposed method. In Table XXX, the first column indicates the reference number of the published article. The data presented in Table XXX show that, except for three instances where all approaches obtained optimal solutions, the proposed method consistently outperforms both LCM and WOC-LCM. It is noteworthy that, in four out of eight instances, the proposed method outperforms VAM, while in the remaining instances, both approaches yield similar solutions.

The numerical experiments indicate that the proposed MWOC-VAM consistently performs as well as or better than both VAM and LCM. Furthermore, VAM requires the calculation of the DI at each iteration, which increases its computational cost. In contrast, the proposed MWOC-VAM only needs to compute the DI and WOC once initially, making it more computationally efficient.

TABLE XXVII. MODIFIED WEIGHTED OPPORTUNITY COST INCLUDED TRANSPORTATION TABLEAU PROBLEM 6

	D_1	D_2	D_3	supply	DI
O_1	108 0 ×	$\frac{13.8}{3}$ 3 ×	$\frac{76}{.5}$ .5 **8**	~~8~~	0.5
O_2	$\frac{12}{3}$ 3 **3**	$\frac{12}{7}$ 7 ×	$\frac{28.5}{10}$ 10 ×	~~3~~	4
O_3	$\frac{6}{1}$ 1 **3**	$\frac{13.8}{.7}$ .7 **6**	$\frac{76}{11}$ 11 ×	~~9~~	.3
Demand	~~6, 3~~	~~6~~	~~8~~		
DI	1	2.3	9.5		

TABLE XXVIII. COMPARISON REGARDING THE FLOW OF ALLOCATIONS AND TOTAL TRANSPORTATION COST TO FIND OUT THE IBFS OF THE PROBLEM 6

Method	**Flow of allocations and IBFS**	**Total Cost**
LCM	$x_{11} \rightarrow x_{13} \rightarrow x_{32} \rightarrow x_{23} \rightarrow x_{33}$ 6 → 2 → 6 → 3 → 3	68.2
VAM	$x_{13} \rightarrow x_{32} \rightarrow x_{31} \rightarrow x_{21}$ 8 → 6 → 3 → 3	**20.2**
WOC-LCM	$x_{11} \rightarrow x_{13} \rightarrow x_{32} \rightarrow x_{33} \rightarrow x_{23}$ 6 → 2 → 6 → 3 → 3	68.2
MWOC-VAM Proposed	$x_{13} \rightarrow x_{32} \rightarrow x_{21} \rightarrow x_{31}$ 8 → 6 → 3 → 3	**20.2**

TABLE XXIX. COMPARISON AMONG LCM, VAM, WOC-LCM, AND PROPOSED MWOC-VAM REGARDING THE IBFS OF SOME RANDOMLY GENERATED NUMERICAL INSTANCES

Ex. No.	**Problem**	**LCM**	**VAM**	**WOC-LCM**	**MWOC-VAM Proposed**	**Opt. Sol.**
1	C_{ij}:{(9,8,5,7); (4,6,8,7);(5,8,9,5)} S: (12,14,16); D: (8,18,13,3)	248	248	240	**241**	240
2	C_{ij}:{(4,3,5);(6,5,4);(8,10,7)} S: (9,8,10); D: (7,12,8)	145	150	145	**144**	**139**
3	C_{ij}:{(2,5,4);(6,1,2);(4,5,2)} S: (4,6,6); D: (3,7,6)	29	29	29	**29**	29
4	C_{ij}:{(14,19,7,5);(16,6,12,9);(6,16,5,20)} S: (10,12,18); D: (9,14,7,10)	243	243	243	**243**	243
5	C_{ij}:{(4,2,1);(3,8,4);(6,5,2)} S: (50,70,45); D: (40,65,60)	605	**490**	605	**490**	475
6	C_{ij}:{(21,16,23,13);(17,18,14,23); (32,27,18,41)} S: (11,13,19);D: (6,10,12,15)	922	**796**	919	**796**	**796**
7	C_{ij}:{(1,16,17);(0,6,8);(3,3,7)} S: (10,2,3); D: (10,3,2)	51	**35**	51	**35**	**35**
8	C_{ij}:{(1,16,17);(0,3,8);(3,3,7)} S: (10,2,3); D: (10,3,2)	60	**36**	60	**36**	**36**
9	C_{ij}:{(1,16,17);(0,6,8);(3,3,7)} S: (30,6,9) ;D: (30,9,6)	153	**105**	153	**105**	105

TABLE XXX. Comparison Among LCM, VAM, WOC-LCM, and Proposed MWOC-VAM Regarding the IBFS of the Published Numerical Instances

Ref No.	Problem	LCM	VAM	WOC-LCM	MWOC-VAM Proposed	Opt. Sol.
[3]	C_{ij}:{(10,2,20,11);(12,7,9,20);(4,14,16,18)} S: (15,25,10); D: (5,15,15,15)	475	475	475	**475**	435
[26]	C_{ij}:{(6,4,1); (3,8,7);(4,4,2)} S: (50,40,60); D: (20,95,35)	555	555	555	**555**	555
[60]	C_{ij}:{(7,5,9,11); (4,3,8,6);(3,8,10,5);(2,6,7,3)} S: (30,25,20,15); D: (30,30,20,10)	435	470	435	**430**	410
[61]	C_{ij}:{(4,3,5); (6,5,4);(8,10,7)} S: (90,80,100); D:(70,120,80)	1450	1500	1450	**1440**	1390
[26]	C_{ij}:{(4,1,2,4,4);(2,3,2,2,2);(3,5,2,4,4)} S: (60,35,40); D: (22,45,20,18,30)	305	273	278	**273**	273
[62]	C_{ij}:{(4,19,22,11); (1,9,14,14);(6,6,16,14)} S: (100,30,70); D: (40,20,60,80)	2090	2170	2160	**2090**	2040
[63]	C_{ij}:{(6,1,9,3); (11,5,2,8);(10,12,4,7)} S: (70,55,90); D: (85,35,50,45)	1165	1220	1165	**1165**	1160
[64]	C_{ij}:{(13,21,14); (8,12,21);(15,17,19)} S: (13,20,5); D: (12,15,11)	473	473	473	**473**	**465**

VI. Conclusion

IBFS is crucial for obtaining an optimal solution in TP. While various approaches exist in the literature to determine IBFS, most are formulated by manipulating the cost matrix to control allocation flow. In this article, we stand out as perhaps the first to consider the impact of node capacity distribution on the flow of allocations in both LCM and VAM. Through numerical experiments, we observed significant changes in output due to the distribution of capacity among nodes, even when the cost matrix and total supply and demand remained constant. For example, if the cost matrix is identical, the flow of allocations for approaches like NWC, LCM, VAM, etc., remains almost unchanged regardless of the distribution of capacity among nodes. However, by addressing this issue in the formulation of the flow allocation matrix in the proposed method, the flow of allocations varies significantly. To leverage this effect, we introduced a novel tool to control allocation flow. To incorporate the influence of node capacity distribution, we developed a capacity-influenced allocation control matrix, termed Capacity-Influenced Distribution Indicator (CI-DI), along with the distribution indicator defined by VAM. Subsequently, we proposed a capacity-influenced algorithm for finding IBFS in balanced TP. It is observed from the numerical experiments that the proposed method is effective to find out better IBFS of TPs. The proposed approach significantly overcomes the limitations of both LCM and VAM concerning the impact of capacity distribution among nodes. Additionally, it demonstrates enhanced computational efficiency compared to VAM. While VAM requires the calculation of the DI for each allocation step, the proposed method only needs to compute the Capacity-Influenced Distribution Indicator (CI-DI) matrix once. Experimental results lead to the conclusion that practitioners in the supply chain and transportation domain should not only consider cost distributions but also recognize the substantial role of capacity distributions among nodes in controlling allocation flow, leading to the identification of better IBFS. The concept of a capacity-influenced flow of allocation is innovative, providing a new perspective or "window" through which researchers can approach transportation problems and other linear programming challenges. In future work, we aim to develop a hybrid algorithm by integrating the proposed approach with fuzzy-based techniques.

Acknowledgment

This work was supported by the King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSP2025R18. Authors also acknowledge the support received from the Department of Computer Science, University of Maine, USA.

References

[1] B. Amaliah, C. Fatichah, and E. Suryani, “Total opportunity cost matrix–Minimal total: A new approach to determine initial basic feasible solution of a transportation problem”, Egyptian Informatics Journal., vol. 20, no. 2, pp. 131-141, 2019.

[2] B. Amaliah, C. Fatichah, and E. Suryani, “A Supply Selection Method for better Feasible Solution of balanced transportation problem”, Expert System with Applications., vol. 203, pp. 117399, oct. 2022

[3] B. Amaliah, C. Fatichah, and E. Suryani, “A new heuristic method of finding the initial basic feasible solution to solve the transportation problem,” Journal of King Saud University–Computer and Information science., vol. 34, no. 5, pp. 2298-2307, 2022.

[4] L. Aizemberg et al., “Formulations for a problem of petroleum transportation”. European Journal of Operational Research, vol. 237 no. 1, pp. 82-90, 2014.

[5] M. A. Babu et al., “Lowest allocation method (LAM): a new approach to obtain feasible solution of transportation model,” International Journal of Scientific and Engineering Research., vol. 4, no. 11, pp. 1344-1348, 2013.

[6] M. A. Babu, M. A. Hoque, and M. S. Uddin, “A heuristic for obtaining better initial feasible solution to the transportation problem,” Opsearch., vol. 57, pp. 221-245, 2020.

[7] A. P. Bhadane, and S. D. Manjarekar, “APB’s method for the IBFS of transportation problems and comparison with least cost method,” 2020.

[8] M. R. Bordón, J. M. Montagna, and G. Corsano, “Solution approaches for solving the log transportation problem,” Applied Mathematical Modelling., vol. 98, pp. 611-627, 2021.

[9] T. Can, and H. Koçak, “Tuncay Can’s Approximation Method to obtain initial basic feasible solution to transport problem,” Applied and Computational Mathematics., vol. 5, no. 2, pp. 78-82, 2016.

[10] S.K. Goyal, “Improving VAM for Unbalanced Transportation Problems,” J. Oper. Res. Soc., vol. 35, pp. 1113–1114, 1984.

[11] M. A. Hakim, and M. R. Kabir, “An Efficient Approach for Finding an Initial Basic Feasible Solution for Transportation Problems,” Progress in Nonlinear Dynamics and Chaos., vol. 5, no. 1, pp. 17-23, 2017.

[12] E. Hosseini, “Three new methods to find initial basic feasible solution of transportation problems,” Applied Mathematical Sciences., vol. 11, no. 37, pp.1803-1814, 2017.

[13] M. Hedid, and R. Zitouni, “Solving the four index fully fuzzy transportation problem,” Croatian Operational Research Review., vol. 11, no. 2, pp. 199-215, 2020.

[14] A. P. P. Htun, and K. T. Kyi, “Analysis of minimizing the transportation cost using least cost and vogel's approximation methods,” Doctoral dissertation, MERAL Portal, 2019.

[15] A. R. M. J. U. Jamali, F. Jannat and P. Akhtar, “Weighted cost opportunity based algorithm for initial basic feasible solution: A new approach in transportation problem,” Journal of Engineering Science., vol. 8, no. 1, pp. 63-70, 2017.

[16] A. R. M. J. U. Jamali, and P. Akhtar, “Find the IBFS of transportation problem by using sequentially updated weighted opportunity cost-based algorithm,” GANIT J. Bangladesh Math. Soc., vol. 38, pp. 47-55. 2018.

[17] A. R. M. J. U. Jamali, and R. R. Mondal, “Modified Dynamically-updated Weighted Opportunity Cost Based Algorithm for Unbalanced Transportation Problem,” Journal of Engineering Science., vol. 12, no. 2, pp. 119-131, 2021.

[18] A. R. M. J. U. Jamali, and M. T. Rahman, “Analysis of pitfalls of VAM for solving transportation problem,” A F Mujibur Rahman-Bangladesh Mathematical Society National Mathematics Conference-2022, 2023 PP. 185-186.

[19] A. R. M. J. U. Jamali, and M. T. Rahman, “Investigating the pitfalls of the least cost and Vogel's approximate methods: understanding the impact of cost matrix patterns,” Journal of Engineering., vol. 14, no. 1, pp.123-135, 2023.

[20] Z. A. M. S. Juman, and M. A. Hoque, “A heuristic solution technique to attain the minimal total cost bounds of transporting a homogeneous product with varying demands and supplies,” European journal of operational research., vol. 239, no. 1, pp. 146-156, 2014.

[21] Z. A. M. S. Juman, and M. A. Hoque, “An efficient heuristic to obtain a better initial feasible solution to the transportation problem,” Applied Soft Computing., vol. 34, pp. 813-826, 2015.

[22] O. Jude et al., “A new and efficient proposed approach to find initial basic feasible solution of a transportation problem,” American Journal of Applied Mathematics and Statistics., vol. 5, no. 2, pp. 54-61, 2017.

[23] F. S. Josephine, A. Saranya, and I. F. Nishandhi, “A dynamic method for solving intuitionistic fuzzy transportation problem,” European Journal of Molecular & Clinical Medicine., vol. 7, no. 11, pp. 5843-5854, 2020.

[24] S. Korukoğlu, and S. Ballı, “An improved Vogel's approximation method for the transportation problem,” Mathematical and Computational Applications., vol. 16, no. 2, pp. 370-381, 2011.

[25] K. Karagul, and Y. Sahin, “A novel approximation method to obtain initial basic feasible solution of transportation problem,” Journal of King Saud University-Engineering Sciences., vol. 32, no. 3, pp. 211-218, 2020.

[26] A. M. Khoso, A. A. Shaikh, and A. S. Qureshi, “Modified LCM’S Approximation Algorithm for Solving Transportation Problems,” Journal of Information Engineering and Applications., vol. 10, no. 3, pp. 7-15, 2020.

[27] G. Krishnaveni, and K. Ganesan, “An effective approach for the solution of fully fuzzy transportation problems,” In IOP Conference Series: Materials Science and Engineering, IOP Publishing, April 2021, Vol. 1130, No. 1, pp. 012065.

[28] R. Kumar, R. Gupta, and O. Karthiyayini, “A new approach to find the initial basic feasible solution of a transportation problem,” Int. J. Res. Granthaalayah., vol. 6, no. 5, pp. 321-325, 2018.

[29] R. R. Lekan, L. C. Kavi, and N. A. Neudauer, “Maximum Difference Extreme Difference Method for Finding the Initial Basic Feasible Solution of Transportation Problems,” Applications and Applied Mathematics: An International Journal (AAM)., vol. 16, no. 1, pp. 18, 2021.

[30] M. Mathirajan, and B. Meenakshi, “Experimental analysis of some variants of Vogel's approximation method,” Asia-Pacific Journal of Operational Research., vol. 21, no. 04, pp. 447-462, 2004.

[31] M. Mathirajan, S. Reddy, and M. V. Rani, “An experimental study of newly proposed initial basic feasible solution methods for a transportation problem,”Opsearch., vol. 59, no. 1, pp. 102-145, 2022.

[32] S. Muthukumar, R. Srinivasan, and V. Vijayan, “An optimal solution of unbalanced octagonal fuzzy transportation problem,” Materials Today: Proceedings., vol.37, pp. 1218-1220, 2021.

[33] J. Pratihar, et al, “Modified Vogel’s approximation method for transportation problem under uncertain environment,” Complex & intelligent systems., vol. 7, no. 1, pp. 29-40, 2021.

[34] A. K. M. S. Reza, A. R. M. J. U. Jamali, and B. Biswas, “A modified algorithm for solving unbalanced transportation problems,” Journal of Engineering., vol. 10, no. 1, pp. 93-101, 2019.

[35] N. M. Sharma, and A. P. Bhadane, “An alternative method to north-west corner method for solving transportation problem,” International Journal for Research in Engineering Application & Management, vol. 1, no. 12, pp. 1-3, 2016.

[36] P. Sumathi, and C. S. Bama, “A Tactical Strategy in Transportation Problems using Statistical Process,” Int. j. eng., vol. 7, no. 4, pp. 473-475, 2018.

[37] V. J. Sudhakar, N. Arunsankar, and T. Karpagam, “A new approach for finding an optimal solution for transportation problems,” European journal of scientific research., vol. 68, no. 2, pp. 254-257, 2012.

[38] A. S. S. G. A. Tularam, and G. M. Bhayo, “A comparative study of initial basic feasible solution methods for transportation problems,” 2014.

[39] M. W. Ullah, M. A. Uddin, and R. Kawser, “A modified Vogel’s approximation method for obtaining a good primal solution of transportation problems,” Annals of Pure and Applied Mathematics., vol. 11, no. 1, pp. 63-71, 2016.

[40] E. R. Wulan, et al, “The New Technique for Solving Transportation Problem,” 2020.

[41] B. Amaliah, C. Fatichah, and E. Suryani,. “Two Highest Penalties: A Modified Vogels Approximation Method to Find Initial Basic Feasible Solution of Transportation Problem,” In 2021 13th International Conference on Information & Communication Technology and System (ICTS), IEEE, October, 2021, pp. 318-323.

[42] Z. S. Mahdi, H. A. Wasi, and M. A. Shiker, “Solving transportation problems by using modification to Vogel’s approximation method,” In AIP Conference Proceedings, AIP Publishing, Vol. 2834, No. 1, December, 2023.

[43] C. S. Ramakrishnan, “An improvement to Goyal's modified VAM for the unbalanced transportation problem,” Journal of the Operational Research Society, vol. 39, no. 6, pp.609-610, 1988.

[44] N.Balakrishnan, “Modified Vogel's approximation method for the unbalanced transportation problem.,” Applied Mathematics Letters., vol. 3, no. 2, pp. 9-11, 1990.

[45] U. K. Das et al., “Logical development of Vogel’s approximation method (LD-VAM): an approach to find basic feasible solution of transportation problem,” International Journal of Scientific & Technology Research (IJSTR)., vol. 3, no. 2, pp. 42-48, 2014.

[46] S. M. A. K. Azad, M. B., Hossain, and M. M. Rahman, “An algorithmic approach to solve transportation problems with the average total opportunity cost method,” International Journal of Scientific and Research Publications., vol. 7, no. 2, pp. 266-270, 2017.

[47] H. Arsham, and A. B. Kahn, “A simplex-type algorithm for general transportation problems: an alternative to stepping-stone.” Journal of the Operational Research Society., vol. 40, pp. 581-590, 1989.

[48] V. N., Maurya et al., “Progressive Review and Analytical Approach for Optimal Solution of Stochastic Transportation Problems (STP)

Involving Multi-Choice Cost," American journal of modeling and optimization, vol. 2, no. 3, pp. 77-83, 2014.

[49] A. Akilbasha, G. Natarajan, and P.Pandian, "Optimising fully fuzzy interval integer transshipment problems," International Journal of Operational Research., vol. 46, no. 1, pp. 1-19, 2023.

[50] G. Xin et al., "A new approach for solving fuzzy transportation problem., In 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications, IEEE, June, 2014, pp. 37-39.

[51] A. Ebrahimnejad, "On solving transportation problems with triangular fuzzy numbers: Review with some extensions," In 2013 13th Iranian Conference on Fuzzy Systems (IFSC) IEEE, August, 2013, pp. 1-4.

[52] V. Tharakeswari, M. Kameswari, and M. Seenivasan, "A New Approach to the Transportation Problem of the Hexagonal Fuzzy Number," In 2023 Fifth International Conference on Electrical Computer and Communication Technologies (ICECCT), IEEE, February, 2023 pp. 1-5.

[53] M. Bisht, I. Beg, and R. Dangwal, "Optimal solution of pentagonal fuzzy transportation problem using a new ranking technique," Yugoslav Journal of Operations Research., 2023.

[54] M. Fegade, and A. Muley, "Optimal Solution to Transportation Problem with Heptagonal Fuzzy Numbers," European Journal of Mathematics and Statistics., vol. 3, no. 4, pp. 1-5, 2022.

[55] K. Kaewfak et al., "A risk analysis based on a two-stage model of fuzzy AHP-DEA for multimodal freight transportation systems," IEEE Access., vol. 8, pp. 153756-153773, 2020.

[56] G. Sharma et al., "Soft set based intelligent assistive model for multiobjective and multimodal transportation problem," IEEE Access, vol. 8, pp. 102646-102656, 2020.

[57] R. Gupta, and N.Gulati, "Survey of transportation problem," In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE, February, 2019, pp. 417-422.

[58] S. Moslem et al., "A Systematic Review of Analytic Hierarchy Process Applications to Solve Transportation Problems: From 2003 to 2019," IEEE Access, 2023.

[59] M. Sivri, I. Emiroglu, C. Guler, and F. Tasci, "A solution proposal to the transportation problem with Arsham and Khan (1989) the linear fractional objective function," In 2011 Fourth International Conference on Modeling, Simulation and Applied Optimization IEEE, April, 2011, pp. 1-9.

[60] S. Madamedon, E. S. Correa, and P. J. Lisboa, "Tiebreaker Vogel's Approximation Method, a Systematic Approach to improve the Initial Basic Feasible Solution of Transportation Problems," In 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE, May, 2022, pp. 211-216.

[61] M. M. Ahmed, et al. "A new approach to solve transportation problems," Open Journal of Optimization., vol. 5, no. 1, pp. 22-30, 2016.

[62] E. EMUSB, et al., "An effective alternative new approach in solving transportation problems," American Journal of Electrical and Computer Engineering., vol. 5, no. 1, pp. 1-8, 2021.

[63] M.I. Sharma. "A review paper on transportation problem for minimum transportation cost," IJESC., vol. 10, no. 1, ISSN 2321 3361, 2020.

[64] M. M. Ahmed, et al "New procedure of finding an initial basic feasible solution of the time minimizing transportation problems," Open Journal of Applied Sciences, vol. 5, no. 10, pp. 634-640, 2015.

RSS-LSTM: A Metaheuristic-Driven Optimization Approach for Efficient Text Classification

Muhammad Nasir[1], Noor Azah Samsudin[2], Shamsul Kamal Ahmad Khalid[3],
Souad Baowidan[4]*, Dr. Humaira Arshad[5], Wareesa Sharif [6]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn (UTHM), Johor Bahru, Malaysia[1, 2, 3]
Faculty of Computing and IT, King Abdulaziz University, Jeddah, Saudi Arabia[4]
Department of Computer Science, The Islamia University of Bahawalpur (The IUB), Bahawalpur, Pakistan[5]
Department of Artificial Intelligence, The Islamia University of Bahawalpur (The IUB), Bahawalpur, Pakistan[6]

***Abstract*—The digital data consumed by the average user daily is huge now and is increasing daily all over the world, which requires sophisticated methods to automatically process data, such as retrieving, searching, and formatting the data, particularly for classifying text data. Long Short-Term Memory (LSTM) is a prominent deep learning model for text classification. Several metaheuristic approaches, such as the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Firefly Algorithm (FF), have also been used to optimize Deep Learning (DL) models for classification. This study introduced an improved technique for text classification, called RSS-LSTM. The proposed technique optimized the hyperparameters and kernel function of LSTM through the Ringed Seal Search (RSS) algorithm to enhance simplification and learning ability. This work was also compared and evaluated against state-of-the-art techniques such as GA-LSTM, PSO-LSTM, and FF-LSTM. The results showed significantly better results using the proposed techniques, with an accuracy of 96%, recall of 96%, precision of 96%, and 95% f-measure on the Reuters-21578 dataset. In addition, it showed an accuracy of 77%, recall of 77%, precision of 78%, and f-measure of 76% on the 20 Newsgroups dataset, while it achieved accuracy, recall, precision, and f-measure of 91%, 91%, 94%, and 90%, respectively, using the AG News dataset.**

Keywords*—Deep learning; text classification; Long Short-Term Memory; Ringed Seal Search; metaheuristic algorithms; Part Swarm Optimization; Genetic Algorithm; Firefly Algorithm; hyperparameter optimization

I. INTRODUCTION

A large amount of textual data and different types of content are distributed to millions of people worldwide on the internet. The significant increase in the size of online data has attracted significant attention nowadays. Because of the large increase in textual data worldwide, the demand for text classification has also increased [1]. Hence, searching for a specific document within a large collection has become a difficult challenge. It examines people's emotions and distinguishes between customer comments on specific topics. Text classification (text categorization or tagging) involves assigning a text document to a set of predefined labels or classes using different machine learning and deep learning methods [2]. Different text classification techniques have been widely employed to categorize and organize content. It is important to gather and categorize documents automatically, according to their content. The primary objective of text classification is to divide unstructured documents into appropriate groups according to their content [3].

Machine Learning (ML) and Artificial Intelligence (AI) have gained significant prominence in recent years, emerging as highly discussed subjects. Numerous machine learning methods have achieved remarkable results in Natural Language Processing (NLP) [4]. However, conventional machine-learning-based text classification techniques have several drawbacks, including dimension explosion, data sparsity, and generalization capacity and selecting the optimal parameters for models. Most earlier research must consider the possibility that data could be misplaced or misconstrued following neural network computations [5]. The development of new machine learning techniques has yielded significant advancements in recent years, and deep learning has received more attention in the context of text categorization [6]. Deep learning techniques have been successful in the past few years, and there has been a significant increase in studies in this field, signifying that deep learning approaches have outperformed traditional machine-learning-based approaches in several text classification tasks, such as sentiment analysis, news categorization, question answering, and natural language inference [7].

Natural Language Processing has prioritized text classification for a long time. Currently, methodologies and techniques constitute integral components of numerous products and are deemed essential for a wide array of applications and devices. Many deep learning architectures, including LSTM Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and more recent transformers, have been applied to attain various state-of-the-art outcomes in NLP tasks [8]. Text classification is a major process in Natural Language Processing, and recent research has focused on deep learning-based neural network techniques that have shown promise. However, previous studies have often overlooked the potential loss or misinterpretation of information in neural network calculations. A study introduced LSTM-Com, a technique that leverages historical information, such as the original text and hidden layer outputs, to address these issues. LSTM-Com dynamically selects important historical information to compensate for the neural network, resulting in improved performance compared with the baseline in the classification experiments. A Long Short-Term Memory approach overcomes these challenges by performing text classification using historical data, including the original text data and output data

*Corresponding Author.

from the hidden layers [9]. Deep learning-based techniques are more beneficial for text classification compared to machine learning that uses traditional text classifiers, which have flaws such as data sparsity, dimension proliferation, and less generality; instead, the classifiers based on deep learning techniques can magnificently overcome these defects because they have strong learning ability with higher prediction accuracy; in contrast, they also avoid a cumbersome feature extraction process [10].

To build an LSTM model for classification, machine learning researchers manually configure certain parameters that are independent of the data. These parameters, such as the network structure and the training process of the LSTM, are known as hyperparameters. A key challenge is to find a set of hyperparameters that yield an accurate model within a reasonable timeframe, which is an integral part of the hyperparameter optimization problem [11]. The optimization of hyperparameters plays an important role in the performance of machine-learning algorithms. The importance of hyperparameter optimization is well recognized; however, there has been limited research to confirm its assumptions. Hyperparameter optimization is crucial for deep learning models because it directly affects their performance and generalization. Without optimal hyperparameters, a model may underfit or overfit, failing to capture the underlying patterns in the data. Tuning helps strike the right balance, maximizing the model's performance on unseen data [12]. Hyperparameter optimization is critical in machine learning. Machine-learning algorithms require the setting of hyperparameters before training the model. These values significantly impact the model’s performance, but finding good ones is complex, which has led machine learning researchers to look into automated methods for hyperparameter searches [13].

A previous study showed that a meta-heuristics-based algorithm has a significant effect on optimization. In addition, when an optimization problem's dimension (the number of routes or furnished) is increased, a study of the Grasshopper Optimization Algorithm (GOA) explains it [14]. The hyperparameters of a CNN significantly affect its performance and adjusting them manually is laborious and ineffective. The Spotted Hyena Optimization (SHO) is a high-level metaheuristic optimization algorithm with sophisticated exploration and exploitation capabilities. SHO produces a set of solutions in the form of hyperparameters that must be tuned. This procedure was repeated until the target optimal solution was attained [15].

Balancing exploration and exploitation play a central role in defining the effectiveness of an evolutionary algorithm. Optimal performance necessitates varying levels of the exploration-exploitation trade-off at different stages of evolution [16]. A study discussed the importance of balancing local exploitation and global exploration in metaheuristic algorithms and elaborated on the effectiveness of the bat algorithm in achieving this balance. Comparing the bat algorithm with the recurrent search approach, the study demonstrates that the bat algorithm is superior; it also explores the consequences of these findings for higher-dimensional optimization problems and applies the bat algorithm to business and engineering design optimization. A healthy search requires balancing exploration and exploitation [17]. A thorough survey is necessary because various techniques, datasets, and evaluation criteria have been proposed in the literature [18]. Firefly cannot offer a robust mechanism for achieving an ideal balance between exploration and exploitation; it cannot set the parameters strongly [19]. The RSS algorithm is a metaheuristic with two searching states (Brownian and Levy) that substitute randomly because of noise and balance exploitation and exploration of the search, the likelihood of finding local optima quickly is very low. Furthermore, RSS uses significantly fewer parameters than GA, PSO, and FF [20].

The performance of LSTM depends on parameter optimization, which is applied to text classification using various metaheuristic algorithms.

Following major contributions through in proposed approach:

- Pre-processing of data to make it efficient for further processing for feature engineering and deep learning model.
- An enhanced LSTM with RSS is proposed for textual data classification. Hyperparameters have a significant impact on the performance of deep learning models. With this measure of data, choosing the appropriate parameters (by optimizing the kernel function) for a neural network has become a huge exploration region in recent research.
- The proposed technique is compared with existing techniques, such as GA, PSO, and FF, using LSTM. To validate the performance of the proposed model for four measuring matrices: Accuracy, Precision, Recall, and F1-Score.
- Evaluate the existing technique and the proposed technique using three benchmark datasets: 20 Newsgroups, Reuters-21578, and AG News.

Section II explains literature, Section III explains the proposed approach, and Section IV explains the results of the experiments and Section V elaborates the conclusion of this study.

II. Literature Review

Text classification applications include spam filtering, contextual search, opinion mining, product review analysis, content management, and text sentiment mining. When tuning the hyperparameter automatically using an algorithm, it is called auto-tuning, which offers an effective way to automatically train the process of a model, although it provides more efficient results [21].

Several studies have described different machine learning and deep learning models, including pre-processing of text classification, related calculations, and test methods [22-24]. A minimalist and multi-propose-based text classification approach, uTC, was tested on 30 different datasets and showed the best accuracy compared with other state-of-the-art classification methods [23]. Another study suggested a method for determining sentiment review comparisons using three feature extraction techniques: Word2vec, Doc2vec, and TF-

IDF. It uses machine learning algorithms, such as SVM, Naive Bayes, and Decision Tree, with a grid search for optimization. The performance of these algorithms was assessed based on accuracy [24]. The advantages and disadvantages of related models are sorted, and the CNN model can capture the important content of the text. By contrast, the RNN model can analyze the context. The deep learning method is applied to text classification, which saves a large amount of workforce and material resources and improves the accuracy of text classification [25].

A sampling technique was proposed to solve the imbalanced class distribution for classifying tweets using Random Forest, Naive Bayes, and XGBoost [26]. In text classification tasks, the long short-term memory network and convolutional neural network models can both achieve high classification accuracy, and different deep learning models propose feature engineering using Term Frequency-inverse Document Frequency (TF-IDF) and also compare it with CNN, LSTM, and LSTM Attention for short- and long-text classification [27]. A hybrid model using CNN and LSTM was proposed for text classification, in which the features of text sentences were extracted using a multi-scale CNN, and the dependence of the text context was then captured using an LSTM model [28]. A supervised weighing scheme called the term frequency-inverse category frequency model proposed for text classification using deep learning was proposed for five different datasets this research covers to overcome the computational cost compared to other deep learning models [29].

Although resource-intensive, hyperparameter optimization is essential for machine learning. A novel technique called AgentHPO, which automates this procedure, examines task data on its own, experiments with different hyperparameters, and optimizes them based on past performance. Compared to conventional methods, this methodology streamlines the setup, lowers the number of trials required, and improves interpretability. Empirical tests reveal that AgentHPO frequently performs better than human trials and yields interpretable outcomes [30].

Multiple hyperparameters need to be set and tuned for the deep learning model's evaluation to predict the early onset of Parkinson's disease using hyperparameter optimization of the deep learning model [31]. The MLearn-ATC algorithm was compared with popular algorithms for classification, including Support Vector Machines, Probabilistic Neural Networks (PNN), K-Nearest Neighbor (KNN), and Naïve Bayes [32], to solve the text categorization issue, they proposed approach was examined using three distinct document datasets: Reuters-21578, 20 Newsgroups, and Real dataset [33].

Comparative investigations on the various feature selection and classification techniques used in sentiment analysis based on Natural Language Processing and contemporary techniques such as the Genetic Algorithm and rough set theory are evaluated. Another study examined the differences between sentiment analysis and standard feature-selection techniques for text categorization [34]. Several studies using supervised and unsupervised learning methods have been conducted to solve the issue of fake news identification. A study was conducted using the ISOT dataset to identify fake news. Long Short-Term Memory is applied in the developed model to distinguish between fake and authentic news, and hyperparameter tuning techniques, including grid search and random search, are proposed to adjust the model's hyper-parameters [35].

A role-based access control (RBAC) strategy is required to precisely identify access permissions to secure data. SQL queries created by authorized users have extremely similar features and are challenging to separate. A CNN-LSTM based on Part Swarm Optimization was proposed for hyperparameter optimization to detect attacks in SQL queries. Stock market uncertainty has a significant impact on many global financial and economic activities. Setting an investment plan or choosing the best time to trade depends on forecasting stock price movement, and a PSO-LSTM-based technique was proposed for stock price forecasting [36]. Convolution layer and Bidirectional LSTM (BiLSTM) with the attention mechanism proposed for text classification, particularly sentiment analysis. However, there is still an issue that LSTM cannot distinguish the different relevance between each part of the document [37].

A cross-entropy trained-based deep learning model called a bidirectional LSTM network is employed to perform text classification, utilizing both supervised and semi-supervised procedures, and an evaluation test using IMDB and AG News Group datasets [38]. The CNN-LSTM-based NC2LO Caledonian crow-optimization-based hybrid approach was applied for short-text classification. It was applied to IMDb, Tagmy News, Twitter, and AG News datasets, for developing tool modeling skills and attainment attracted, this Caledonian crow optimization model employs both social and asocial learning [39]. Another study reviewed metaheuristic optimization algorithms for power systems, focusing on their ability to solve complex optimization problems in environments with limited information and computational resources. It discusses six key challenges in power systems and evaluates the effectiveness of various metaheuristic algorithms in addressing them, evaluating their importance in promoting environmental sustainability and supporting renewable energy sources. The effectiveness of a metaheuristic algorithm is mostly determined by how well it balances globally diversified exploration and local intensive exploitation [40].

A hybrid technique using bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT)-based approach was proposed for text mining to understand Chinese railway incidents caused by electromagnetic interference. A text mining technique using TextBlob for sentiment score with TF-IDF vectorization and a Linear SVC classification model was proposed for text mining the Covid-19 vaccination Twitter dataset [41]. Table I gives a comprehensive review of related work.

Ship pilots must have a thorough understanding of the future positions of their ships and their target ship at a given time. However, there are now important problems that need to be resolved regarding forecast accuracy and computing efficiency. The deep, long, short-term memory network architecture and genetic algorithm were developed in this study to address these issues and predict the shipping route of inland water. The GA-LSTM model effectively increased the precision and speed of trajectory prediction [56]. Convolutional neural networks

(CNNs) have gained recognition for their promising performance in text categorization and sentiment analysis because they can preserve a document's 1D spatial orientation, where the order of words is crucial. Research has been conducted using genetic algorithms to automatically determine the ideal network architecture without the need for any intervention from experts [57]. Researchers use machine learning, and some use deep learning models to solve the classification issue, and Artificial Neural Networks (ANN), which are implemented with GA for text classification, although ANN also has the main problem of tuning hyperparameters [53].

A study investigated the impact of LSTM parameter optimization with meteoritic algorithms on text classification performance. The GA-LSTM automatically chooses settings to create the best gene subset. The original Cuckoo Search [58] optimization enters the local optimum because of the high dimensions of the early convergence of complicated issues. CS, PSO, and GA dominate global optimization algorithms in scientific and technological applications. These algorithms have limitations in the development of novel solutions to preserve the equilibrium between exploration and exploitation [59]. Choosing the best hyperparameter of a model has an immediate effect on the model's performance, and another study showed that the Bayesian Optimization [60] technique is a more viable technique for the performance of the K-Nearest Neighbor model than other models [61].

TABLE I. LITERATURE REVIEW

Year	Ref.	Technique	Limitations	Outcome(s)	Dataset
2024	[42]	CNN-Bi-LSTM	Improvements could involve incorporating additional variables	CNN-Bi-LSTM model adapts and achieves coefficients of determination, RMSE, and RMAE of 0.95, 37.94, and 5.27, respectively.	Gold prices dataset
2024	[43]	GS-CNN-LSTM	Probable overfitting of the model. Model Complexity. The test model generalizes well to unseen data	Hybrid model with grid search achieves 91.67% accuracy, 89.66% recall, 93.55% specificity, 92.86% precision, 91.23% f1-score, and 0.9310 AUC.	Heart disease Cleveland
2023	[44]	CNN-LSTM	Limited comparison. Hybridization models complexity	In this study hybrid model achieves an accuracy of 93.51%, outperforming traditional ML models in detecting PD using dynamic features.	PC-GITA disease
2023	[45]	LSTM-RNN-GRU	Economic indicators, humidity, and seasonal factors could also significantly impact electrical load that is not considered here.	In this study, Deep learning models, including LSTM, GRU, and RNN, are used for load forecasting. GRU model achieves the best performance.	Forecast electricity load in Palestine based on a novel real dataset
2022	[46]	LSTM-AE-TPE	Model complexity. Computational Cost.	The proposed model in this study achieves an R-square of over 0.9, indicating its effectiveness in indoor temperature prediction	Temperature dataset
2021	[47]	GA- Deep Long Short-Term Memory	Increased computation time. Increased complexity.	DLSTM model that achieves RMSE using Dynamic-Adam as of 0.026 and using Dynamic-Adamax 0.006	Power load dataset
2021	[48]	BO-PSO-RNN	Model complexity concerns. Limited comparison. The model has not yet been tested in high-dimensional spaces.	In this study RNN and LSTM models demonstrate their effectiveness compared to other methods like BO-L-BFGS-B and BO-TNC	Stock market price data. Oilfield production
2021	[49]	Hyperparameter Exploration LSTM-Predictor (HELP-LSTM)	Sequence length and number of fully connected layer units can impact performance. The HELP algorithm might experience collapse due to its extreme hyperparameter	This study utilizes probability-based exploration with LSTM-based prediction to improve hyperparameter exploration in neural network training.	MNIST
2021	[50]	GA-DNN	Computational cost. Kind of black-box mature. Model complexity	GA-based approach achieves 75.86% for RNNs and 41.12% for DNNs.	Sample streaming-data, Indian stock market, MNIST, CIFAR10
2020	[51]	PSO-LSTM, PSO-ANN	Limited country level dataset is used, and validation of dataset may also require	The PSO-LSTM model improves prediction accuracy and stability for water level forecasting compared to ANN. It enhances flood prediction at varying lead times, aiding future flood risk mitigation efforts in the study region.	Watersheds dataset used in this study
2020	[52]	GWO-LSTM	Address only global optima. Limited comparison.	The Grey Wolf Optimizer (GWO) algorithm to optimize the hyperparameters of Long Short-Term Memory models for language modeling tasks performed in this study.	Penn treebank dataset
2019	[53]	GA-LSTM	Computational cost. Performance issues for finding the global optimum	GA-LSTM optimization technique and achieve a maximum of 55% accuracy.	
2018	[54]	Differential Evolution-LSTM	DE algorithm took more processing time	LSTM hyperparameters improve emotion recognition. The proposed framework achieved 77.68% accuracy.	the dataset collected from wireless wearable sensors (Emotive and Expatica E4)
2018	[55]	CSA-LSTM	The hybrid model has not been evaluated on various datasets containing different text types.	Competitive search algorithm (CSA) used with LSTM and shows higher results.	Reuters-21578, RCV1-v2 and EUR-Lex

III. Proposed Approach

An improved text classification technique is proposed using Ringed Seal Search, Long Short-Term Memory and the model named RSS-LSTM, which demonstrated a considerable impact of LSTM hyperparameter optimization with RSS by optimization of the kernel. The RSS-LSTM technique proposed herein achieves a balanced approach between exploitation and exploration. The proposed research suggests a strong technique for hyperparameter optimization for text classification that produces more optimized results than the existing methods. This study explicitly compared using 20 Newsgroups, Reuters-21578, and AG News datasets for multiple labeled text classification. In this study, three datasets were taken from the Kaggle and UC Irvin machine learning repositories: 20 Newsgroups, Reuters-21578, and AG News. The datasets were used to evaluate the effectiveness of the RSS-LSTM. The experiment was programmed using the Jupyter Notebook for the proposed RSS-LSTM, at HP Xeone Workstation z440, 32gb RAM and 2.4 processor. The datasets used in this experiment were selected based on the extent to which they liked. There are two parts to the datasets; 30% of the data were chosen for testing, while 70% were used for training.

Fig. 1 shows the proposed model stages, which consist of three major stages: Data Pre-processing, Optimization of LSTM parameters, and performance measurement criteria. The details of the three stages are as follows.

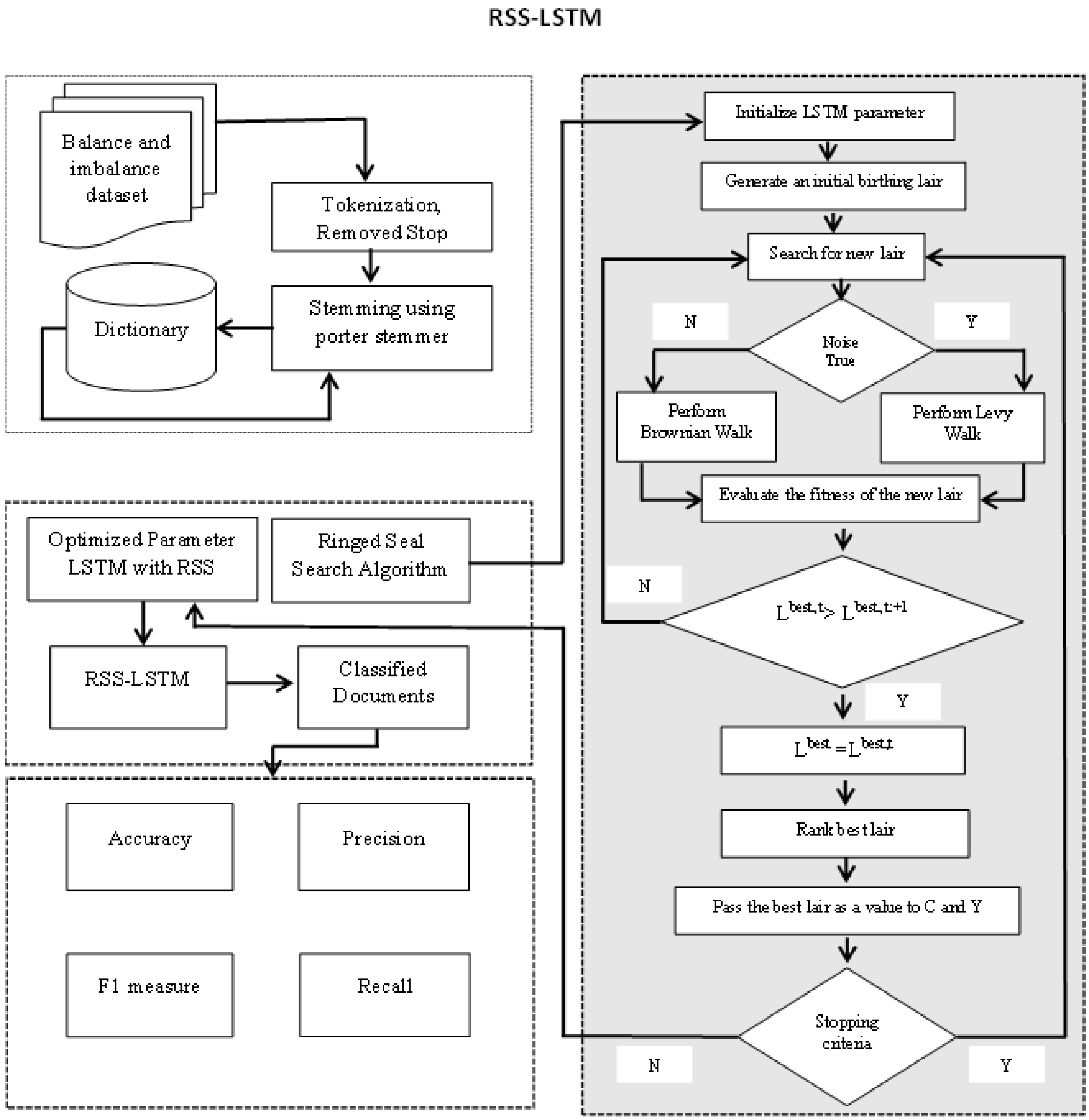

RSS-LSTM
Balance and imbalance dataset
Tokenization, Removed Stop
Dictionary
Stemming using porter stemmer
Optimized Parameter LSTM with RSS
Ringed Seal Search Algorithm
RSS-LSTM
Classified Documents
Accuracy
Precision
F1 measure
Recall
Initialize LSTM parameter
Generate an initial birthing lair
Search for new lair
N
Noise True
Y
Perform Brownian Walk
Perform Levy Walk
Evaluate the fitness of the new lair
N
$L^{best,t}> L^{best,t+1}$
Y
$L^{best} = L^{best,t}$
Rank best lair
Pass the best lair as a value to C and Y
Stopping criteria
N
Y

Fig. 1. Proposed model RSS-LSTM.

A. *Stage 01*

Pre-Processing: In this stage 01 textual dataset, such as 20 Newsgroups, Reuters-21578, and AG News Group, are pre-processed and used for further experiments. Below list of preprocessing steps performed at this stage:

- Tokenization of datasets and feature extraction.
- Removing spaces and punctuation.
- Removing unnecessary words to proceed with meaningful words.
- Removing emojis and stop words.
- Porter stemmer to remove inflection.

B. *Stage 02*

Optimization of LSTM Parameters: An enhanced method using RSS based on LSTM was implemented in this stage. The performance of the proposed RSS-LSTM method was measured and compared with existing methods, such as GA-LSTM, PSO-LSTM, and FF-LSTM. The proposed RSS-LSTM showed improved results compared to existing techniques.

C. *Stage 03*

Performance Measuring: In this stage, the performance of the proposed model is evaluated, which comprises four measuring criteria: accuracy, recall, precession, and F1-score. The results section of the proposed comparison model explains these measurement criteria and their results. The proposed models focus on optimizing deep learning techniques, such as LSTM, to achieve optimal results. Fig. 1 shows the proposed model. Algorithm 1 shows the Pseudo-code of proposed RSS-LSTM technique.

Algorithm 1: Pseudo-code of RSS-LSTM

```
Start …
1.  Set the initial parameter of LSTM
2.  Producing starting number of lairs $L_1 = (f = 1,2,3,\ldots,n)$
3.  While (Stopping measure)
4.  If noise = false
5.  Search the nearness for a new layer with a Brownian walk
6.  Else
7.  Expend the search for a process for a new layer by using
    levy walk
8.  End if
9.  Evaluating the fitness of each new lair and comparison with
    the previous lair
10. If
11.   $L^{\text{best,k}} > L^{\text{best,k+1}}$
12. Select the new lair
13.   $L^{\text{best}} > L^{\text{best,k}}$
14. Else
15.   Go at 4
16. End if
17. Rank the solutions.
18. Return the best lair of execution
19. The global finest lair is fed LSTM classifier for training
20. Training the classifier (LSTM)
21. End while
22. End
```

D. *Proposed Approach RSS-LSTM Explanation*

A metaheuristic technique called RSS is suggested to address optimization issues. To escape predators, the RSS method relies on the foundation of seal pup behavior to find the best lair. This technique divides the search space into two states: ordinary, routine, and urgent or fast. Under ordinary and urgent states, an intensive and extensive search is performed to find good-quality air and move in it to escape predators [62].

If it identifies the location of the exploration space where ω = 1 (ω= 0 indicate the normal state), ∂ is notified that **Ω** contains β, a predator that is moving and making noise pointed as ω. A state (**Ω**, ρ) for an E event, where **Ω** is referred to as an urgent state if **Ω** comprises β and ∂ members of the event in the exploration space that has noise ω. Let A be an event and the search space be (**Ω**, β, ∂, ρ).

If the search space is now ρ is ω in a state where ω= 0 (shows the outside noise), then ∂ is assumed to be not informed **Ω**, contains β, and (**Ω**, ρ) represents a normal condition, executes a Levy-walk for an urgent ∂ situation and Brownian-walk for a typical normal condition ∂. The proposed RSS-LSTM is discussed in the next section.

Text Classification: The proposed research developed an enhanced technique using Ringed Seal Search and Long Short-Term Memory. The Hyperparameters of LSTM are optimized using the RSS one of the metaheuristic techniques, which overcomes the efficiency of the other techniques PSO-LSTM, GA-LSTM, and FF-LSTM algorithms. The selection of parameters for the classification problem presents one of the primary obstacles to the optimal LSTM. It is a frequent practice to improve LSTMs by utilizing metaheuristic search algorithms that are inspired by nature to achieve better classification outcomes. A brief description of the LSTM classifier is provided below:

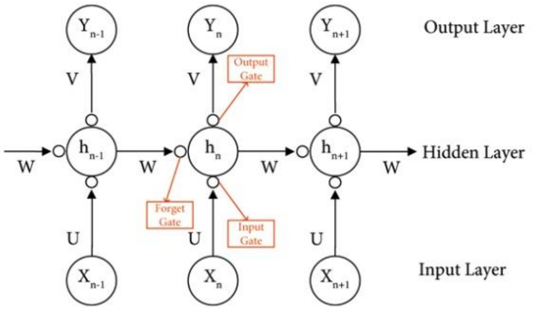


Fig. 2. LSTM Classifier.

As shown Fig. 2 describes the LSTM classifier basic architecture, the input and output data are represented by Xn and Yn, respectively, the weight coefficients are represented by U, V, and W, he is the hidden layer status, and hn is related to the current input Xn input and the previous R hidden layers.

$$h_1 = UX_n + W_{n-1}h_{n-1} + W_{n-2}h_{n-2} + \cdots + W_{n-R}h_{n-R} \quad (1)$$

Finding a hyperplane that appropriately divides the training dataset into two groups and optimizes the LSTM is the main goal. The LSTM classification problem is combined as follows:

$$f_t = \sigma_g(W_f x_t + U_{\mathfrak{f}} h_{t\,-1} + b_f)$$

$$\iota_t = \sigma_g\ (W_i x_t\ +\ U_i\ h_{i-1} + b_i)$$

$$O_t = \sigma_g\ (W_o a_t\ +\ U_o h_{t\,-1} + b_o) \tag{2}$$

$$c_t\ =\ f_t\ \circ c_{t-1}\ + \iota_t \circ \sigma_c\ (W_c x_t\ +\ U_c h_{t-1} + b_c)$$

$$h_t\ = \sigma_t \circ \sigma_h\ (c_t)$$

Hyperparameters of the error terms were represented by C > 0. The normal vector and offset of the separating hyperplane are the variables denoted by the letters w and b in Eq. (2), respectively. The LSTM parameters were optimized using the Ringed Seal Search. The RSS uses how seal pups search to find the best-hidden place to avoid predators. The proposed RSS algorithm presents a sensitive search paradigm inspired by the movement of the seals. Seal pups are always relocated to high-quality lairs. These lairs offer both thermal shelter against low air temperatures and strong wind chills, as well as shielding themselves from predators, similar to bears. A complex lair can be placed in one place for a seal owing to its close movement. When the seal pup moves throughout its multi-chamber cave and searches for a new one, a sequence of events can be recounted. Evolution was accomplished by altering a random value. The starting population of LSTM parameters is represented by a matrix, the chosen parameters are placed in a vector form, and the vector has evolved to find the optimal combinations of parameters in each iteration.

Inspired by nature, when addressing an optimization problem, the RSS always begins with initial values that can be used as the initial state. The first answer is represented by a vector of values (L_i,i =1, 2,3..k_n.) during the optimization process. The RSS algorithm always begins with a multi-chambered initial number of birthing lairs n. Puppies enter a search area to locate new, better lairs to hide themselves. The formation of an array from these starting values in the search space is required to locate a better search space. Eq. (3) and (4) define the number of lairs in the RSS algorithm that corresponds to the lairs for seal pups. Most metaheuristic algorithms start with the initial population, which can be named as starting values or initial values, because to solve an optimization problem, it is necessary to start with some initial values. Eq 3 represents the initial lair

$$L_i, i = 1,2,3,\ldots,n \tag{3}$$

There are chambers m in every lair, arranged at random. Each L contained m chambers. As an example, consider an array of L = [I multiply m] that represents the lair i of a habitat that is now in use,

$$L = [i \times m] \tag{4}$$

The values range from a predetermined bottom bound, L_bj, to the upper limit, U_bj, randomly and consistently in the search space, as described in Eq. (5).

$$L_i = L_b + (U_b +).rand\ \left(size(L_b)\right) \tag{5}$$

$$i = 1,2,3,\ldots,n \tag{6}$$

where i is the lair number, and n is the number of initialized pairs. The seal travels from one lair to another in a particular search pattern, producing new solutions (new lairs) x t+1 for seal *i*. A new lair is located in the Eq. (7)

$$\chi_i^{t+1}\ = \chi_i^{t+1} + \alpha \times \Delta x \tag{7}$$

where a indicates the size of the step in urgent or normal states.

$$\Delta x =\ {}^{\lambda}levy \quad \text{where}\, w\ =\ 1 \tag{8}$$

where ω represents the uniform discrete distribution shown in Eq. (8), (ω = 1 denotes the external noise). For the Levy walk, the random walk is typified by a step size that is determined using an inverse power-law tail probability distribution, as shown in Eq. (9).

$$Levy \sim u\, \overline{t^{-\lambda}} \tag{9}$$

where t is the length and 1< λ < 3. When λ is less than or equal to 3, there is no heavy tail in the distribution, and the sum of all the lengths approaches a distribution.

Anomalous diffusion, in which the mean squared displacement increases linearly with time, characterizes a Levy walk. The Brownian walk, in contrast to the Levy walk, is typified by normal diffusion, where the mean-squared displacement increases linearly.

Eq. (10) illustrates the structure of the Brownian walk search for a new chamber inside a multichambered lair structure.

$$equal\Delta x =\ {}^{\lambda}brownian \quad \text{where}\, \omega\ =\ 0 \tag{10}$$

The search is characterized by the step size described in Eq. (11).

$$S = K\ \times\ rand(d,\ Ndot) \tag{11}$$

K represents the standard deviation of the regular distribution of the diffusion rate coefficient, d denotes the dimensions of the problem, and N dots symbolize the quantity of Brownian particles within the search space.

The proposed RSS-LSTM approach responds to variations in hyperparameters through its ringed seal search optimization process. This process iteratively explores the hyperparameter space to find the optimal set of hyperparameters that minimizes the objective function, which in this case is the performance of the LSTM model in classification. The RSS algorithm perform a balanced exploration that tries new hyperparameters and exploitation that exploits known good hyperparameters to efficiently search the hyperparameter space for optimal solution.

The Brownian Walk function of the algorithm generates a random walk for a specified step size. Given a current value x in the range of the lower and upper bounds [lower_bound, upper_bound], from the uniform distribution, it adds a random value walk [step_size, step_size] to x. Then, the function checks if the new value [new_value] is within the bounds [lower_bound, upper_bound]; if it exceeds the upper bound then it returns the upper_bound, or if it falls below the lower bound then it returns the lower_bound. Otherwise, it returns to a new value, this function is mathematically represented as:

$$\text{new\_value} = \text{walk} + \text{lower\_bound}$$

where,

$$\text{walk} \sim \text{Uniform}(-\text{step\_size}, \text{step\_size}) \tag{13}$$

The final value is then paired to ensure that it remains within the specified bounds.

Using Pareto distribution, the Levy walk function generates a random walk. First, it samples a value r from a Pareto distribution with the shape parameter beta. It then samples an angle u from the uniform distribution [0, 2/pi] and computes a walk value using r×cos(u). Similar to the Brownian walk, the function ensures that the new value remains within the specified bounds [lower-bound, upper-bound]. Mathematically it can be represented as

$$\text{r} \sim \text{Pareto}(\beta)$$

$$\text{u} \sim \text{Uniform}(0{,}2\pi)$$

$$\text{walk} = \text{r} \times \cos(\text{u})$$

$$\text{new\_value} = \text{walk} + \text{lower\_bound} \tag{14}$$

E. Performance Criteria

Accuracy, precision, recall, and f-measure are all important factors to consider when evaluating the efficacy of a classification model in classifying text. Precision is calculated using Eq. (15), and recall is specified in Eq. (16). Eq. (17) and (18) display the Accuracy and F-measure, respectively.

$$Precision\ = \frac{t_p}{t_p + f_p} \tag{15}$$

In the Eq. (15) t_p denotes the true positive rate and f_p shows the false positive rate in precision.

$$Recall\ = \frac{t_p}{t_p + f_n} \tag{16}$$

where t_p describes the true positive rate and f_n denotes the false negative rate in the recall.

Accuracy is defined as the ratio of the number of correctly classified objects to the total number of objects. Inaccuracy and true positive (t_p), true negative (t_n), false negative (f_n) and false positive (f_p) values are calculated as in Eq. (17):

$$Accuracy = \frac{t_p + t_n}{t_p + \ f_p + t_n + \ f_n} \tag{17}$$

The F-measure is the harmonic mean in which, precision and recall are combined, and the traditional f-measure is calculated as in Eq. (18):

$$F - measure\ = \ 2 \times \frac{p \times r}{p + r} \tag{18}$$

where p denotes the precision and r is the recall in the F-measure.

IV. Results

Different experiments were conducted to analyse the data, and the performance of RSS-LSTM was compared with different metaheuristic algorithms associated with LSTM such as GA-LSTM, PSO-LSTM, and FF-LSTM., for three datasets: 20 Newsgroups, Reuters-21578, and AG News were used to test the performance using several measurement parameters, including accuracy, F-measure, precision, and recall. The proposed model was tested using different sets of classes with different iterations.

A. Reuters-21578 Dataset Results

As shown in Table II, it is demonstrating that the proposed technique using RSS-LSTM performs more effectively than the other techniques. For Reuters-21578, RSS-LSTM's accuracy is superior to that of GA-LSTM, PSO-LSTM, and FF-LSTM. Compared to earlier methods, RSS-LSTM significantly outperformed the other methods on the entire dataset. On the Reuters-21578 text dataset, RSS-LSTM produced an accuracy of 96%, GA-LSTM produced 78%, firefly produced 56%, and PSO-LSTM produced 87% as shown in Fig. 3. Compared to the F-measure, RSS-LSTM outperformed GA-LSTM, PSO-LSTM and FF-LSTM, and achieved 95%, 64%, 86% and 54% respectively as shown in Fig. 4. Fig. 5 shows that RSS-LSTM delivered improved results in terms of precision compared to the existing techniques, RSS-LSTM achieved 96%, whereas GA-LSTM, PSO-LSTM and RR-LSTM achieved 72%, 86% and 71% respectively. In the recall scenario, Fig. 6 shows that RSS-LSTM again performs better than GA-LSTM, PSO-LSTM, and FF-LSTM and achieved results as of 96%, 68%, 87% and 56% respectively. The accuracy of the proposed method was superior to that of the other mentioned techniques, as demonstrated in Table II, for the entire dataset. When compared to the GA-LSTM, PSO-LSTM, and FF-LSTM algorithms, the RSS-LSTM approach performed better in terms of accuracy, precision, recall, and f-measure.

TABLE II. Performance of RSS-LSTM among PSO-LSTM and GA-LSTM using Reuters-21578 Dataset

Classifier	Measure criteria			
	Accuracy	F-measure	Precision	Recall
GA-LSTM	0.78	0.64	0.72	0.68
PSO-LSTM	0.87	0.86	0.86	0.87
FF-LSTM	0.56	0.54	0.71	0.56
RSS-LSTM	**0.96**	**0.95**	**0.96**	**0.96**

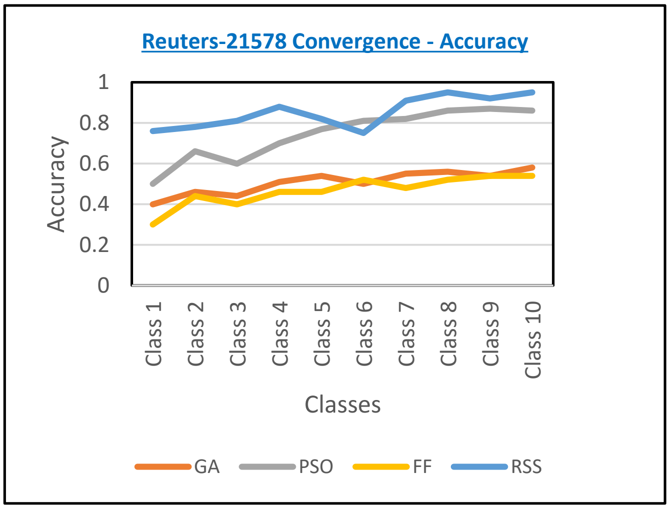


Fig. 3. Reuters-21578 convergence - accuracy.

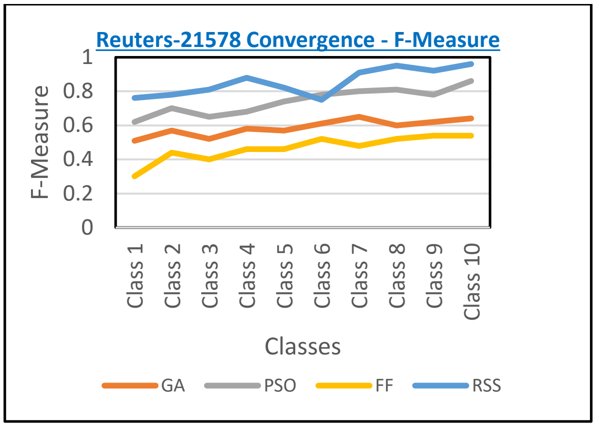


Fig. 4. Reuters-21578 convergence - f-measure.

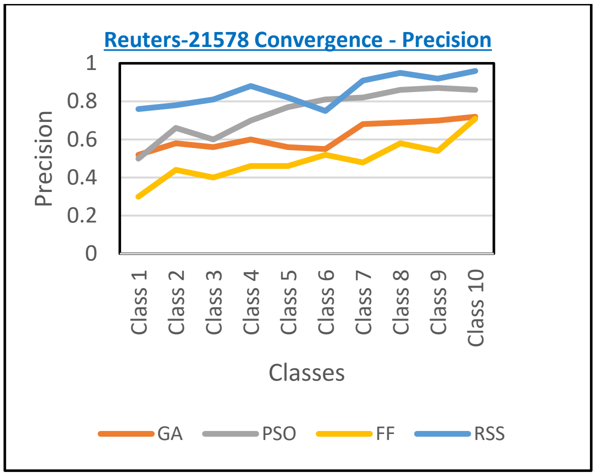


Fig. 5. Reuters-21578 - precision.

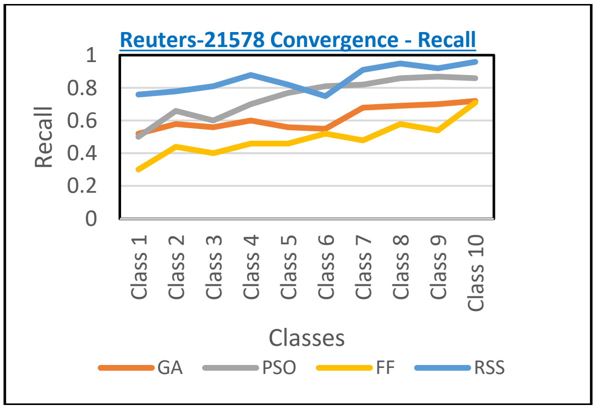


Fig. 6. Reuters-21578 convergence - recall.

B. *Result of 20 Newsgroups Dataset*

An experiment using 20 Newsgroups text datasets demonstrated that the proposed RSS-LSTM technique outperformed the GA-LSTM, PSO-LSTM, and FF-LSTM strategies already in use. Fig. 7 shows that the RSS-LSTM generated high accuracy of 77%. While GA-LSTM produced 49%, PSO-LSTM and FF-LSTM provided results of 58% and 21%, respectively as shown in Fig. 7 and Table III. In addition, the F-measure was 77% for RSS-LSTM, compared to 41%, 56%, and 35% for GA-LSTM, PSO-LSTM, and FF-LSTM, respectively as shown in Fig. 8. The RSS-LSTM technique provided better results as of 78% than the existing techniques for 20 Newsgroups dataset in precision comparison as shown in Fig. 9, while GA-LSTM, PSO-LSTM, and FF-LSTM achieved precisions as 50%, 73%, and 21% respectively. Similarly, Fig. 10 shows that RSS-LSTM achieved a higher recall as of 77% in comparison to GA-LSTM, PSO-LSTM and FF-LSTM. RSS-LSTM generated the best accuracy of 77%. The outcome for RSS-LSTM is superior to that of the GA-LSTM, PSO-LSTM, and FF-LSTM methodologies, as shown in Table III and described below.

TABLE III. PERFORMANCES OF RSS-LSTM RESULTS AMONG GA-LSTM, PSO-LSTM, AND FF-LSTM USING THE 20-NEWSGROUP DATASET

Classifier	**Measure criteria**			
	Accuracy	**F-measure**	**Precision**	**Recall**
GA-LSTM	0.49	0.41	0.50	0.49
PSO-LSTM	0.58	0.56	0.73	0.58
FF-LSTM	0.21	0.35	0.21	0.30
RSS-LSTM	**0.77**	**0.77**	**0.78**	**0.77**

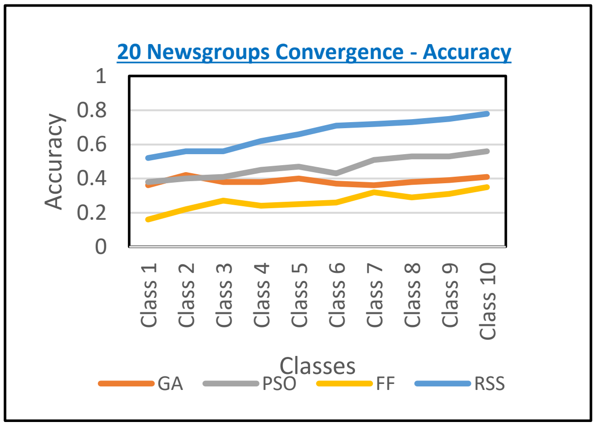


Fig. 7. 20 Newsgroups convergence – accuracy.

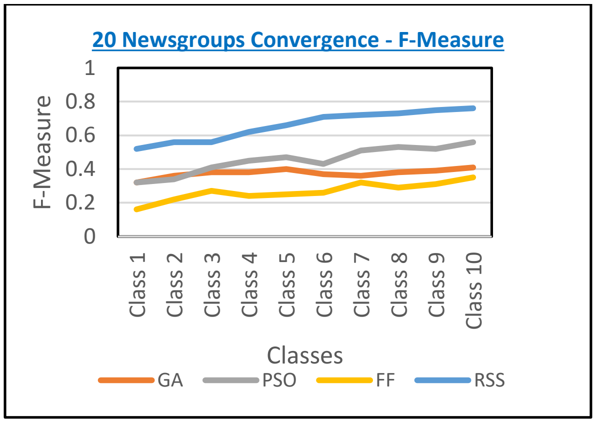


Fig. 8. 20 Newsgroups convergence - f-measure.

C. *Result of AG News Dataset*

Table IV shows the performance measured using the AG News dataset. The performance of the RSS-LSTM optimization approach was tested against those of existing GA-LSTM, PSO-LSTM, and FF-LSTM techniques. The study was carried out for the evaluation matrix, as accuracy, F-measure, precision, and recall are among the metrics used to evaluate RSS-LSTM. Compared to existing GA-LSTM, PSO-LSTM, and FF-LSTM techniques, the RSS-LSTM technique produced greater accuracy than other comparing techniques. Fig. 11 shows that GA-LSTM achieved 86% accuracy, PSO-LSTM and FF-LSTM produced 88% and 80% accuracy, respectively, while the proposed RSS-LSTM produced 91% accuracy.

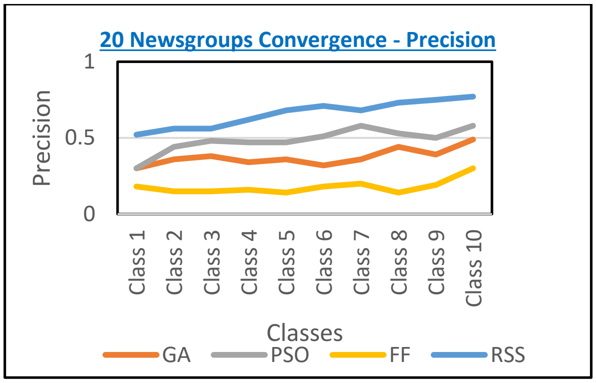


Fig. 9. 20 Newsgroups convergence – precision.

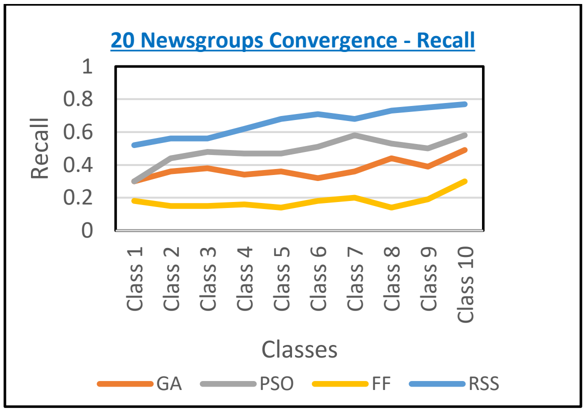


Fig. 10. 20 Newsgroups convergence – recall.

TABLE IV. PERFORMANCES RSS-LSTM RESULT AMONG GA-LSTM, PSO-LSTM, AND FIREFLY-LSTM USING THE AG NEWS DATASET

Classifier	**Measuring Criteria**			
	Accuracy	**F-measure**	**Precision**	**Recall**
GA-LSTM	0.86	0.85	0.86	0.86
PSO-LSTM	0.88	0.84	0.89	0.89
FF-LSTM	0.80	0.78	0.88	0.88
RSS-LSTM	**0.91**	**0.90**	**0.94**	**0.91**

To evaluate the F-measure score, significant results were obtained for RSS-LSTM, GA-LSTM, PSO-LSTM, and FF-LSTM, which were 91 %, 85%, 84%, and for FF-LSTM, 78% shown in Fig. 12. Additionally, RSS-LSTM achieved higher precision and Fig. 13 demonstrated 94% precision compared to 86%, 89%, and 88% precision for GA-LSTM, PSO-LSTM, and FF-LSTM, respectively. The outcome of RSS-LSTM was also measured for recall, and it provided a result of 91%, compared to 86%, 89%, and 88% for GA-LSTM, PSO-LSTM, and FF-LSTM, respectively also shown in Fig. 14. The table below shows that the overall performance of RSS-LSTM is superior to that of the GA-LSTM, PSO-LSTM, and FF-LSTM techniques.

Table V presents the combined results of the proposed RSS-LSTM model compared with GA-LSTM, PSO-LSTM, and GA-LSTM. Additionally, Fig. 15, 16, and 17 provide graphical representations of the results achieved by the proposed model and the compared techniques.

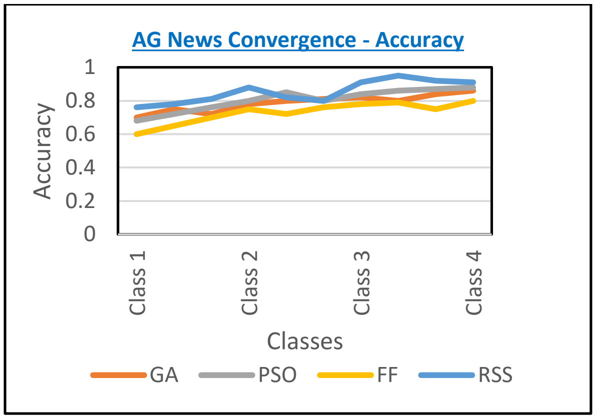


Fig. 11. AG News convergence – accuracy.

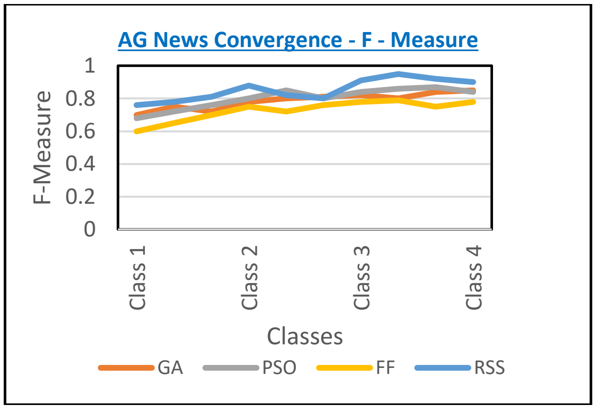


Fig. 12. AG News convergence - f – measure.

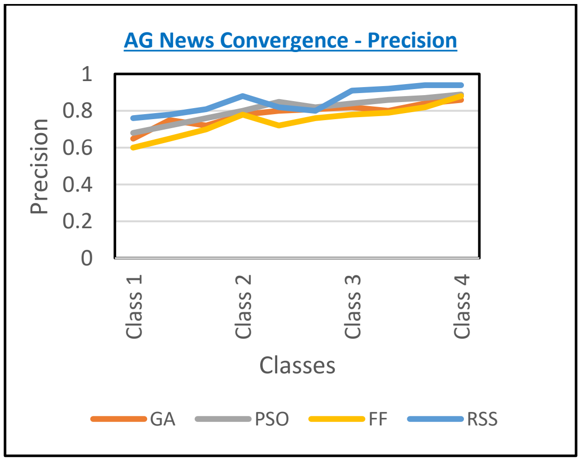


Fig. 13. AG News convergence - precision.

TABLE V. PERFORMANCES RSS-LSTM WITH RESPECT TO THREE DATASETS

Dataset	Technique	Measuring Matrix			
		Accuracy	**F-Measure**	**Precision**	**Recall**
Reuters-21578	GA-LSTM	0.78	0.64	0.72	0.68
	PSO-LSTM	0.87	0.86	0.86	0.87
	FF-LSTM	0.56	0.54	0.71	0.56
	RSS-LSTM	**0.96**	**0.95**	**0.96**	**0.96**
20 Newsgroups	GA-LSTM	0.49	0.41	0.50	0.49
	PSO-LSTM	0.58	0.56	0.73	0.58
	FF-LSTM	0.21	0.35	0.21	0.30
	RSS-LSTM	**0.77**	**0.77**	**0.78**	**0.77**
AG News	GA-LSTM	0.86	0.85	0.86	0.86
	PSO-LSTM	0.88	0.84	0.89	0.89
	FF-LSTM	0.80	0.78	0.88	0.88
	RSS-LSTM	**0.91**	**0.90**	**0.94**	**0.91**

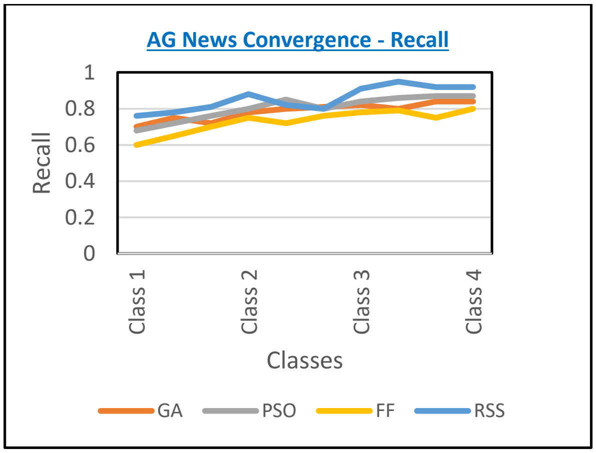


Fig. 14. AG News convergence – recall.

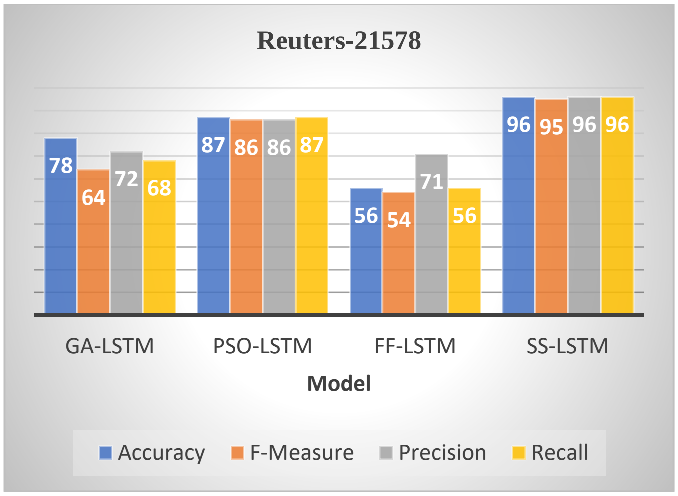


Fig. 15. Comparison with reuters-21578.

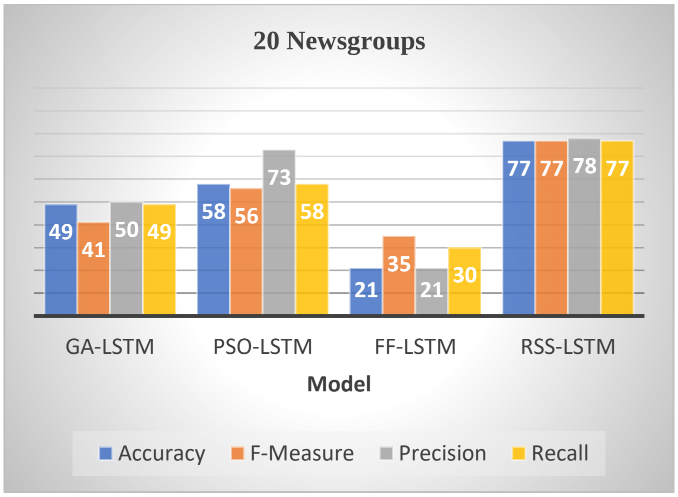


Fig. 16. Comparison with 20 newsgroups.

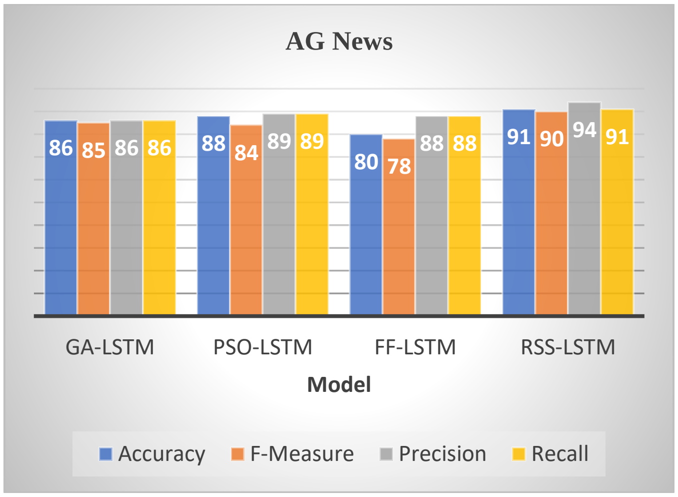


Fig. 17. Comparison with AG news.

D. Comparison of Models

The proposed RSS-LSTM model was also compared with state-of-the-art deep learning models, such as Support Vector Machine, Stochastics Gradient Descent (SGD), Random Forest (RF), Logistics Regression (LR), K-nearest Neighbour (KNN), Naïve Base (NB), Decision Tree (DT), Autor encoder (AE), AdaBoost (AB) using Reuters-21578, 20 Newsgroups and AG news dataset, where Proposed approached shows the significant results compare to mentioned techniques. Table VI describes the comparison of proposed technique with state of art models used for hyperparameter optimization using different datasets.

Above mentioned Table VII lists the hyperparameters with their ranges that we use in this study.

TABLE VI. Performances RSS-LSTM According with respect to Three Datasets

Ref	Year	Technique	Dataset	Findings
[63]	2023	Support Vector Machine, Stochastic Gradient Descent (SGD), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN)	Reuters-21578	Accuracy achieves using Reuters-21578 dataset as 0.8516, 0.8476, 0.8470, 0.8110, 0.8183, 0.8135
[64]	2022	OPT 175b, Bloom 176B, OPT 30b, OPT 1.3b	AG News	AG news achieve the accuracy as follows using different techniques 68.7, 39.5, 60.7, 37.6
[65]	2020	Naïve Base, SVM, Gradient, Boosting, Random Forest, Logistics Regression	20NewsGroups	Achieve maximum accuracy 67.3, 65.3,59.5,60.1 and 67.4 respectively
[66]	2020	Logistic Regression (LR), Decision Trees (DT), Support Vector Machine, AdaBoost (AB), Random Forest (RF), Multinomial Naïve Bayes (MNB), Multilayer Perceptron (MLP), Gradient Boosting (GB).	20NewsGroups	68.28, 44.44, 70.03, 45.61, 62.28 60.62, 60.12, 69.46
[67]	2017	Autoencoder	20NewsGroups	Achieve accuracy 73.78
Proposed Technique		**RSS-LSTM**	**Reuters, 20Newsgroups, AG News**	**Achieve maximum accuracy of 96%, 77%, and 91% respectively**

TABLE VII. Hyperparameter Ranges

Model	Hyperparameter	Optimization Range
GA-LSTM	Dense units	16 to128
	Learning rate	0.001 to 0.1
	Dropout rate	0.1 to 0.05
PSO-LSTM	Dense units	16 to 128
	Learning rate	0.001 to 0.1
	Dropout rate	0.1 to 0.05
FF-LSTM	Dense units	16 to 128
	Learning rate	0.001 to 0.1
	Dropout rate	0.1 to 0.05
	Dense units	16 to 128
RSS-LSTM	Dense units	16 to 128
	Learning rate	0.001 to 0.1
	Dropout rate	0.1 to 0.05
	Dense units	16 to 128

TABLE VIII. Time Constraints of Hyperparameter Optimization

Algorithm	Complexity	Dataset	Time (std dev)
PSO	O(I×(P×L+P))	Reut-21578	132.49
		20NG	71.23
		AG News	448.02
GA	O(G×P×L)	Reut-21578	220.28
		20NG	119.23
		AG News	500.64
FFA	O(num iterations ×num_fireflies^2 × len(firefly-bounds))	Reut-21578	279.25
		20NG	248.31
		AG News	2253.90
RSS	O(max-iterations × num-lairs × (len(search-space) + max-len))	Reut-21578	110.05
		20NG	67.23
		AG News	836.20

Table VIII describes the complexity equivalences, for PSO I, P, and L indicating the total number of iterations, number of particles in a swarm, and search space (dimensioned of parameters) respectively. For the GA algorithm, G indicates the generation (iteration), P resents the Population size and L resents the number of genes in an individual. The Firefly Algorithm contains a total number of iterations, the number of flies, and the length of parameters for a search space. Ringed Seal Search (RSS) consists of its number of iterations, the number of lairs that describe the search areas, the search space that defines the number of dimensions, and at last, it adds a maximum number of lairs. Additionally, Table VIII shows the standard deviation of temporal demands for each algorithm according to the dataset, low standard deviation values indicate more consistent performance, and they shed light on how variable or consistent the algorithm's execution times are across various datasets, however, these values may also vary depending at the factors such as algorithm’s number of iterations, number of epochs, batch size and other parameters.

Table IX demonstrates the variation of hyperparameters for different datasets.

TABLE IX. VARIATION OF HYPERPARAMETERS USING DIFFERENT DATASETS

Dataset	Technique	Dense Unit	Dropout	Learning Rate
Reuters-21578	GA-LSTM	118	0.3914	0.0087
	PSO-LSTM	16	0.2271	0.0364
	FF-LSTM	37	0.2221	0.0031
	RSS-LSTM	27	0.3654	0.01
20 Newsgroups	GA-LSTM	89	0.1527	0.0942
	PSO-LSTM	120	0.3515	0.0090
	FF-LSTM	34	0.2316	0.0093
	RSS-LSTM	54	0.2681	0.01
AG News	GA-LSTM	61	0.1635	0.0150
	PSO-LSTM	91	0.3527	0.0014
	FF-LSTM	104	0.3998	0.0087
	RSS-LSTM	122	0.15739	0.0069

V. DISCUSSION AND CONCLUSION

In diverse fields such as bioinformatics, sentiment analysis, online handwritten recognition, and text classification, LSTM is used to apply diverse classification issues. One area where academics are attempting to increase classification accuracy is text classification. Different experiments were conducted to analyse the data, and the performance of RSS-LSTM was compared with that of GA-LSTM, PSO-LSTM, and FF-LSTM. Three datasets, including 20 Newsgroups, Reuters-21578, and AG News, were used to test the performance using several measurement parameters, including accuracy, F-measure, precision, and recall. The proposed model was tested using different sets of classes.

The results presented in Fig. 3, 4, 5, and 6 demonstrate that the proposed approach RSS-LSTM outperforms existing methods, achieving 96% accuracy, 96% F-score, 95% precision, and 96% recall on the Reuters-21578 dataset. Similarly, as shown in Fig. 7, 8, 9, and 10, the proposed method RSS-LSTM outperforms existing approaches on the 20 News dataset, achieving 77% accuracy, 77% F-score, 78% precision, and 77% recall. Furthermore, Fig. 11, 12, 13, and 14 indicate that the proposed approach surpasses existing methods on the AG News dataset, with 91% accuracy, 90% F-score, 94% precision, and 91% recall.

According to the literature review, search methods affect LSTM performance when solving text classification optimization problems. Therefore, to improve the LSTM parameters for enhanced text classification accuracy, this research presented an enhanced technique called RSS-LSTM, conducted using datasets Reuter-21578, 20 Newsgroups, and AG News Dataset, which was used to evaluate the effectiveness of the proposed model. The simulation results demonstrated that in terms of Accuracy, F-measure, Precision, and Recall, the proposed RSS-LSTM surpasses existing techniques. The experimental results on different classes of these three datasets showed that the proposed model performed well in terms of term precision, F-value, precision, and recall. The proposed model also compares with LSTM addresses different types of text classification problems in various fields such as bioinformatics, opinion mining, handwriting, and online recognition. One of the areas where scholastics are endeavouring to increase characterization accuracy is text classification.

A. Future Work

To evaluate the performance of the proposed model with different hyperparameter and ranges. To evaluate the proposed models at images datasets. To assess the effect of different iterations of different algorithms. The proposed technique performs very well as per the given measuring matrix, however detailed temporal demands may also be required in future work using different iteration and parameter settings evaluate the proposed technique with other deep learning models such as Recurrent Neural Network (RNN), Feedforward Neural Network (FNN), Gated Recurrent Unit (GRU) Autoencoders (AE).

REFERENCES

[1] Hassan, S.U., J. Ahamed, and K. Ahmad, Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, 2022. 3: p. 238-248.

[2] Roul, R.K., S.R. Asthana, and G. Kumar, Study on suitability and importance of multilayer extreme learning machine for classification of text data. Soft Computing, 2017. 21(15): p. 4239-4256.

[3] Janani, R. and S. Vijayarani, Automatic text classification using machine learning and optimization algorithms. 2021. 25(2 %J Soft Comput.): p. 1129–1145.

[4] Kowsari, K., et al., Text Classification Algorithms: A Survey. 2019. 10(4): p. 150.

[5] Wanjale, K., et al. Analyzing Machine Learning Algorithm for Breast Cancer Diagnosis. in International Conference on Computer Vision and Robotics. 2023. Springer.

[6] Liu, C. and X. Wang, Quality-related English text classification based on recurrent neural network. Journal of Visual Communication and Image Representation, 2020. 71: p. 102724.

[7] Minaee, S., et al., Deep learning--based text classification: a comprehensive review. 2021. 54(3): p. 1-40.

[8] Joshi, R., P. Goel, and R. Joshi. Deep learning for hindi text classification: A comparison. in Intelligent Human Computer Interaction: 11th International Conference, IHCI 2019, Allahabad, India, December 12–14, 2019, Proceedings 11. 2020. Springer.

[9] Huang, W., et al., LSTM with compensation method for text classification. 2021. 20(2): p. 159-167.

[10] Cai, J., et al. Deeplearning model used in text classification. in 2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP). 2018. IEEE.

[11] Wistuba, M., N. Schilling, and L. Schmidt-Thieme. Hyperparameter Optimization Machines. in 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016.

[12] Shankar, K., et al., Hyperparameter Tuning Deep Learning for Diabetic Retinopathy Fundus Image Classification. IEEE Access, 2020. 8: p. 118164-118173.

[13] Claesen, M. and B.J.a.p.a. De Moor, Hyperparameter search in machine learning. 2015.

[14] Meraihi, Y., et al., Grasshopper Optimization Algorithm: Theory, Variants, and Applications. IEEE Access, 2021. 9: p. 50001-50024.

[15] Fatyanosa, T.N. and M. Aritsugi. Effects of the Number of Hyperparameters on the Performance of GA-CNN. in 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT). 2020. IEEE.

[16] Sun, J., et al. Balancing exploration and exploitation in multiobjective evolutionary optimization. in Proceedings of the genetic and evolutionary computation conference companion. 2018.

[17] Yang, X.-S., S. Deb, and S.J.a.p.a. Fong, Bat algorithm is better than intermittent search strategy. 2014.

[18] Qian, L., et al., A new method of inland water ship trajectory prediction based on long short-term memory network optimized by genetic algorithm. 2022. 12(8): p. 4073.

[19] Hall, M., et al., The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009. 11(1): p. 10-18.

[20] Sharif, W., et al., An Optimised Support Vector Machine with Ringed Seal Search Algorithm for Efficient Text Classification. Journal of Engineering Science and Technology, 2019. 14: p. 1601-1613.

[21] Wang, Z., et al. Automatic hyperparameter tuning of machine learning models under time constraints. in 2018 IEEE international conference on big data (Big Data). 2018. IEEE.

[22] Elgeldawi, E., et al. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. in Informatics. 2021. MDPI.

[23] Tellez, E.S., et al., An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems, 2018. 149: p. 110-123.

[24] Isa, S.M., et al., Optimizing the hyperparameter of feature extraction and machine learning classification algorithms. 2019. 10(3): p. 69-76.

[25] Wu, H., et al., Review of Text Classification Methods on Deep Learning. 2020. 63(3).

[26] Duan, H.K., et al., Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. 2023. 48: p. 100600.

[27] Zhou, H. Research of text classification based on TF-IDF and CNN-LSTM. in Journal of Physics: Conference Series. 2022. IOP Publishing.

[28] Zhai, Z., et al., Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model. 2023: p. 1-14.

[29] Attieh, J. and J.J.K.-B.S. Tekli, Supervised term-category feature weighting for improved text classification. 2023. 261: p. 110215.

[30] Liu, S., C. Gao, and Y.J.a.p.a. Li, Large Language Model Agent for Hyper-Parameter Optimization. 2024.

[31] Kaur, S., et al., Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. 2020. 31: p. 1-15.

[32] Springenberg, J., et al., Striving for Simplicity: The All Convolutional Net. 2014.

[33] Janani, R. and S.J.S.C. Vijayarani, Automatic text classification using machine learning and optimization algorithms. 2021. 25: p. 1129-1145.

[34] Ahmad, S.R., A.A. Bakar, and M.R. Yaakub. Metaheuristic algorithms for feature selection in sentiment analysis. in 2015 Science and Information Conference (SAI). 2015. IEEE.

[35] Vyas, P., J. Liu, and O. El-Gayar, Fake news detection on the web: An LSTM-based approach. 2021.

[36] Ji, Y., A.W.-C. Liew, and L.J.I.A. Yang, A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast. 2021. 9: p. 23660-23671.

[37] Liu, G. and J.J.N. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification. 2019. 337: p. 325-338.

[38] Singh Sachan, D., M. Zaheer, and R.J.a.e.-p. Salakhutdinov, Revisiting LSTM Networks for Semi-Supervised Text Classification via Mixed Objective Function. 2020: p. arXiv: 2009.04007.

[39] Sendhilkumar, S.J.E.S.w.A., Developing a conceptual framework for short text categorization using hybrid CNN-LSTM based Caledonian crow optimization. 2023. 212: p. 118517.

[40] Nassef, A.M., et al., Review of Metaheuristic Optimization Algorithms for Power Systems Problems. 2023. 15(12): p. 9434.

[41] Qorib, M., et al., Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. 2023. 212: p. 118715.

[42] Amini, A. and R. Kalantari, Gold price prediction by a CNN-Bi-LSTM model along with automatic parameter tuning. 2024. 19(3): p. e0298426.

[43] Maulani, A.A., et al., Comparison of Hyperparameter Optimization Techniques in Hybrid CNN-LSTM Model for Heart Disease Classification. Sinkron : jurnal dan penelitian teknik informatika, 2024. 8(1): p. 455-465.

[44] Lilhore, U.K., et al., Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease. Scientific Reports, 2023. 13(1): p. 14605.

[45] Abumohsen, M., A.Y. Owda, and M. Owda, Electrical Load Forecasting Using LSTM, GRU, and RNN Algorithms. Energies, 2023. 16(5): p. 2283.

[46] Ben, J., et al., Attention-LSTM architecture combined with Bayesian hyperparameter optimization for indoor temperature prediction. Building and Environment, 2022. 224: p. 109536.

[47] Bakhashwain, N. and A. Sagheer, Online Tuning of Hyperparameters in Deep LSTM for Time Series Applications. International Journal of Intelligent Engineering & Systems, 2021. 14(1).

[48] Li, Y., Y. Zhang, and Y. Cai, A New Hyper-Parameter Optimization Method for Power Load Forecast Based on Recurrent Neural Networks. Algorithms, 2021. 14(6): p. 163.

[49] Li, W., et al., HELP: An LSTM-based Approach to Hyperparameter Exploration in Neural Network Learning. Neurocomputing, 2021. 442.

[50] Kumar, P., S. Batra, and B. Raman, Deep neural network hyper-parameter tuning through twofold genetic approach. Soft Computing, 2021. 25(13): p. 8747-8771.

[51] Andonie, R. and A.-C. Florea, Weighted random search for CNN hyperparameter optimization. arXiv preprint arXiv:2003.13300, 2020.

[52] Aufa, B.Z., S. Suyanto, and A. Arifianto. Hyperparameter Setting of LSTM-based Language Model using Grey Wolf Optimizer. in 2020 International Conference on Data Science and Its Applications (ICoDSA). 2020.

[53] Gorgolis, N., et al. Hyperparameter optimization of LSTM network models through genetic algorithm. in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). 2019. IEEE.

[54] Nakisa, B., et al., Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access, 2018. 6: p. 49325-49338.

[55] Loussaief, S. and A. Abdelkrim, Convolutional neural network hyper-parameters optimization based on genetic algorithms. International Journal of Advanced Computer Science and Applications, 2018. 9(10).

[56] Qian, L., et al., A New Method of Inland Water Ship Trajectory Prediction Based on Long Short-Term Memory Network Optimized by Genetic Algorithm. 2022. 12(8): p. 4073.

[57] Andersen, H., et al. Evolving neural networks for text classification using genetic algorithm-based approaches. in 2021 IEEE Congress on Evolutionary Computation (CEC). 2021. IEEE.

[58] Razno, M.J.C.L. and I. Systems, Machine learning text classification model with NLP approach. 2019. 2: p. 71-73.

[59] Mohapatra, P., et al., An improved cuckoo search based extreme learning machine for medical data classification. 2015. 24: p. 25-49.

[60] Chacra, D.A. and J. Zelek. Road Segmentation in Street View Images Using Texture Information. in 2016 13th Conference on Computer and Robot Vision (CRV). 2016.

[61] Aghaabbasi, M., et al., On hyperparameter optimization of machine learning methods using a Bayesian optimization algorithm to predict work travel mode choice. 2023. 11: p. 19762-19774.

[62] Saadi, Y., et al., Ringed seal search for global optimization via a sensitive search model. 2016. 11(1): p. e0144371.

[63] Daud, S., et al., Topic Classification of Online News Articles Using Optimized Machine Learning Models. 2023. 12(1): p. 16.

[64] Gonen, H., et al., Demystifying prompts in language models via perplexity estimation. 2022.

[65] Singh, A.K., <Text-classification-performance-analysis-using-dif-ferent-supervised-machine-learning-models.pdf>. International Journal of Scientific & Engineering ResearchISSN 2229-5518, September-2020. Volume 11(Issue 9).

[66] Alyasiri, O., et al., Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review. IEEE Access, 2022. 10: p. 39833-39852.

[67] Aziguli, W., et al., A Robust Text Classifier Based on Denoising Deep Neural Network in the Analysis of Big Data. Scientific Programming, 2017. 2017: p. 1-10.

Construction of Image Retrieval Module for Cultural and Creative Products Based on DF-CNN

Meng Jiang
School of Art and Design, Henan University of Engineering, Zhengzhou, 451191, China

***Abstract*—With the growth of the cultural and creative product industry, more and more cultural and creative products have been designed and published in different channels. A method based on image retrieval module is proposed to address the search problem of Chinese creative products in online channels. During the process, a cascaded forest is proposed to achieve layer by layer processing, with class vectors as the main transfer content in the entire forest system. An image attribute feature extraction process that introduces extreme gradient enhancement is designed, and the aggregation of multi-scale and multi-region features is utilized to improve image retrieval performance. The experimental results showed that in the similarity test of extracting image feature information when the image contained three composite cultural and creative objects and the total pixel amount of the image reached 7M, the similarity of image feature information was 97.6%. In the analysis of running time, the research method only took 7.4ms to generate search results in seven fields. In the analysis of the proportion of false search content, the research method maintained a false search proportion of within 6.0% when searching for a single cultural and creative product object. This indicates that the research method has higher accuracy and efficiency in image retrieval of cultural and creative products. Research methods can provide certain technical support for the development of the cultural and creative industry.**

Keywords—Image retrieval; class vector; extreme gradient enhancement; Chinese creative products; layer by layer processing

I. Introduction

In today's rapidly developing era of digitization and informatization, the cultural and creative industry has become an important force driving economic growth and cultural inheritance. Cultural and creative products (CCPs), with their unique design and cultural connotations, not only satisfy people's pursuit of a better life but also become an important medium for spreading culture and shaping brand image [1]. However, with the increasing variety of CCPs, efficiently retrieving and identifying these products has become a problem that needs to be solved [2]. Image retrieval is an important retrieval method, and some scholars have conducted relevant research on image retrieval technology. Zhang et al. proposed a research method based on generative adversarial learning for image retrieval in text writing. The process enhanced retrieval efficiency by enhancing the semantics of generated images, and the resolution of generated images was determined by local discriminators. The results indicated that the designed method was efficient and accurate for image retrieval [3]. Humenberger et al. proposed a visual localization-based research method for image retrieval in autonomous driving. The method prioritized the judgment of image types by setting scene road conditions and then used fuzzy matching in dynamic scenes for image localization. The results indicated that the proposed method had good fitness for image retrieval in dynamic scenes [4]. Salih et al. proposed a research method based on double-layer feature judgment for multimedia image retrieval problems. The method filtered out images with deviant types by setting a coarse search layer, and then locked in the required images through subdivision class retrieval. The outcomes indicated that the designed method had high accuracy for image retrieval on multimedia [5]. Wang et al. proposed a research method based on spatial and exchange domains for texture image retrieval. The process used dual-tree complex wavelet transform to decompose the image and re-model it, achieving a complementary global retrieval structure. The outcomes indicated that the designed method improved the retrieval efficiency in the database [6]. Li et al. proposed a research algorithm based on a hash method for image retrieval on the Internet. The algorithm retrieved data patterns by collecting image data in a stationary environment, and simulating image transformations in non-stationary environments to adapt to its retrieval environment. The results indicated that the proposed method was efficient for image retrieval in large datasets [7].

Traditional image retrieval techniques, although achieving certain results in certain scenarios, often fail to meet the high demands of users for retrieval accuracy and efficiency. With the advancement of computer technology, deep learning technology has emerged and made revolutionary progress in the field of image recognition and retrieval [8]. Convolutional neural network (CNN) has become an important technology in the field of image processing due to its powerful feature extraction ability and ability to learn complex patterns. Some scholars have conducted relevant research on CNN. Xin's team proposed a CNN-based feature extraction method for the diagnosis of ocean turbine attachments. This method consists of three steps: data preprocessing, feature extraction, and fault diagnosis. The results indicated that the method could work smoothly in harsh environments [9]. Kumar et al. put forward a training algorithm using Dolphin-SCA to address the compression problem of CNN models applied in tumor classification. During the process, a fuzzy deformable fusion model was used for image segmentation. The experiment findings showed that the proposed method had good diagnostic accuracy [10]. Thirusangu et al. proposed a deep CNN based on the U-Net architecture to address the issue of low resolution in transcranial ultrasound images. During the process, filters were used for preprocessing and the effects of other architectures were compared. The results indicated that the U-Net architecture had better accuracy in identifying SN [11]. Ma proposed a method combining CNN to address the design issue of facial feature tracking systems. During the process, heterogeneous convolution was used to reduce the parameters of the convolution kernel. The search box mechanism was inserted into the network acceptance domain adjustment module, and the attention dispersal mechanism was used to standardize the arrangement of heterogeneous convolutions. The experiment outcomes indicated that the proposed method had good tracking accuracy and interpretability [12]. Ashtiani et al. proposed a method combining CNN technology to address the issue of medical image classification. During the process, the light waves incident on the pixel array were directly processed to obtain image classification results, and linear calculations were performed optically to reduce time consumption. The experiment outcomes indicated that the proposed method had good classification accuracy [13].

In summary, there have been many techniques using CNN for image retrieval and processing, but existing image retrieval techniques often struggle to accurately identify and match complex design elements and cultural connotations in CCPs, resulting in low accuracy of retrieval results. CCP images often contain rich color and texture information, and existing technologies may not be able to fully explore and utilize the useful information in the images when processing such high-dimensional data. Moreover, existing research has mostly focused on specific types of images, lacking the generalization ability for image retrieval in the special field of CCPs, which limits the universality of the technology. A single CNN model may have biases in extracting global and local features, and cannot fully express the complexity and diversity of CCPs. Moreover, a single CNN model usually requires a lot of computing resources. For large-scale image retrieval tasks, a single model may be difficult to meet the real-time requirements [14]. Deep forest (DF) is an ensemble learning method based on decision trees, which inherits the advantages of decision trees in processing high-dimensional data and improves the model's generalization ability and robustness by integrating multiple decision trees. The problems that the research attempts to solve and the gaps that will be achieved with other studies include: (1) developing an image retrieval technology that can accurately identify the design elements and cultural connotations of CCP images, to improve the relevance of retrieval results and make the research method more adaptable to CCP images than other methods. (2) Design an efficient image retrieval algorithm to meet the retrieval needs of large-scale image data and achieve rapid response. (3) Research on an image retrieval model with good generalization ability, which can adapt to CCP images from different sources and styles, and enable the research method to intelligently adapt to the constantly changing content of CCP images. In this context, the study attempts to innovatively combine DF and CNN, and constructs a cascaded forest structure. By combining feature vectors and similarity calculation, a new image retrieval technology for CCPs is designed. It is expected to provide certain technical references for the CCP industry.

The research is mainly conducted in four sections. Section II part is the design of image attribute feature extraction technology based on DF algorithm and CCP image retrieval method combined with CNN. Section III is to analyze the effectiveness of the research method through performance testing and application analysis. Section IV discusses and summarizes the entire text.

II. Methods and Materials

A. *Design of Image Attribute Feature Extraction Technology Based on DF Algorithm*

When designing, searching, and purchasing CCPs, a large amount of retrieval is required, among which image retrieval is a commonly used retrieval method [15]. However, CCPs often have rich designs and diverse styles, which may make it difficult to accurately match and recognize images during image retrieval. The DF algorithm can handle high-dimensional data and complex problems during computation, and can extract complex and diverse attribute features of images [16]. Research uses DF algorithm to design image attribute feature extraction technology for CCP image retrieval technology. In response to the complexity of information contained in CCPs, a cascaded forest is proposed to implement layer by layer processing of DF networks, and its structure is constructed as shown in Fig. 1.

In Fig. 1, each layer in the cascaded forest receives the feature information from the previous layer and the original input feature information, and each layer is an integrated structure of a decision tree forest. To obtain better diversity, a study is conducted to construct a composite forest structure using random forests and completely random forests. The construction of each decision tree begins with a randomly selected set of features. At each node of the tree, the system randomly selects a feature to segment the dataset. The ultimate goal is to grow each leaf node to be pure, meaning that the samples in the leaf nodes belong to the same category. The number of decision trees in the forest is a hyperparameter that can be changed according to different tasks of image retrieval for CCPs. In the entire forest system, class vectors are studied as the main transfer content, and the generation method is shown in Fig. 2.

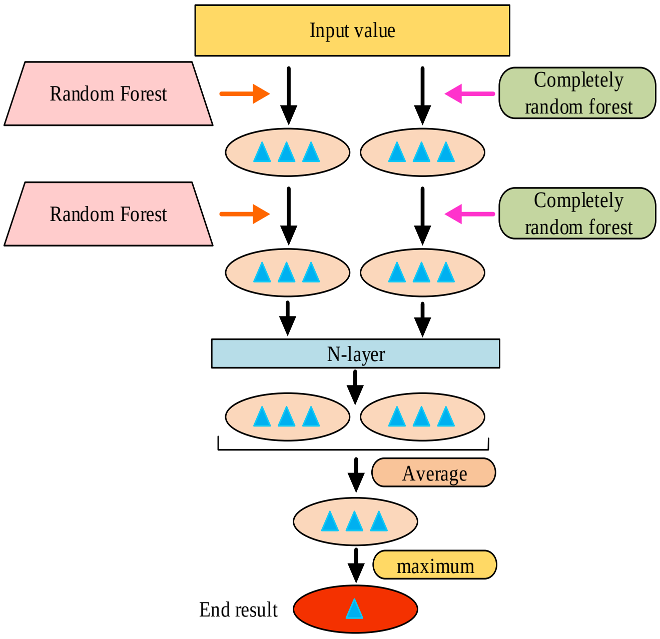


Fig. 1. Cascade forest structure.

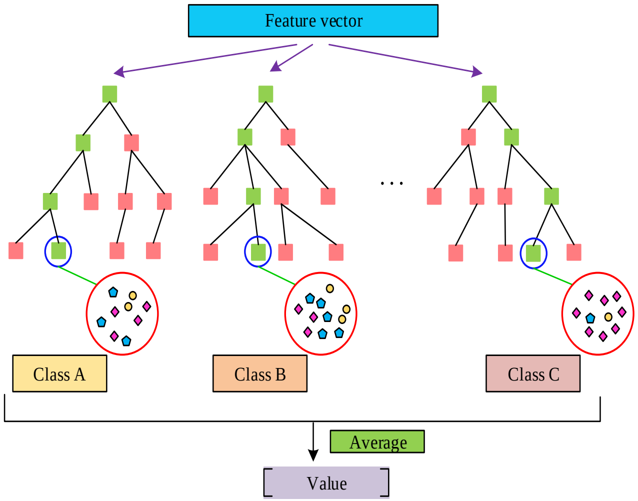


Fig. 2. Class vector generation method.

In Fig. 2, when generating class vectors, at each leaf node of each tree, the class vector is calculated based on the class distribution of the training instance, representing the relative proportion of each type in that node. For each tree in the forest, the separately calculated class vector will be averaged to form the final output of the tree. In addition to the initial layer, the forest at subsequent levels will fuse the class vectors generated by the previous layer with the original input features to form new inputs [17, 18]. Starting from the second layer of forest, regardless of how the number of trees in the forest changes, the dimension of the generated class vector will remain unchanged. The calculation method is to multiply the number of categories by the number of trees in the forest, and then add the dimension of the original features. If there is no significant improvement in performance on the validation set, it will stop the iteration and automatically select the optimal number of layers. When extracting image attribute features, involves the extraction of different attribute features. To improve the universality of the method, extreme gradient enhancement is introduced to optimize the DF algorithm. After a given training set, a tree ensemble model is generated, and then the expected value is obtained, as shown in Eq. (1).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{1}$$

In Eq. (1), $\hat{y}_i$ represents the expected value. F stands for regression tree. k represents the number of trees. x_i represents the content of the training set. f_k represents leaf nodes. The objective function is calculated from the expected value, as shown in Eq. (2).

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

In Eq. (2), $L(\phi)$ represents the objective function. l represents a differentiable loss function. y_i represents compliance true value. Ω is a regularization function. The loss function is minimized after generating a new tree, as shown in Eq. (3).

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) \tag{3}$$

In Eq. (3), $L^{(t)}$ represents the loss function after generating a new tree. f_t stands for new tree. Taylor expansion is performed and different steps of the original function are calculated to simplify the loss function, as shown in Eq. (4).

$$\tilde{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{4}$$

In Eq. (4), $\tilde{L}^{(t)}$ represents the simplified loss function. g_i represents the first-order degree of the original loss function. h_i represents the second-order degree of the original loss function. The optimal weight of the leaf node is calculated, as shown in Eq. (5).

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{5}$$

In Eq. (5), w_j^* represents the optimal weight of the leaf node. λ represents a constant term that prevents the denominator from being 0. Then the quality of the evaluation tree is calculated, as shown in Eq. (6).

$$\tilde{L}^{i}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{6}$$

In Eq. (6), $\tilde{L}^{i}(q)$ represents the quality value of the tree. T represents the number of leaf nodes in the tree. γ represents the leaf node coefficient. The process of extracting image attribute features by introducing extreme gradient enhancement is shown in Fig. 3.

In Fig. 3, when extracting image attribute features for CCPs, the original image data is first read and preprocessed before starting feature extraction. The preprocessed image data is input into the DF model for preliminary feature extraction. The initially extracted features are combined with the original image data to form an enhanced feature class vector, which is then fed into the next layer of DF model. After each feature extraction, the effectiveness of the features is evaluated through classification testing. After averaging and maximizing operations, the final feature extraction result of CCP image attributes is obtained.

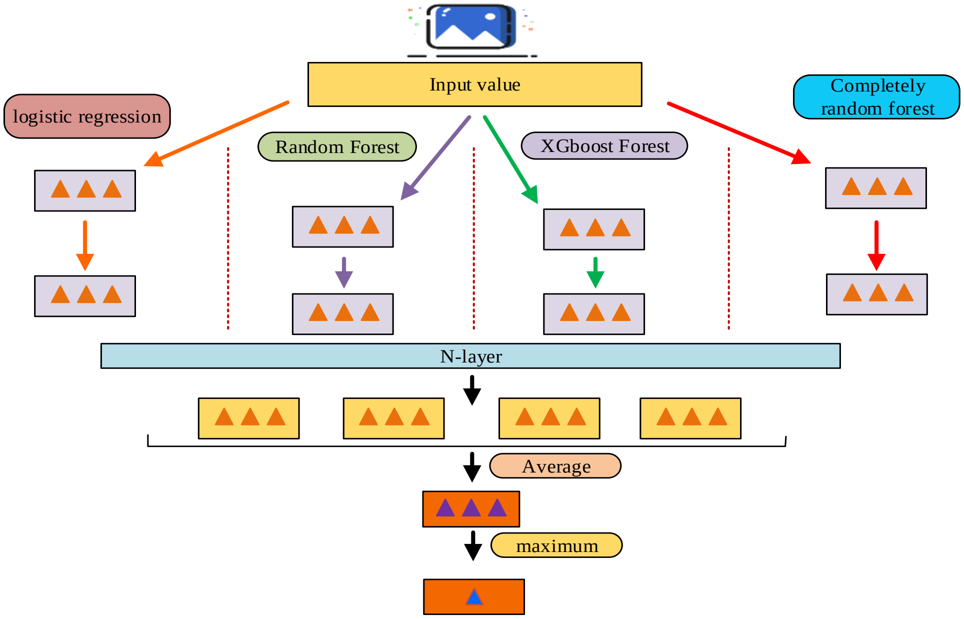


Fig. 3. Image attribute feature extraction process.

B. Image Retrieval Method for Cultural and Creative Products Combined with CNN

After extracting image attribute features, image attribute features can be used as the core data for image retrieval. CNN deep features are a data-driven high-level image feature with good image feature expression ability. They have certain advantages over traditional methods in extracting global, local features, and contextual information from images [19, 20]. Based on the DF algorithm, research designs the image retrieval methods for CCPs combined with CNN. The research uses the aggregation method of multi-scale and multi-region features in CNN and combines them with the spatial and channel weights of depth features to achieve multi-dimensional feature aggregation. When performing multi-region feature aggregation on CCP images, it needs to define several regions and calculate the aggregated feature vector for each region, as shown in Eq. (7).

$$f_{\Re} = \left[f_{\Re,1}, \dots, f_{\Re,\ell}, \dots, f_{\Re,\Im} \right]^{tra} \tag{7}$$

In Eq. (7), $f_{\Re}$ represents the region aggregation feature vector. tra stands for transpose matrix. $f_{\Re,\ell}$ represents the maximum value in the $\Re$ region on the ℓ th feature map. After obtaining the feature vectors of all regions, the normalized features are summed to form a multidimensional aggregated feature vector. When aggregating multi-dimensional features through spatial and channel weights of deep features, feature vectors are generated as shown in Eq. (8).

$$F_{crow} = \left[f_1, \dots, f_\ell, \dots, f_\Im \right], f_\ell = \sum_{y=1}^{H} \sum_{x=1}^{W} \alpha_{xy} \beta_\ell S_\ell (x, y) \tag{8}$$

In Eq. (8), F_{crow} represents the deep feature aggregation feature vector. α_{xy} represents spatial weight. β_ℓ represents channel weight. The feature aggregation process established in the study is shown in Fig. 4.

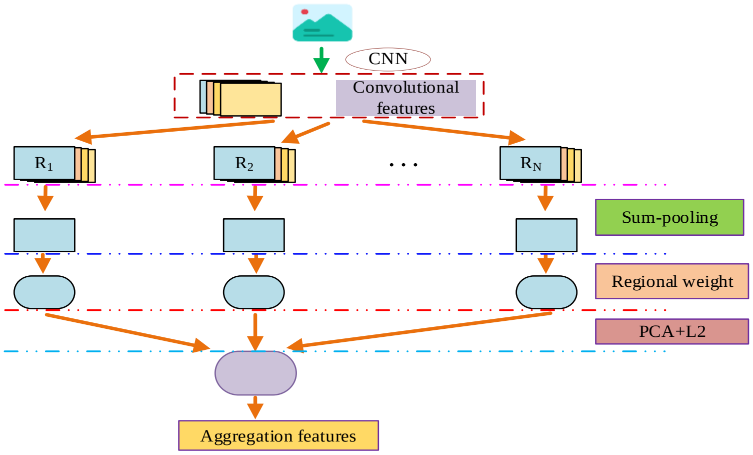


Fig. 4. Characteristic polymerization process.

In Fig. 4, when aggregating image features, the feature maps are first divided into sub regions, and the channel sensitivity weights are calculated based on the sparsity and intensity of the non-zero response values of the channels. After obtaining the significance weights of different regions, it calculates the feature vectors of different regions. The feature vectors of each region are normalized, one principal component analysis and whitening process are completed, and then a second normalization is performed. Sum-pooling is performed on the region features that have been normalized twice, and the aggregation results are normalized once to obtain the aggregated feature vectors. To more effectively extract effective information from feature maps, research is being conducted to incorporate channel sensitivity weights for optimization. The channel sensitivity is shown in Eq. (9).

$$\mathbb{N} = 1 - \gamma_\ell \tag{9}$$

In the above equations, $\mathbb{N}$ represents channel sensitivity. γ_ℓ represents the sum of the intensity amplitudes and positive response values of all non-zero response values on each channel. In the image of CCPs, the importance of different regions varies, and the corresponding weights also vary, as shown in Fig. 5.

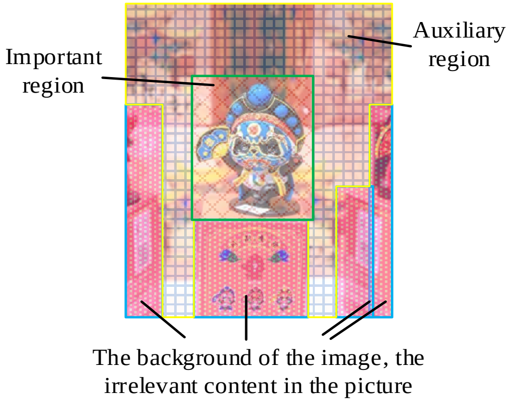

Important region
Auxiliary region
The background of the image, the irrelevant content in the picture

Fig. 5. Image area weight difference example.

As shown in Fig. 5, there are important areas in the image of CCPs that express the main body of the product, occupying the highest importance. In addition to important areas, there are auxiliary areas that play a secondary role in creating a visual atmosphere and enriching image details. The second most important factors are the background of the image and irrelevant content in the image. The significance weight of a region is defined as shown in Eq. (10).

$$\beta_r = \frac{\sqrt{\sum_{i \in R_r} \sum_{j \in R_r} S^{'}\left(i, j\right)^2}}{\sum \sum_{ij} S^{'} * A_r} \tag{10}$$

In Eq. (10), β_r represents the significance weight of the region. A_r represents the proportion of the size of the target area to the entire feature map. $S^{'}$ represents the sum of the values of all feature maps at each position. (i, j) represents the location of the feature map. The proportion of the target area to the entire feature map is calculated as shown in Eq. (11).

$$A_r = \frac{w_r \times h_r}{W \times H} \tag{11}$$

In Eq. (11), $W \times H$ represents the size of the feature map. $w_r \times h_r$ represents the size of the target area. Region feature vectors are generated as shown in Eq. (12).

$$\hat{f}_r^{\,\ell} = \left(\sum_{i \in R_r} \sum_{j \in R_r} \lambda_\ell S_\ell \left(i, j\right) \right)^\alpha \tag{12}$$

In Eq. (12), $\hat{f}_r^{\,\ell}$ represents the numerical value of the region feature vector on the ℓth feature map. λ_ℓ represents channel sensitivity weight. S_ℓ represents the sum of the values of the feature map at its position. The calculation of channel sensitivity weights is shown in Eq. (13).

$$\lambda_\ell = log\left(\frac{L\mathbb{C} + \sum_{\ell=1}^{L} \gamma_\ell}{\mathbb{C} + \gamma_\ell}\right) \tag{13}$$

In Eq. (13), $\mathbb{C}$ represents a small constant that ensures numerical stability. After extracting the image features of CCPs, the image retrieval module is constructed through feature semantic similarity measurement. In image retrieval, precise matching is often difficult to achieve due to the rich design elements and cultural connotations of CCPs. Fuzzy semantics provides a more flexible matching method, allowing the system to retrieve based on the similarity between image features and query conditions. The semantic similarity between two samples are calculated using fuzzy semantics as shown in Eq. (14).

$$s_{i,j} = \sum_{k=1}^{K} \left(\mu_{\xi_{P_i}} \left(P_j^k \right) + \mu_{\xi_{P_j}} \left(P_i^k \right) \right) \tag{14}$$

In Eq. (14), $s_{i,j}$ represents semantic similarity. $\mu_{\xi_{P_i}}$ represents the fuzzy semantics of templates. P_j^k represents the k nearest neighbor sample among the sample neighbors. The retrieval results are output based on semantic similarity. The operation process of the deep forest convolutional neural network (DF-CNN) CCP image retrieval module designed for research is shown in Fig. 6.

In Fig. 6, when conducting image retrieval for CCPs, after preprocessing the image and query information, the feature representation of the image is extracted. Image features are input into the feature space and an index with the image library is established. The similarity between query information and image library images is measured and matched to obtain corresponding retrieval images and complete sorting. At the same time, the results are fed back to the image library to enrich the search information of the image library. Finally, a list of content for image retrieval of CCPs is obtained, and the retrieval is completed.

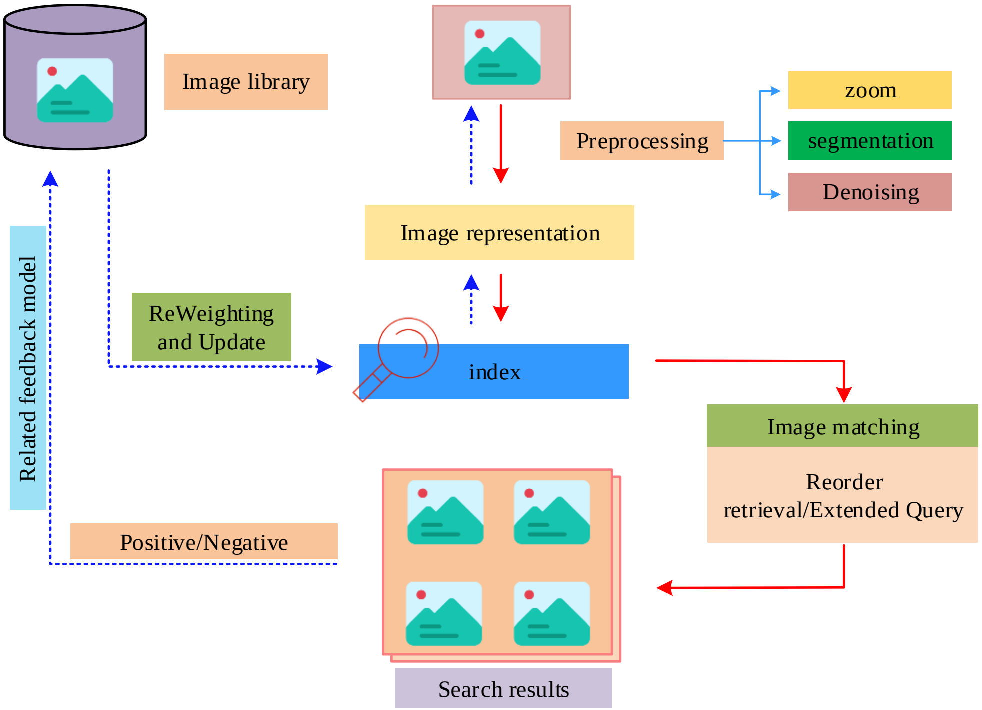


Fig. 6. Cultural and creative products image retrieval module operation process.

III. RESULTS

A. Performance Testing of Image Retrieval Technology for Cultural and Creative Products

To analyze the performance of the image retrieval technology for CCPs designed for research at runtime, ImageNet and Open Images Dataset datasets were selected for testing. When conducting performance testing, the software and hardware environments are shown in Table I.

During testing, the core objects in the dataset images were equated with important areas in the CCP images, and images containing only a single core object and multiple core objects were divided. DF-CNN was compared with the currently advanced Inception Networks and Speeded Up Robust Features methods during testing. The convergence performance of different methods was compared, as shown in Fig. 7.

TABLE I. EXPERIMENTAL SOFTWARE AND HARDWARE ENVIRONMENTAL PARAMETERS

Software and hardware environment name	**Parameter specification**
Device type	Deep learning server
Processor	13th Gen Intel(R) Core(TM) i5-13490F 2.50 GHz
Graphics card	Nvidia GTX 4080
Internal memory	64GB
Hard disk space	1TB
Operating system	Ubuntu 14.04
Experimental platform	Caffe, Python 2.7.12, Matlab R2015b

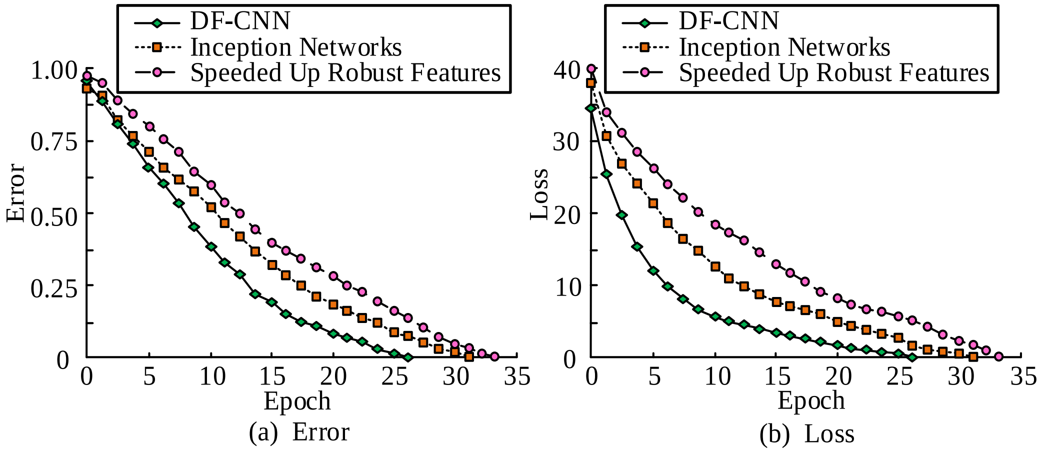


Fig. 7. Convergence performance test.

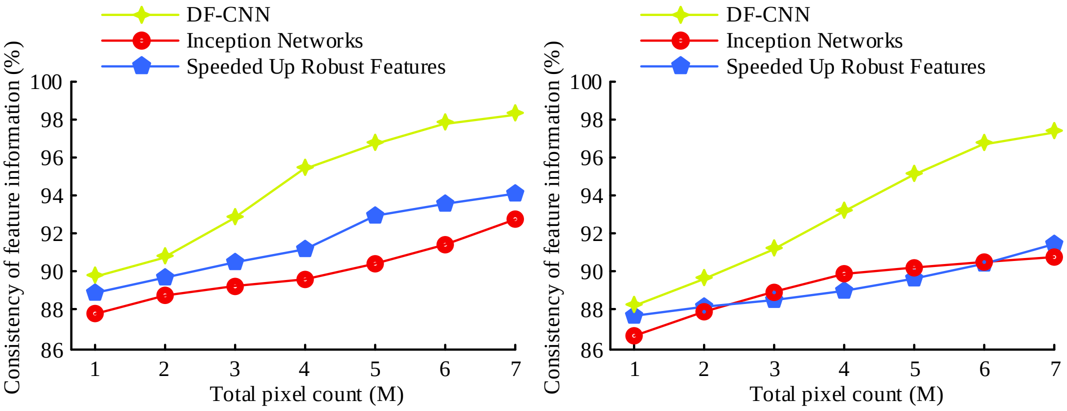
In Fig. 7, the overall convergence trend of performance for different methods during training was consistent. Fig. 7(a) shows that during the error convergence process, the error value of DF-CNN at the beginning of the iteration was lower than that of Speeded Up Robust Features and higher than that of Inception Networks. Within 5 iterations, the error value of DF-CNN rapidly decreased and was already lower than Inception Networks by the 5th iteration. In Fig. 7 (b), during the loss convergence process, the error value of DF-CNN at the beginning of the iteration was 34, which was lower than the Speeded Up Robust Features and Inception Networks. The loss value of DF-CNN decreased rapidly in the early stage and gradually slowed down in the later stage until the training was completed. The DF-CNN proposed in the study only required 26 iterations of training to achieve the optimal error and loss values, which were 7 and 10 fewer than Speeded Up Robust Features and Inception Networks, respectively. This indicates that the research method has better convergence performance and high training efficiency. The consistency of extracting image feature information was tested using different methods in images containing different numbers of pixels, as shown in Fig. 8.

From Fig. 8, the consistency of the results obtained by different methods in extracting image feature information increased with the total number of pixels in the image. Fig. 8 (a) shows that when the image was set to only contain a single cultural and creative object, the image feature information consistency of Speeded Up Robust Features was 94.1% when the total pixel amount of the image reached 7M. The similarity of image feature information in DF-CNN increased rapidly as the total pixel size of the image increased from 1M to 4M, and then gradually slowed down the rate of increase. During the entire process, the similarity of image feature information was 89.8% when the total pixel size of the image was 1M. When the total pixel count of the image reached 7M, it increased to 98.4%. From Fig. 8(b), when the image contained three composite cultural and creative objects, the image feature information consistency of Speeded Up Robust Features was 91.6% when the total pixel amount of the image reached 7M. The image feature information consistency of Inception Networks was 90.7% when the total pixel size of the image reached 7M. The image feature information consistency of DF-CNN was 97.6% when the total pixel size of the image reached 7M. The research method experienced a slight decrease in performance when extracting image features from multiple objects, but the magnitude of the decrease was smaller than other methods. This indicates that the research method has good performance in extracting image feature information.

B. Analysis of the Practical Application Effect of Image Retrieval Technology for Cultural and Creative Products

When analyzing the practical application of image retrieval technology for CCPs designed in research, the study extracted CCP images from a CCP design website as a retrieval image library. The running time of different methods was analyzed, and to improve the accuracy of statistical results, each scheme was repeated 10 times, presented in the form of mean and standard deviation. Seven different search fields were set and abbreviated as A, B, C, D, E, F, and G. The runtime was divided into two different stages: feature extraction and retrieval result generation, as shown in Fig. 9.

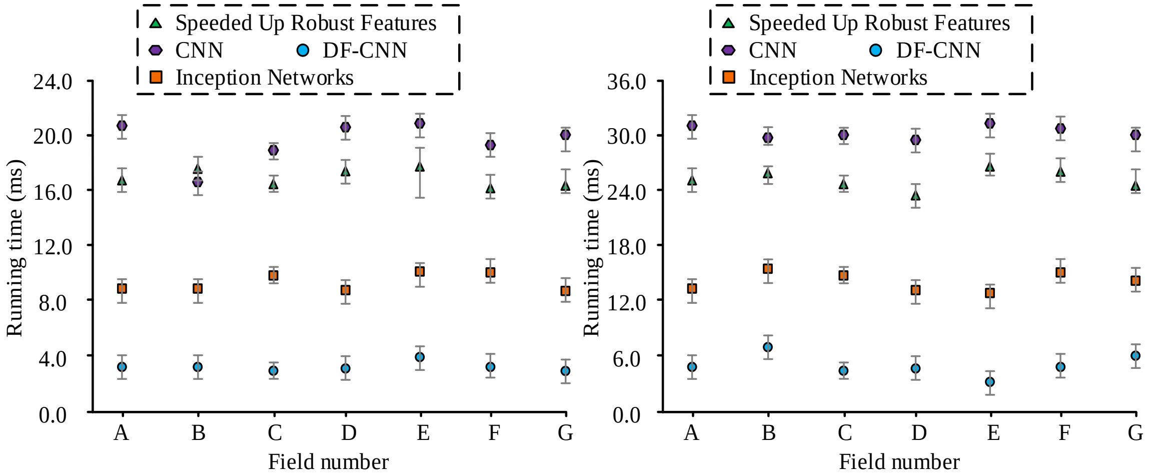
In Fig. 9, the runtime time of different methods was related to the fields and generally remained within a certain range. Fig. 9 (a) shows that during feature extraction, CNN had significantly higher runtime in six out of seven fields compared to other methods. The running time of Inception Networks was significantly lower than that of Speeded Up Robust Features, reaching its lowest point in the D field, between 7.0ms and 8.0ms. The running time difference of DF-CNN in the seven fields was relatively small, with a minimum of 2.3ms and a maximum of only 4.7ms. The fluctuation in time during multiple runs was also maintained within 2.0ms. As shown in Fig. 9 (b), different methods exhibited significant efficiency stratification when generating search results. Among the seven fields, the results were ranked according to the longest running time of CNN, the second longest running time of Speeded Up Robust Features, the third longest running time of Inception Networks, and the shortest running time of DF-CNN. The search result generation time of DF-CNN in seven fields was the highest at only 7.4ms, and the lowest at 1.6ms. The research method had higher operational efficiency at runtime and could complete image retrieval of CCPs at a faster speed. The proportion of false retrieval content in the retrieval results of different methods in practical applications was analyzed, as shown in Fig. 10.



(a) Single cultural and creative object

(b) 3 composite cultural and creative objects

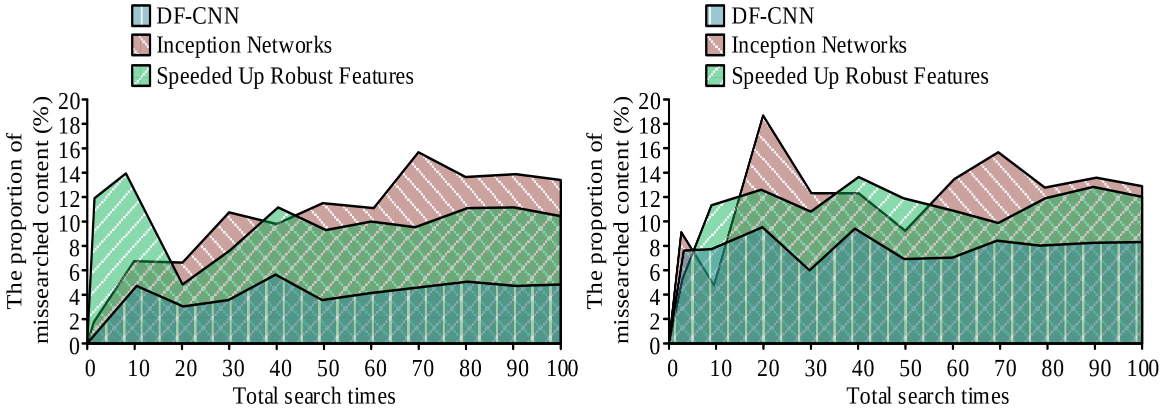
Fig. 8. Extracting the coincidence degree of image feature information.



(a) Feature extraction

(b) Search result generation

Fig. 9. Runtime analysis.



(a) Single cultural and creative object

(b) 3 composite cultural and creative objects

Fig. 10. Analysis of the proportion of false retrieved content.

As shown in Fig. 10, different methods tended to stabilize the proportion of false retrieval content as the number of searches increased when generating retrieval content. Fig. 10 (a) shows that when searching for a single CCP object, Inception Networks had the highest proportion of false searches in 100 searches, reaching 15.8%. When the number of searches reached 80, it tended to stabilize and fluctuated slightly around 13.7%. The false search rate of Speeded Up Robust Features reached a maximum of 14.1% in 100 searches, and stabilized when the search frequency reached 50, mainly fluctuating in the range of 10.0% to 11.0%. The proportion of false searches in DF-CNN remained within 6.0% during 100 search cycles, and tended to stabilize at around 5.0% when the search cycle reached 50. From Fig. 10(b), when searching for three comprehensive CCPs, the proportion of false searches by Inception Networks in 100 search processes ultimately fluctuated around 13.4%. The proportion of false searches in Speeded Up Robust Features ultimately fluctuated around 12.3%. The proportion of false searches in DF-CNN tended to stabilize when the number of searches reached 50, fluctuating around 8.2%. The research method can better exclude irrelevant content when conducting image retrieval of CCPs. The overall content matching of the generated search result list was analyzed, and 20 searches on each key field were performed, as shown in Fig. 11.

In Fig. 11, in the analysis of content matching in search results, the matching degree of the results obtained from multiple searches could be integrated into a range of intervals. When the length of the search result list was 20, the content matching degree of Speeded Up Robust Features mainly fluctuated around 78%. The content matching degree of the retrieval results of Inception Networks mainly fluctuated around 85%. The content matching degree of DF-CNN search results mainly fluctuated around 92%. As the length of the search result list increased, there was a certain degree of decrease in the content matching of the search results for Speeded Up Robust Features and Inception Networks. However, the degree of decrease in the content matching degree of DF-CNN search results was minimal. When the length of the search result list reached 100, the content matching degree of the search results still fluctuated around 91%. The research method could maintain high operational performance and retrieval accuracy for a long time when conducting a large number of CCP image retrieval tasks.

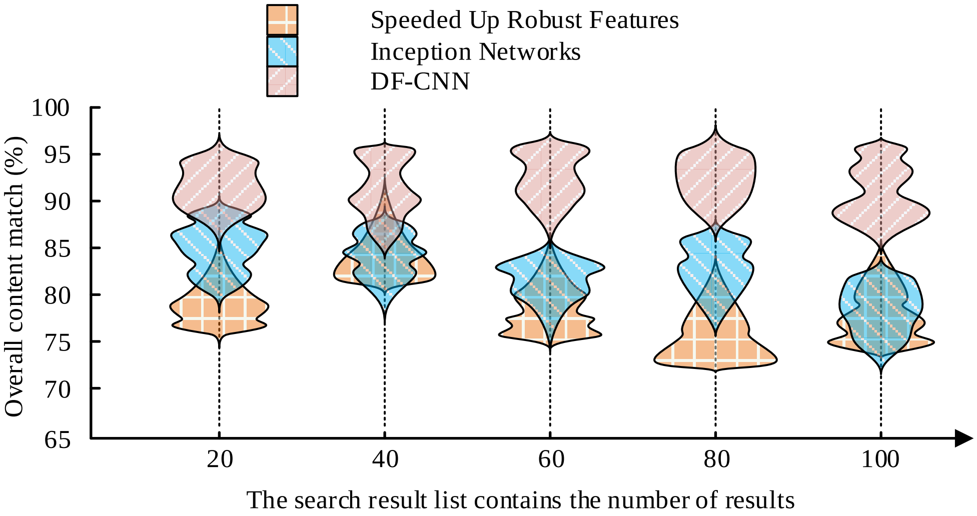


Fig. 11. Search results content matching degree analysis.

IV. Discussion and Conclusion

A CCP search assistance technology based on DF-CNN image retrieval module was studied and designed to enhance the exposure ability of CCPs. In the process, the number of decision trees in the forest was used as a hyperparameter to evaluate the effectiveness of features through classification testing. The aggregated feature vectors were calculated for each region, and region features which completed twice normalization were conducted sum-pooling aggregation. The significance weight of the region was defined. After extracting the image features of CCPs, the image retrieval module was constructed using feature semantic similarity measurement. The similarity measurement was performed on the query information and image library images to match the corresponding retrieval images. Finally, the effectiveness of the method was analyzed. The experimental results showed that in convergence performance testing, the research method only required 26 iterations of training to achieve the optimal error and loss values. When setting the image to only contain a single cultural and creative object for image feature information matching analysis, it increased to 98.4% when the total pixel size of the image reached 7M, indicating that DF-CNN could effectively capture the detailed features of the image when processing high-resolution images, providing richer information for image retrieval. When searching for three comprehensive CCPs, the proportion of false searches in the research method ultimately fluctuated around 8.2%, indicating that DF-CNN could effectively handle the complexity and diversity of features in multi-object images. The content matching degree of research method retrieval results has been fluctuating around 91% for a long time, which means that when using DF-CNN for image retrieval, users can obtain highly relevant results to the query, thereby improving the user experience of retrieval. The research method could generate higher quality retrieval results when conducting image retrieval of CCPs. With the development of technology, the application of 3D images and dynamic images was becoming increasingly widespread. Future work can consider extending DF-CNN to these new image types to meet a wider range of application requirements. However, in actual deployment and application, the scalability and robustness of the model also need to be considered. Future work can explore the combination of image retrieval technology and natural language processing to achieve text-based image retrieval, to improve the accuracy of retrieval and the naturalness of user interaction. At the same time, research can also be conducted on retrieval algorithms that are suitable for 3D models and video content, to meet the market's demand for dynamic and stereoscopic visual content retrieval. User privacy protection in the process of image retrieval should be strengthened and how to provide personalized retrieval services without leaking user data needs to be studied.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] Zhao W X, Liu J, Ren R, Wen J R. Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems, 2024, 42(4): 1-60.

[2] Rana M, Bhushan M. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications, 2023, 82(17): 26731-26769.

[3] Zhang F, Xu M, Xu C. Tell, imagine, and search: End-to-end learning for composing text and image to image retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2022, 18(2): 1-23.

[4] Humenberger M, Cabon Y, Pion N, Weinzaepfel P, Lee D, Guérin N, et al. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. International Journal of Computer Vision, 2022, 130(7): 1811-1836.

[5] Salih S F, Abdulla A A. An effective bi-layer content-based image retrieval technique. The Journal of Supercomputing, 2023, 79(2): 2308-2331.

[6] Wang H, Qu H, Xu J, Wang J, Wei Y, Zhang Z. Texture image retrieval based on fusion of local and global features. Multimedia Tools and Applications, 2022, 81(10): 14081-14104.

[7] Li Q, Tian X, Ng W W Y, Kwong S. Recent development of hashing-based image retrieval in non-stationary environments. International Journal of Machine Learning and Cybernetics, 2022, 13(12): 3867-3886.

[8] Khan S U, Hussain T, Ullah A, Baik S W. Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance. Multimedia Tools and Applications, 2024, 83(5): 15079-15100.

[9] Xin B, Zheng Y, Wang T, Chen L, & Wang Y. A diagnosis method based on depthwise separable convolutional neural network for the attachment on the blade of marine current turbine. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 2021, 235(10):1916-1926.

[10] Kumar S, Mankame D P. Optimization driven deep convolution neural network for brain tumor classification. Biocybernetics and Biomedical Engineering, 2020, 40(3): 1190-1204.

[11] Thirusangu N, Subramanian T, Almekkawy M. Segmentation of induced substantia nigra from transcranial ultrasound images using deep convolutional neural network. The Journal of the Acoustical Society of America, 2020, 148(4):2636-2637.

[12] Ma Y, Song Q, Hu M, Zhu X. Correction: A Lightweight Neural Learning Algorithm for Real-Time Facial Feature Tracking System via Split-Attention and Heterogeneous Convolution. Neural Processing Letters, 2023, 55(2): 1581-1581.

[13] Ashtiani F, Geers A J, Aflatouni F. An on-chip photonic deep neural network for image classification. Nature, 2022, 606(7914): 501-506.

[14] Naeem A, Anees T, Ahmed K T, Naqvi R A, Ahmad S, Whangbo T. Deep learned vectors’ formation using auto-correlation, scaling, and derivations with CNN for complex and huge image retrieval. Complex & Intelligent Systems, 2023, 9(2): 1729-1751.

[15] Rajwar K, Deep K, Das S. An exhaustive review of the metaheuristic algorithms for search and optimization: Taxonomy, applications, and open challenges. Artificial Intelligence Review, 2023, 56(11): 13187-13257.

[16] Gheisari M, Hamidpour H, Liu Y, Saedi P, Raza A, Jalili A, Rokhsati H, Amin R. Data Mining Techniques for Web Mining: A Survey. Artificial Intelligence and Applications, 2023, 1(1): 3-10.

[17] Bhosale Y H, Patnaik K S. Application of deep learning techniques in diagnosis of covid-19 (coronavirus): a systematic review. Neural processing letters, 2023, 55(3): 3551-3603.

[18] Godwin E C, Izuchukwu C, Mewomo O T. Image restorations using a modified relaxed inertial technique for generalized split feasibility problems. Mathematical Methods in the Applied Sciences, 2023, 46(5): 5521-5544.

[19] Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 2023, 35(2): 757-774.

[20] Wu C, Khishe M, Mohammadi M, Taher Karim S H, Rashid T A. RETRACTED ARTICLE: Evolving deep convolutional neutral network by hybrid sine–cosine and extreme learning machine for real-time COVID19 diagnosis from X-ray images. Soft Computing, 2023, 27(6): 3307-3326.

Method for Mission Analysis Using ToT-Based Prompt Technology Utilized Generative AI

Satellite Mission Analysis for SaganSat-0 of Remote Sensing Satellite

Kohei Arai
Dept. of Intelligent Information Science, Saga University, Saga City, Japan

***Abstract*—Method for mission analysis using ToT: Tree of Thought-based prompt technology utilized generative analysis AI is proposed. Mission analysis needs methods for simulation of the supposed images which will be acquired with the imaging mission instruments, and the other mission instruments. In order to create simulation images, ToT-based prompt technology utilized generative AI is used. An application of the proposed method is shown for a mission analysis for SaganSat-0 of remote sensing satellite which will carry three mission instruments, a 720-degree camera, a thermal infrared camera and a Geiger counter. The simulated images and the Geiger counter sounds created by the proposed method are shown here together with analyzed results.**

***Keywords*—CoT; ToT; AI; Mission analysis; prompt technology; generative AI; SaganSat-0; remote sensing satellite; 720-degree camera; thermal infrared camera; Geiger counter**

I. Introduction

In recent natural language processing, Chain-of-Thought (CoT) prompting for language models is a technique that includes a series of steps to solve a problem in the prompt (i.e., the text input to the language model) [1]-[10]. The basic CoT prompting mimics the "step-by-step thought process" and includes the flow of reasoning steps as examples/samples for few-shot learning. Namely, LLMs (large-scale language models) such as ChatGPT are not good at inference. If you want them to think logically, you have to ask them questions in a way that encourages logical thinking. This is called CoT prompting. Few-Shot prompting is a method of generating answers by giving a few examples. The idea is to lay the groundwork for thinking by giving examples and having ChatGPT think about it. Zero-Shot prompting is used for difficult questions, but in the sense of giving a few examples, CoT prompting is similar in some ways.

In problem solving, if questions are one of those of which someone has already answered, ChatGPT will give a clear answer. However, if the question has never been answered by anyone, ChatGPT will immediately give you an ambiguous answer. This is true not only for "unknown problems", but also for personal problems that cannot be generalized, such as your own or your company's. For such problems, we need to narrow down the scope of the problem and show the logical path. In short, if we can cover his weakness (inability to deduce or construct logic) on the human side, we can get a very useful answer because ChatGPT has knowledge from all over the world that we personally cannot have. This is the true meaning of "prompt engineering".

According to a paper published by a team at Google's AI research lab in January 2023 [11], "CoT prompts are a technique that does not immediately ask for the answer to a problem, but rather involves intermediate inference steps and then layering step-by-step inferences (chains of thought) to arrive at the answer." In this way, the CoT prompt shows the flow of thought to solve the problem logically and deductively, but because it is one-way, it may lead to the wrong answer in the case of more difficult problems. Therefore, the Self-consistency with CoT prompt was devised, which presents multiple answers and allows the user to choose the appropriate answer from among them. However, for problems that involve complex processes, verifying and evaluating all the answers one by one will not be enough resources no matter how many there are. The Tree of Thoughts (TOT) prompt, announced in May 2023 by a joint research team from Google's DeepMind team and Stanford University, is effective for problems that have multiple paths to the goal and are difficult to arrive at the correct answer [12]-[22].

Like COT, thinking deductively is desired, step by step, but in this case, it is unknown what order to do it in. Even with real problems, it's not uncommon to not know where to start. In such cases, we use the concept of "decision tree" to think about it. There are many options at each stage to get to the goal. This is a way to think about which route to take. This TOT prompt can be applied to solving complex problems. First, we ask participants to come up with some solutions in ChatGPT, and then we evaluate them based on criteria such as feasibility and budget. (For example, (certain/possible/impossible) or "on a 5-point scale.") Then, we dig deeper into the top ideas. A good question would be to ask participants to "dig deeper from a certain point of view." After that, we ask them to summarize their ideas in ChatGPT. "Based on your analysis so far, please rank your ideas in order of highest to lowest." This ToT prompt is close to human thinking, and as the DeepMind team suggested it, it is the logic used in games such as Go-play, Shogi, and autonomous driving.

The purpose of this paper is to present usefulness and effectiveness of the ToT for generation of images, moving pictures, and so on in comparison to the other CoT, plain generative AIs. By using ToT prompt technology, mission analysis (simulation of mission instrument data) is conducted. As an example, mission analysis for SaganSat-0 of remote sensing satellite which carries three mission instruments, 720-degree camera, thermal camera and Geiger counter is conducted. ToT prompts are used for creation of simulation imagery data of the three mission instruments. Furthermore, mission analysis is conducted successfully. SaganSat-0 was launched in the midnight on 5 August 2024 and will be put into the orbit of ISS: International Space Station on 29 August 2024. Therefore, the simulation data will be verified after the SaganSat-0 data acquisition and also confirmed mission analyzed results after that.

In the next section, SaganSat-0 mission is described together with three mission instruments. Then, simulation of these mission instruments data is described with the proposed ToT-based prompt technology. After that, some mission analysis results are described followed by conclusion with some discussions.

II. SaganSat-0

SaganSat0 is a small satellite of the CubeSat (1U) type supported by the Saga Prefectural Space Science Museum. Below is more information about SaganSat0:

Satellite details:

Name: SaganSat0 (SaganSat No. 0)

Type: CubeSat (1U)

Status: Not launched, scheduled for 2024

Launch rocket: Falcon 9 (Cygnus, NG-21)

Project overview: Three mission instruments, 720-degree camera, thermal infrared camera, Geiger counter

Support organization: Saga Prefectural Space Science Museum

Contact: a-ito@yumeginga.jp.nospam

Website:

SaganSat's website aims to develop nanosats and create support contacts

Japanese news:

The Saga Shimbun newspaper has extensive coverage of news about SaganSat0

From these information, SaganSat0 is a small satellite project led by the Saga Prefectural Space Science Museum, scheduled for launch on 5 August 2024 to ISS and will be released from the ISS and putted into ISS orbit in the near future (late of August 2024).

A. Mission Instruments

The SaganSat-0 satellite is equipped with three mission instruments. One of them (Mission 1) is two 360-degree cameras (720-degree cameras) that test the Earth-pointing direction and the deep space direction. The 360-degree cameras take images with a resolution of 800 by 600, and the image size of the two images is about 60 KB (0.45 degrees/pixel) in total. If the radius of the Earth is R, the satellite altitude (ISS altitude = 400km) is H, and the instantaneous field of view is θ, the resolution of the Earth's surface is expressed by Formula (1). The resolution of the Earth's surface is about 3142m.

$$R=H\cdot\tan(\theta) \quad (1)$$

The Earth's spheroid used here is the standard spheroid called GRS80, and the geodetic coordinate system using it has a polar radius of 6356.752 km in ITRF84. The infrared camera for Mission 2 will take images at a resolution of 800 by 650, with an image size of about 40KB; the resolution is 200m, and since the down-link is at 4.8kbps, one set of Mission 1 and 2 data can be down-linked in 21 seconds. Mission 3 data can be down-linked at 42B/21 seconds.

Since the altitude of the International Space Station (ISS) is about 400 km, the range over which a ground station can communicate with the ISS is affected by the curvature of the Earth and the characteristics of the antenna of the ground station, and an approximate method for calculating the basic range is expressed by Formula (2), which calculates the range over which a ground station can communicate with the ISS (line of sight).

$$d=\sqrt{(2hR+h^2)} \quad (2)$$

where: d is the distance above the horizon (km), h is the altitude of the ISS (km), and R is the radius of the Earth (about 6371 km). Therefore, the range over which a ground station can communicate with the ISS is about 2293 km. Within this range, direct communication with the ISS is possible, and since the speed of the ISS is about 28,000 km/h (7.8 km/s), the visible range is about 4.91 minutes = 294.6 seconds. Therefore, about 14 frames of Mission 1, 2, and 3 data can be downloaded.

B. 720-Degree Camera

The 720-degree camera is composed of two 360-degree cameras. Moving pictures and still pictures can be acquired with the different operation modes. Instant Field of View: IFOV of the camera is 0.45 degrees.

C. Thermal Infrared Camera

Arducam：2MP Global Shutter OV2311 Mono Camera Modules Pivariety (NoIR), compatible with Raspberry Pi ISP, and Gstreamer Plugin-Arducam is used for mission #2 of thermal infrared camera. Moving pictures and still pictures can be acquired as well. Swath width is 163 km, Instantaneous field of view is 102m, Frame rate is 30fps for moving picture acquisition, The number of pixels can be selected (1) 1600 by 1300, (2) 800 by 650, Quantization bit rate is 8bit, Transmission bandwidth is 499.2 Mbps, Camera clock is 27MHz in maximum, and it can be transmitted one frame image in every 18.5 seconds for 1600 by 1300 mode, and in 4.6 seconds for 800 by 650 mode, respectively.

D. Geiger Counter

The ISS is located at an altitude of about 400 km, outside the Earth's magnetosphere, making it an environment that is easily affected by cosmic rays: a Geiger counter makes a "click" sound each time it detects radiation (alpha particles, beta particles, gamma rays, etc.). On Earth, the amount of radiation is low, so the sound is heard sporadically, but on the ISS, the intensity of radiation is much higher, so the detection sound is more frequent. For example, natural radiation levels on Earth would make a Geiger counter click a few dozen times per minute, but on the ISS, hundreds to thousands of clicks per minute would be heard.

III. Simulation of Mission Instrument Data with a Simple Prompt of the Generative AI

A. 720-Degree Camera Images

360-degree camera images which will be acquired with 720-degree camera of Mission #1 are tried to create with a simple prompt of the generative AI (ChatGPT, for instance). Two examples are shown in Fig. 1. The prompt is as follows, “Please create 360-degree camera image of the Earth from the altitude of ISS (400km)”. As is shown in Fig. 1, it seems like nighttime scenes of cloudy South American continent with the thin atmosphere and shining the sun.

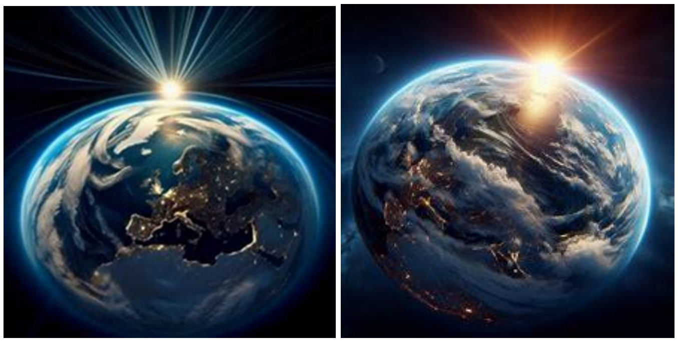


Fig. 1. Two example images which will be acquired with 720-degree camera of Mission #1 created with a simple prompt of the generative AI (ChatGPT).

The following ToT of prompt (“Please decide to choose the Pika[1] generated images or Tenbin AI generated images for creation of 360-degree camera image of the Earth from the altitude of ISS (400 km)”) is attempted to the creation of 360-degree camera images. Tenbin AI is created by GMO[2], Japan which allows the use of more than six generative AI at once free of charge. Such that the decision tree selection of the prompt does work to generate many appropriate images. Fig. 2 shows the images created by the ToT prompt proposed here.

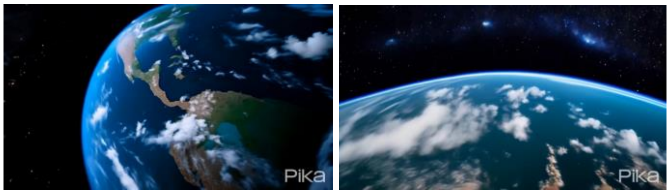
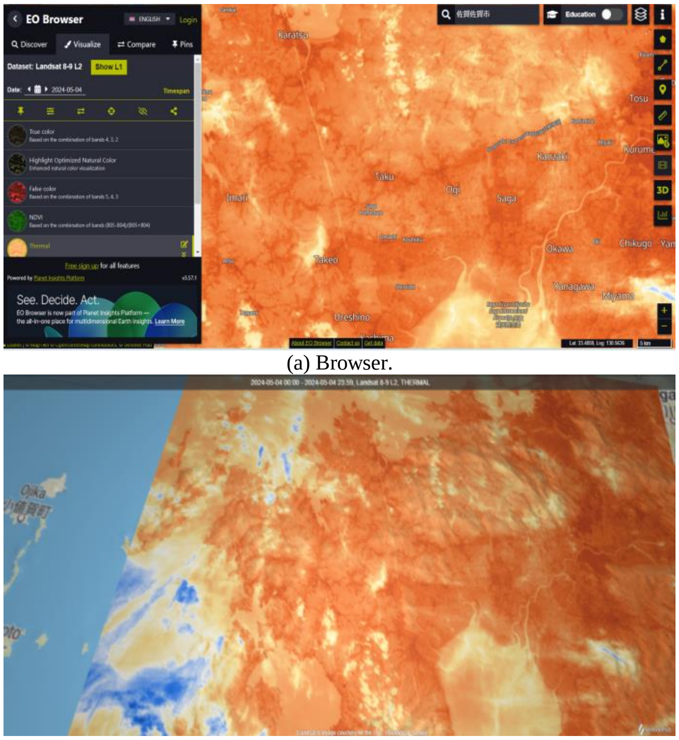


Fig. 2. Two example images which will be acquired with 720-degree camera of Mission #1 created with the proposed ToT prompt of the generative AI (ChatGPT).

Curvature of the Earth can be seen. At an altitude of 400 km, the curvature of the Earth is clearly visible. The curved horizon is a distinctive feature. Continents and the oceans can also be seen. The vast oceans can be seen together with the spreading continents and their boundaries. During the day, the sunlight reflects beautifully off the land and water surfaces. The thin blue layer of atmosphere that surrounds the Earth can be seen as well. This is the Earth's atmosphere. On the night side of the Earth, scattered city lights can be identified. The boundary between day and night (terminator) divides the Earth into two halves. The night sky features a large number of stars and a clear view of the Milky Way.

B. Thermal Infrared Camera Images

Landsat-9/TIRS-2(Thermal Infrared Sensor-2), OLI-2(Operational Land Imager-2) can be used to simulate the mission #2 of thermal infrared camera images. By using EO (Earth Observation) Browser[3] provided by European Space Agency, simulation images can be simulated. Fig. 3 shows an example of the simulated thermal infrared images with 200 m spatial resolution of Saga city and its surroundings which was acquired on 4 May 2024.



(a) Browser.

(b) 200m resolution of the simulated thermal infrared camera image.

Fig. 3. EO browser and example of the simulated thermal infrared camera image.

The following ToT of prompt (“Please decide to choose the EO browser derived Pika generated images or Tenbin AI-generated images for creation of thermal infrared camera image of a land containing flooding areas are included from the altitude of ISS (400km)”) is attempted to the creation of thermal infrared camera images. Such that decision tree selection of prompt does work to generate many appropriate images. Fig. 4 shows the image created by the ToT prompt proposed here.

[1] https://pika.art/

[2] https://tenbin.ai/

[3] https://apps.sentinel-hub.com/eco-browser/?zoom=11&lat=33.29182&lng=130.03693&themeId=DEFAULT-THEME&visualizationUrl=https%3A%2F%2Fservices.sentinel-hub.com%2Fogc%2Fwms%2Ffa073661-b70d-4b16-a6a9-e866825f05fd&datasetId=AWS_LOTL2&fromTime=2024-05-04T00%3A00%3A00.000Z&toTime=2024-05-04T23%3A59%3A59.999Z&layerId=THERMAL&demSource3D=%22MAPZEN%22

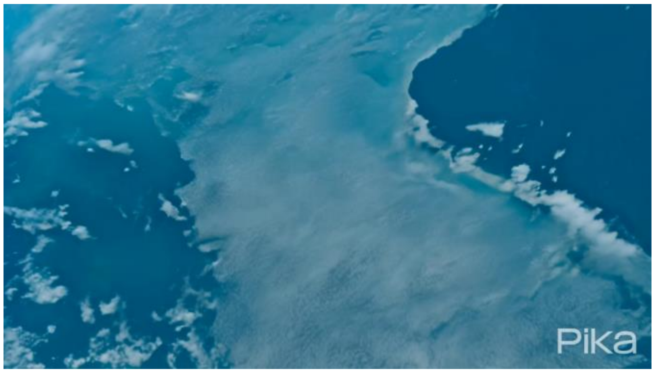


Fig. 4. Example of the simulated thermal infrared camera image.

C. Geiger Counter Sounds

The sound of natural radiation levels on Earth would be tick... tick... tick... tick... tick... tick... tick... whereas the sound of cosmic rays on the ISS would be tick-tock-tock-tock-tock-tock-tock... Thus, while on the ISS in orbit, the sound of the Geiger counter can be heard very frequently, and almost continuously, due to the frequent detection of high energy particles.

Famous links to actual audio data recording the sounds of cosmic rays in ISS orbit are (1) JAXA Kibo Public Relations and Information Center[4] and (2) NASA Glenn Research Center[5]. Listening to these audio recordings gives you an idea of what life is like on the ISS and makes the presence of cosmic rays seem more familiar to you. Geiger counters can detect not only cosmic rays but also other types of radiation such as beta rays and gamma rays, so not all sounds are necessarily caused by cosmic rays.

By using ToT-based prompt technology, the sound of cosmic rays detected by a Geiger counter at an altitude of 400 km on the ISS can generally be simulated with the following prompt. “Choose the generate sounds as a quiet crackling when the satellite altitude is low or popping sound for the relatively higher satellite altitude”. Like a crackling, similar to a light sound like a spark flying. It occurs every time a cosmic ray strikes the detector of a Geiger counter. As a popping sound, a sound similar to a small popping sound (occurs every time a high-energy cosmic ray strikes the detector) is heard. The loudness of the sound depends on the energy of the cosmic ray and the sensitivity of the detector, but generally, the more energetic the cosmic ray, the louder it will be. Also, a more sensitive detector can detect even weaker cosmic rays.

IV. Mission Analysis with the Simulated Images and the Sounds

A. 720-Degree Camera Images

By using generative AI (Tenbin AI) with the following simple prompt of “What insights can you glean from the images you've uploaded?” the following response is obtained.

The uploaded image shows the Earth from space, particularly the North American continent. Key insights from the image include are as follows, (1) Roundness of the Earth: From space, the Earth is clearly spherical. (2) Atmosphere: The atmospheric layer surrounding the Earth is depicted as a beautiful blue atmospheric color. (3) Clouds: Clouds of various shapes are spread across the Earth's surface, indicating the movement of the atmosphere. (4) Land and Ocean: The land of North America and the ocean surrounding it are clearly distinguishable. (5) Stars in the night sky: Countless stars shine against the background of the Earth. (6) Word "Pika" in the bottom left of the image: The word "Pika" is written in the bottom left of the image, which may indicate that the image is part of some kind of content.

B. Thermal Infrared Camera Images Containing Flooding Areas

Thermal infrared imagery creates images based on the temperature of objects. Flooded or submerged areas have a lower temperature than the surrounding land, and therefore appear darker in thermal infrared imagery. For this image, the following methods are used to detect flooded or submerged areas: Simple thresholding method compares the brightness values of all pixels in the image and considers pixels that are lower than a certain threshold as flooded or submerged areas; Object-based segmentation method identifies objects in the image and determines whether they are flooded or submerged based on their temperature and shape; and Machine Learning method uses historical flood data to train a model to automatically detect flooded or submerged areas from thermal infrared imagery.

For the simple threshold method, the histogram shows that the brightness values in the image are relatively evenly distributed. There is a small peak on the left side of the image that could be a flooded or submerged area. Also, all pixels with a brightness value below 100 are considered to be flooded or submerged areas. There are some areas on the left side of the image that could be flooded or submerged areas.

Using Tenbin of generative AI with the prompt of “Extract flooding areas from the simulated thermal infrared image”, mission analysis of flooding area detection is conducted.

The simulated thermal image is converted to gray-scale and has a size of 1125 by 2000 pixels. Fig. 5 shows the gray-scale image. Next, we proceed to detect flooded areas. In thermal imagery, flooded areas usually appear cooler (darker in gray-scale) due to lower water temperature compared to the surrounding land. A thresholding technique is applied to highlight these cooler regions. In the thresholded image, the potentially flooded areas are highlighted in black. These regions correspond to the cold areas in the thermal imagery, indicating the accumulation of water due to flooding. Fig. 6 shows the detected flooding areas.

Other flooding area detection method with the following simple prompt “Detect flooded or submerged areas from uploaded thermal infrared images” using Tenbin of generative AI is attempted. Fig. 7 (a) shows the uploaded thermal infrared camera image with the histogram while Fig. 7 (b) shows the histogram equalized image with the histogram. On the other hand, Fig. 7 (c) shows gray-scale image of the uploaded thermal infrared camera image while Fig. 7 (d) shows

[4] https://humans-in-space.jaxa.jp/iss/

[5] https://www.nasa.gov/missions/analog-field-testing/why-space-radiation-matters/

histogram stretched image of the gray-scale image, respectively.

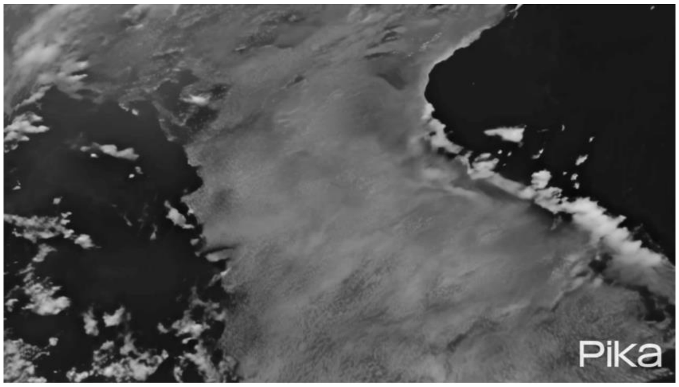


Fig. 5. Gray-scale of the simulated thermal infrared camera image.

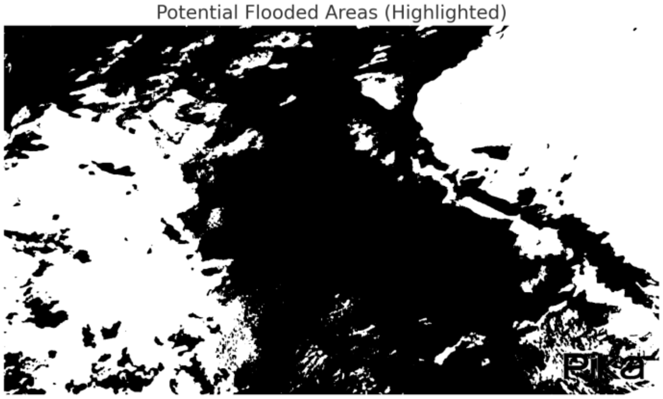


Fig. 6. Detected flooding areas from the gray-scale image.

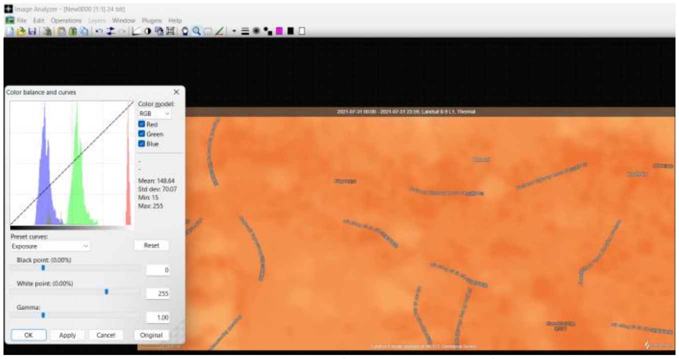


Image Analyzer - [New0000 [1:1] 24 bit]
File Edit Operations Layers Window Plugins Help
Color balance and curves
OK
Apply
Cancel
Original
Reset

Color model: RGB
☑ Red
☑ Green
☑ Blue
Mean: 148.64
Std dev: 70.07
Min: 15
Max: 255
Preset curves: Exposure
Black point: (0.00%) 0
White point: (0.00%) 255
Gamma: 1.00

(a) Uploaded simulation image from a thermal infrared camera.

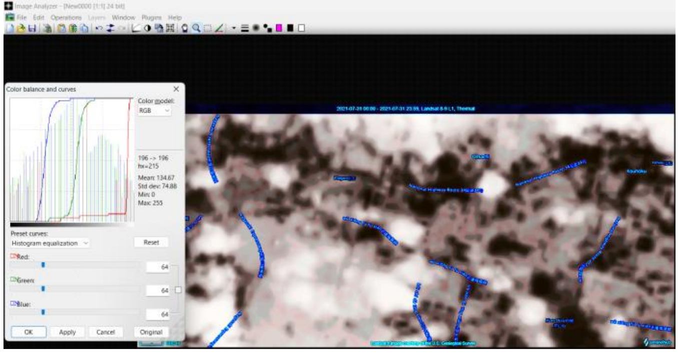


Image Analyzer - [New0000 [1:1] 24 bit]
File Edit Operations Layers Window Plugins Help
Color balance and curves
OK
Apply
Cancel
Original
Reset

Color model: RGB
196 -> 196
hx=215
Mean: 134.67
Std dev: 74.88
Min: 0
Max: 255
Preset curves: Histogram equalization
Red: 64
Green: 64
Blue: 64

(b) Histogram equalized image.

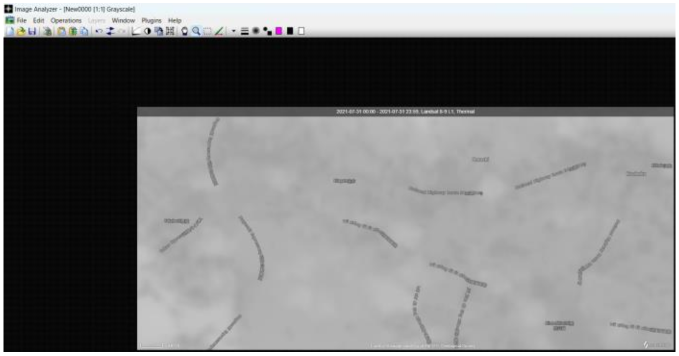


Image Analyzer - [New0000 [1:1] Grayscale]
File Edit Operations Layers Window Plugins Help

(c) Gray-scale image.

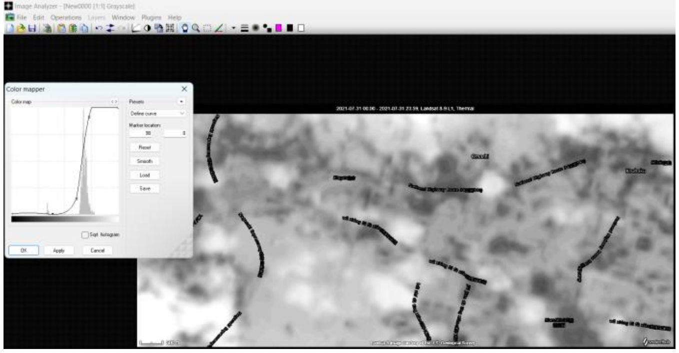


Image Analyzer - [New0000 [1:1] Grayscale]
File Edit Operations Layers Window Plugins Help
Color mapper
OK
Apply
Cancel

(d) Histogram stretched image.

Fig. 7. Flooding area detection with the simulated thermal infrared camera image.

C. New Insights Gained by Combining 360-degree Camera Images and Thermal Infrared Camera Images

Using the Tenbin AI with the simple prompt "Combining visible light images from a 360° camera taken at ISS altitude with thermal infrared images at 200 m resolution allows for many interesting applications" the following responses are obtained,

1) Analyzing the urban heat island effect: By identifying urban structures with visible light images and observing their heat distribution with thermal infrared images, you can investigate temperature rises in urban areas in detail.

2) Early detection and monitoring of forest fires: By identifying smoke and flames with visible light images and identifying high-temperature areas with thermal infrared images, forest fires can be detected earlier and more accurately.

3) Observing ocean surface temperature distribution: By visually identifying ocean currents and eddies with visible light images and observing the distribution of sea surface temperatures with thermal infrared images, studying climate change and its impact on marine ecosystems can be done.

4) Assessing agricultural productivity: Identifying crop types and distribution with visible light images and evaluating crop stress and irrigation efficiency with thermal infrared images can be performed.

5) Volcanic activity monitoring: By observing the shape and plumes of volcanoes with visible light imagery and detecting thermal anomalies with thermal infrared imagery, you can monitor changes in volcanic activity.

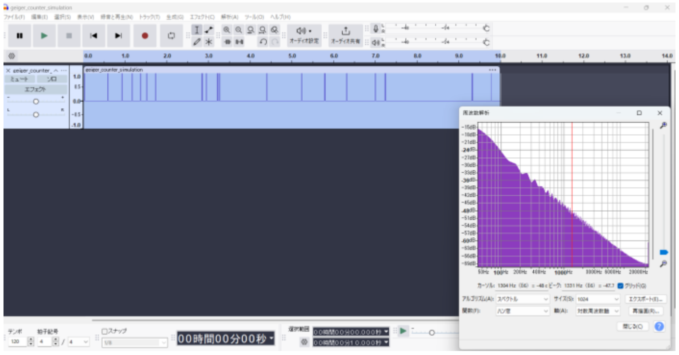
6) Urban planning and sustainability research: By examining urban structures and green spaces with visible light imagery and identifying energy-efficient buildings and areas with thermal infrared imagery, you can use this data for urban planning.

7) Observing changes in glaciers and polar regions: By examining the shape and extent of glaciers with visible light imagery and observing temperature changes with thermal infrared imagery, you can study the effects of climate change.

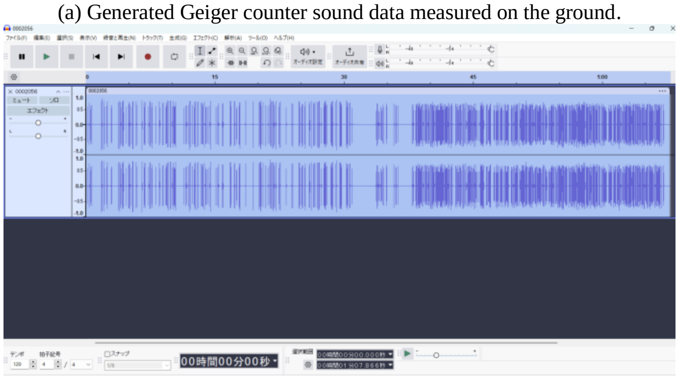
8) Air pollution research: By observing smog and atmospheric conditions with visible light imagery and observing temperature gradients in the atmosphere with thermal infrared imagery, you can study the distribution and dynamics of air pollution.

D. Insights from the Simulated Geiger Counter Sound Data

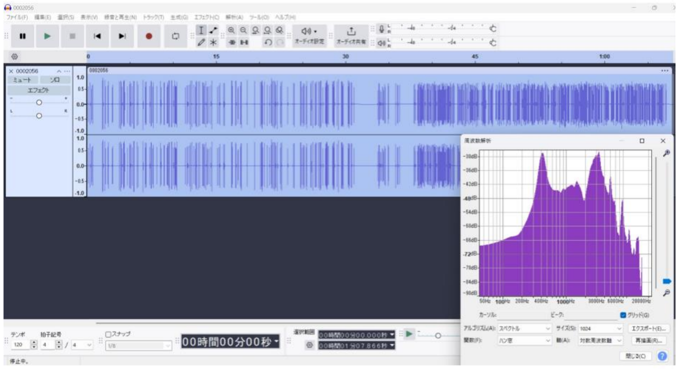
Geiger counter sound data on the Earth’s surface is simulated by using the Tenbin AI with the simple prompt “Generate Geiger counter sound data which is measured on the ground and the 400 km altitude”. Then the frequency components of the Geiger counter sound data are investigated with Audacity of sound data analysis software tool[6]. Fig. 8 (a) shows the generated Geiger counter sound data measured on the ground and its frequency components while Fig. 8 (b) shows the generated Geiger counter sound data at the 400 km altitude. Meanwhile, Fig. 8 (c) shows its frequency components.



(a) Generated Geiger counter sound data measured on the ground.



(b) Generated Geiger counter sound data measured at the 400 km altitude.



(c) Frequency components of the generated Geiger counter sound data measured at the 400 km altitude.

Fig. 8. Generated Geiger counter sound data measured on the ground and at 400 km altitude (SaganSat-0 altitude) and their frequency components.

V. Conclusion

Method for mission analysis using ToT: Tree of Thought-based prompt technology utilized generative analysis AI is proposed. Mission analysis needs methods for simulation of the supposed images which will be acquired with the imaging mission instruments, and the other mission instruments.

In order to create simulation images, ToT-based prompt technology utilized generative AI. An application of the proposed method is shown for a mission analysis for SaganSat-0 of remote sensing satellite which will carry three mission instruments, a 720-degree camera, a thermal infrared camera and a Geiger counter. The simulated images and the Geiger counter sounds created by the proposed method are shown here together with analyzed results. In summary, it is found that ToT-based prompt technology is superior to the other CoT and plain generative AIs.

VI. Future Research Works

SaganSat-0 of remote sensing satellite was launched with launching vehicle Falcon-9 on 5 August 2024 and will be put into orbit at 400 km altitude just same as ISS orbit after being released from the ISS in the near future. Then, three mission instrument data will be acquired after all. Therefore, mission analysis will be made after the actual three mission instrument data are obtained. ToT-based prompt technology will be evaluated by using the actual three mission instrument data through a comparison. Then improvements will take place for ToT prompt technology.

Acknowledgment

The author would like to thank Prof. Dr. Hiroshi Okumura for his valuable discussions for this research works. Also, the author would like to thank Sho Kobayashi and M. Tanaka of Saga Prefectural Space Science Museum Yumeginga for their providing data and information.

References

[1] Jason Wei, Xuezhi Wang, and Dale Schuurmans, Chain of Thought Prompting: A Simple yet Effective Method for Improving Reasoning in Language Models, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022), 2022.

[6] https://forest.watch.impress.co.jp/library/software/audacity/download_10718.html

[2] Yixin Nie, Pengcheng Yin, and Graham Neubig, "Improving Chain of Thought Prompting with Recursive Decomposition", 2022. https://arxiv.org/abs/2209.07141, Accessed on 15 August 2024.

[3] Yufei Wang, et al., "Chain of Thought Prompting for Solving Math Word Problems" 2022. https://arxiv.org/abs/2209.13341, Accessed on 15 August 2024.

[4] Yixin Nie, et al., "Chain of Thought Prompting for Multimodal Reasoning", 2022. https://arxiv.org/abs/2210.02419, Accessed on 15 August 2024.

[5] Xuezhi Wang, et al., "Chain of Thought Prompting for Commonsense Reasoning", 2022. https://arxiv.org/abs/2210.03541, Accessed on 15 August 2024.

[6] Yixin Nie, et al., "Analyzing the Effectiveness of Chain of Thought Prompting", 2022. https://arxiv.org/abs/2210.05191, Accessed on 15 August 2024.

[7] Pengcheng Yin, et al., "Evaluating the Reasoning Ability of Language Models with Chain of Thought Prompting", 2022. https://arxiv.org/abs/2210.06241, Accessed on 15 August 2024.

[8] Yufei Wang, et al., "Chain of Thought Prompting for Dialogue Systems", 2022. https://arxiv.org/abs/2210.08191, Accessed on 15 August 2024.

[9] Xuezhi Wang, et al., "Chain of Thought Prompting for Explainable AI", 2022. https://arxiv.org/abs/2210.09241, Accessed on 15 August 2024.

[10] Yixin Nie, et al., "Chain of Thought Prompting for Multitask Learning", 2022. https://arxiv.org/abs/2210.10341, Accessed on 15 August 2024.

[11] Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma, Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou, “Google Research”, Brain Team {jasonwei,dennyzhou}@google.com, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, 2023, https://ar5iv.labs.arxiv.org/html/2201.11903, Accessed on 15 August 2024.

[12] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, Karthik Narasimhan, “Tree of Thoughts: Deliberate Problem Solving with Large Language Models, Computer Science > Computation and Language”, [Submitted on 17 May 2023 (v1), last revised 3 Dec 2023 (this version, v2)], https://arxiv.org/abs/2305.10601, Accessed on 15 August 2024.

[13] Pengcheng Yin, et al., "Think-Then-Act: Bridging the Gap between Reasoning and Acting in Language Models", 2022. https://arxiv.org/abs/2209.07131, Accessed on 15 August 2024.

[14] Yixin Nie, et al., "Improving Think-Then-Act Prompting with Recursive Reasoning", 2022. https://arxiv.org/abs/2210.02411, Accessed on 15 August 2024.

[15] Yufei Wang, et al., "Think-Then-Act Prompting for Solving Math Word Problems", 2022. https://arxiv.org/abs/2210.13331, Accessed on 15 August 2024.

[16] Yixin Nie, et al., "Think-Then-Act Prompting for Multimodal Reasoning", 2022. https://arxiv.org/abs/2210.04411, Accessed on 15 August 2024.

[17] Xuezhi Wang, et al., "Think-Then-Act Prompting for Commonsense Reasoning", 2022. https://arxiv.org/abs/2210.05511, Accessed on 15 August 2024.

[18] Yixin Nie, et al., "Analyzing the Effectiveness of Think-Then-Act Prompting", 2022. https://arxiv.org/abs/2210.06111, Accessed on 15 August 2024.

[19] Pengcheng Yin, et al., "Evaluating the Reasoning Ability of Language Models with Think-Then-Act Prompting", 2022. https://arxiv.org/abs/2210.07211, Accessed on 15 August 2024.

[20] Yufei Wang, et al., "Think-Then-Act Prompting for Dialogue Systems", 2022. https://arxiv.org/abs/2210.09111, Accessed on 15 August 2024.

[21] Xuezhi Wang, et al., "Think-Then-Act Prompting for Explainable AI", 2022. https://arxiv.org/abs/2210.10211, Accessed on 15 August 2024.

[22] Yixin Nie, et al., "Think-Then-Act Prompting for Multitask Learning", 2022. https://arxiv.org/abs/2210.11311, Accessed on 15 August 2024.

Author’s Profile

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan (Current JAXA) from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of Brawijaya University, Kurume Institute of Technology and Nishi-Kyushu University. He also was Vice Chairman of the Science Commission “A” of ICSU/COSPAR for 2008-2016 then he is now award committee member of ICSU/COSPAR. He wrote 87 books and published 710 journal papers as well as 650 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

IoT-Based Integrated Heads-Up Display for Motorcycle Helmet

L. Raj[1], M. Batumalay[2], C. Batumalai[3], Prabadevi B[4]
Faculty of Data Science and Information Technology, INTI International University, Malaysia, Nilai, Malaysia[1, 2, 3]
School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India[4]

***Abstract*—The prevalence of visual impairment among the global population is a growing concern, with rates continuing to rise at an alarming pace. According to statistics from the World Health Organization (WHO), an estimated 2.2 billion people globally live with some form of visual impairment. Several methods exist to aid the blind in everyday navigation, such as walking sticks and guide dogs. However, these aids do not come without their drawbacks. For instance, using traditional guide dogs may not be suitable for some individuals due to allergies, cultural beliefs, or being unable to take care of a living animal due to the level of responsibility required. Innovations such as smart walking sticks and robotic guide dogs are continually being developed to overcome these gaps and cater to the unique requirements of the visually impaired. Hence, this proposed system is equipped with a joystick-controlled robotic guide that mimics the responsibilities of a traditional guide dog. The proposed system features an obstacle avoidance feature that will detect obstacles in its environment to avoid collisions. It will also provide audio feedback through a Bluetooth-connected mobile application when an obstacle has been detected. The proposed system is a product innovation which can be targeted to benefit visually impaired users by providing them with more independence as well as convenience in terms of mobility. Upon performing acceptance testing with the target audience, the system has been found to achieve its target in aiding the guidance of blind individuals.**

***Keywords*—Heads-up display; motorcycle helmet; Internet of Things (IoT); android application; Raspberry Pi; product innovation**

I. Introduction

In 2015, the United Nations (UN) created the 17 interrelated global goals, also referred to as the Nation Goals or Sustainable Development Goals (SDGs). By 2030, a more equitable and sustainable future is to have been established, and many of the world's social, economic, and environmental problems are to have been resolved. SDG 9 is about infrastructure, industry, and innovation; HUD is in line with this objective. HUD technology integrated into motorcycle helmets can provide real-time navigation, weather updates, and speed information directly in the rider's line of sight. This not only enhances the riding experience but also promotes road safety by minimizing the need for riders to look away from the road [1-2].

Helmets are protective headwear that shields the head and brain from injury during a range of activities, especially ones that include the risk of impact or falls. As previously indicated [3 - 5], the present generation of helmets is lacking significantly, especially considering the increasing prevalence of the Internet of Things (IoT) [6-9] and its use in the automobile sector, including automated braking, head-up displays, and other features. Helmets are used for many different purposes and can be found in many different forms. Motorcycle riders make up the majority of road users, however, because there is a lack of innovation and products in the ASEAN market, the helmet is seen as a less important part of the IoT. IoT are deployed in several applications for the convenience of users [10-12].

The idea for heads-up display systems was initially created for use in automobiles, including commercial aircraft, cars, and military vehicles. To make sure they are free of any obstructions, any user can flip between the two cameras that surround the car and the aeroplane. Operating the gadget is as simple. The navigation system is one of its main benefits; it keeps the pilot or driver focused on manoeuvring the vehicle rather than getting distracted [3,13]. The LCI-HUD Helmet System makes use of the same capability. The main CPU board is the Raspberry Pi, which is also equipped with Bluetooth.

A camera mounted behind the rider's helmet does double duty as a rearview mirror and, more importantly, closes the blind spot for motorcyclists. On the other hand, the helmet's built-in headset allows the rider to talk hands-free while answering calls or listening to music. A button on their visor controls the display.

A map navigation display appears on the device once the user uses the mobile application to set the destination on the helmet. Using the music application, the rider can also play music via a playlist display on the visor. Rather than pulling over to the side of the road to answer a call, the rider can use [8] to answer hands-free. Information about the caller will be shown on the visor using the HUD system. This reduces the possibility that the rider will drive while distracted.

Fig. 1 illustrates the proposed work, which is an improved low-cost integrated head-up display (LCI-HUD). A multitude of information can be projected onto the HUD by integrating the LCI-HUD with currently running third-party applications. One crucial element is the concept of an inexpensive, multipurpose LCI-HUD. Users can retain their user data using LCI-HUD after installing and registering mobile apps. APIs are used to connect the music and additional third-party extensions. LCI-HUD integrates as a result. LCI-HUD interacts with third-party applications to provide users with immediate access to a range of services, including social media notifications, navigation, and weather updates, on the HUD. This seamless connectivity offers convenience while driving, enhancing the user experience all around. The use of APIs ensures that the LCI-HUD stays up to date with the latest features and advancements in third-party apps.

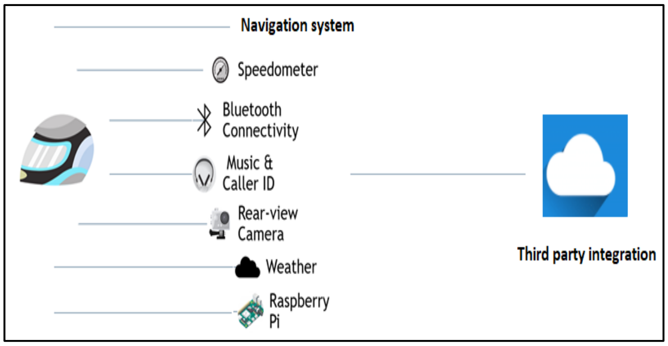


Fig. 1. Integrated heads-up display.

To check blind spots or view dashboard information like the speedometer, riders donning the existing (non-HUD) helmet are compelled to glance away from the road. Additionally, call receiving and navigation are done via external devices like cell phones. Roadside incidents are dangerous because they require quick attention and are easy to miss; using these devices could cause a shift in focus. Driving will become safer and less frequent if drivers maintain their attention on the road. To monitor blind areas or view dashboard information like the speedometer, riders using the existing (non-HUD) helmet are forced to turn their heads and look away from the road. In addition, calls and navigation are handled by external devices, like smartphones. Because road incidents are time-sensitive and easy to overlook, these devices will cause a shift in attention and pose a risk. The number of incidents will go down if drivers keep their eyes on the road for a safer driving experience.

The transparent display of HUD technology allows riders to view data without having to shift their typical range of vision, which is why it is recommended. Driving safety could be increased with the use of HUDs. Head-up display (HUD) navigation on a motorcycle is safer than smartphone navigation, according to a previous study. Additionally, by connecting the LCI-HUD with pre-existing third-party applications, it may show a variety of data on the HUD. There are several features available on the market for a fair price for the proposed LCI-HUD. The LCI-HUD can save user data after users install and register it. The integration of LCI-HUD follows. The small, lightweight additional hardware on the helmet won't restrict rider mobility.

To reduce distraction, motorcyclists can designate their preferred destination on their phone through the navigation module, and it will appear on the visors of their helmets [6]. To lessen the additional weight required to construct a bespoke board with the navigation module and to lower the cost of the hardware, the navigation module will retrieve navigational data from the smartphone. A weather module is added to assist the riders with changes in the weather. When motorcyclists are aware of the weather ahead of time, they may carefully plan their route.

The caller ID module enables the rider to receive a phone call with a pop-out on the rider’s visor. Using a button mounted on the handlebar of the bike, riders can answer or reject phone calls. Passengers can use the Music Module to adjust the music on their phones. To forward or rewind the music, a button is attached. Information is continuously shown on the visor via the Rear-View Camera Module, which is mounted to the back of the helmet and functions like a rear-view camera. By doing this, bikers will be able to avoid their blind spot and feel more confident when changing lanes without looking back.

As seen in Fig. 2, the user must first register for the mobile application. Using Firebase, the user will be able to access and exit the database remotely, and this data will be stored online. The rider and the Raspberry Pi must be connected via Bluetooth. The mobile application primarily controls two functions: music and navigation. The LED button on the Raspberry Pi allows the rider to select the display on the visor. The modular HUD's hardware will feature a camera affixed to the back of the helmet to function as the rear view and more sensors that the user can add in the future. The weather conditions displayed on the visor and data access are also guaranteed by the Raspberry Pi.

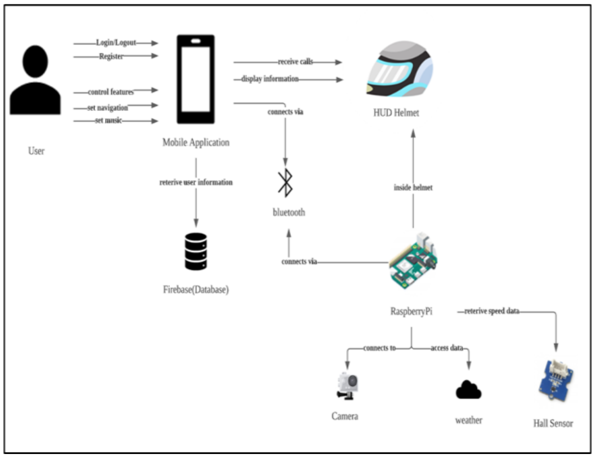


Fig. 2. Rich picture diagram.

II. Literature Review

There are two methods utilised to collect the necessary data: surveys and interviews. The target user base for data gathering is riders, to compare the proposed HUD helmet system with the existing helmet. The plan aims to boost safety while highlighting life experience in contrast to the existing helmet. The surveys will provide quantitative data on user preferences and feedback, while the interviews will offer qualitative insights into user experiences and suggestions for improvement. By targeting riders specifically, the research focused on the primary users of the HUD helmet system to ensure that safety and user experience are prioritized in the development process.

A. *Requirements Phase*

The interviews reveal information on the riders' perspectives and experiences with the existing helmet, as well as any prospective upgrades. The riders' preferences and perceptions of safety elements were quantified using questionnaires. Better user experience and more safety precautions were ensured by customizing the proposed HUD helmet system to the needs and preferences of the target users through their involvement in the data collection process. Motorbike riders were provided with a web-based questionnaire comprising multiple choice questions as part of the quantitative approach. Fig. 3 illustrates riding conditions that can cause distraction. Fig. 3 also presents the data collected on "distraction resulting from current riding experience.

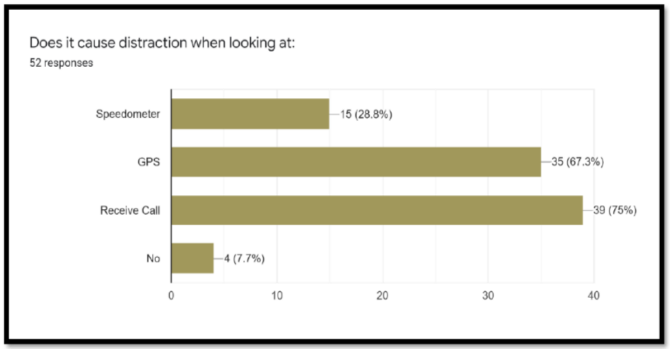


Fig. 3. Distraction due to current riding experience.

The intended LCI-HUD system as well as traditional helmets were the subjects of interviews conducted using qualitative methods. The purpose of the information gathering was to raise awareness about the HUD system in helmets and study any extra features that might be needed, in addition to the basic sensors and features. The HUD Helmet System can help when riding in accordance with the feedback that was received. Since it has features that make riding more convenient, the HUD Helmet System is preferred for use. It also helps to integrate everything together. The interviewee also offers suggestions for future developments that might be made to the HUD Helmet, including voice command, predictive braking distance, lane departure warning, and lane maintain assist. Beyond the advantages of using the existing helmet, the HUD Helmet System also offers the user benefits. To lessen the necessity of constantly checking his phone while riding, another interviewee indicated that he currently wears Bluetooth headphones. In line with other interviewees' responses, the information the respondent desired to have in the helmet—namely, navigation and display speed was also advantageous. Since it needs to determine the speed limit on the road where the rider is, the additional features, like the over-speeding warning message, could be a future enhancement. As conclusion, the interviewee believes that having a HUD helmet system is a good concept since it will assist motorcyclists stay focused because information is displayed on the visor.

B. Design Phase

Fig. 4 displays the activity diagram flow that outlines the riders' involvement, beginning with opening the LCI-HUD app and identifying whether they are first-time users. Once the registration procedure is complete, riders can log in. Riding the LCI-HUD system requires a Bluetooth connection after logging in. Following connection identification, riders can exchange visor information by pressing the LED button. Navigating while maintaining eye contact with traffic is made possible for riders by HUD helmet technology. By putting notifications, speed, and navigational data in real-time right in the rider's line of sight, this technology improves safety. Rider customization and distraction-free connectivity are made possible by the LED button, which allows riders to quickly switch between visor information.

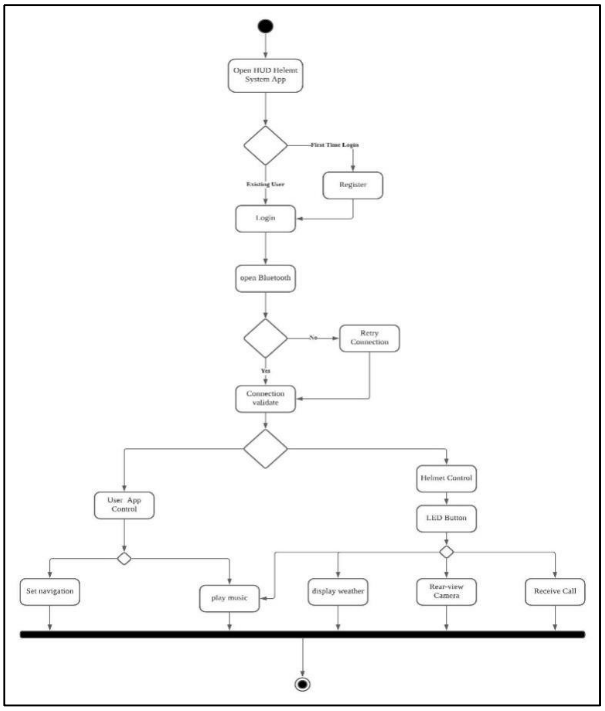


Fig. 4. Activity diagram.

As shown in Fig. 5, the characteristics that are used to specify objects form the basis of the class diagram. The UserID, a primary key that is automatically assigned to the rider at registration, is one of the features in the user database. The Bluetooth host is identified as the user's mobile phone. The further prerequisites are a password, phone number, and email address. The rider and the LCI-HUD system table will be connected via the database table. There are many relations; they are switched between one-to-many in the database and one-to-one in the LCI-HUD system by the database table. This suggests that a single user might be a part of many LCI-HUD systems and accounts. The table contains the board Bluetooth receive address and unique main key for HUD-based helmet system identification. Every user in the database table has a unique identity according to the primary key. There is an assurance that no two users are utilizing the same ID. To enable secure communication and data transfer between the user's helmet system and the LCI-HUD system, the Bluetooth receive address is utilized.

Fig. 6 shows the proposed LCI-HUD system's conceptual configuration. The schematic arrangement of the proposed LCI-HUD system is shown in Fig. 6, which also gives an idea of how the various parts are connected. This diagram facilitates the process of implementing and debugging the system by helping to comprehend its general structure and functionality. The brown cable is a camera module that is attached directly to the CSI port. The black GrovePi+ cable allows the sensor module to be operated directly from a socket. The orange wire, which is connected to the hall sensor and speedometer, respectively, is what powers the LED button. The last cable, which is blue, connects to the headset via the 3.5mm audio connector.

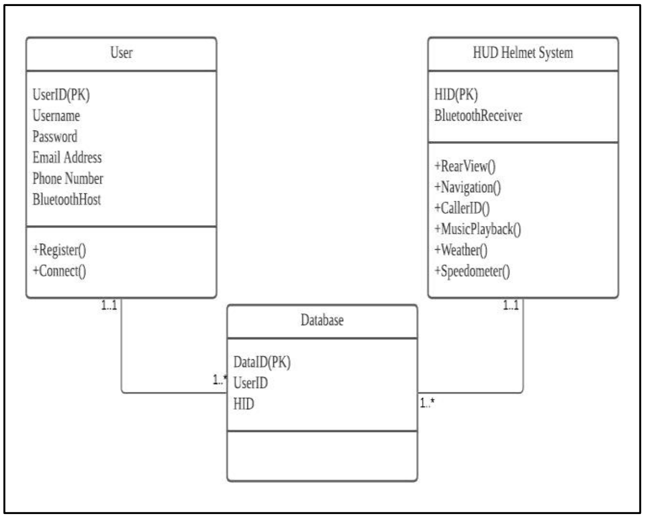


Fig. 5. Class diagram.

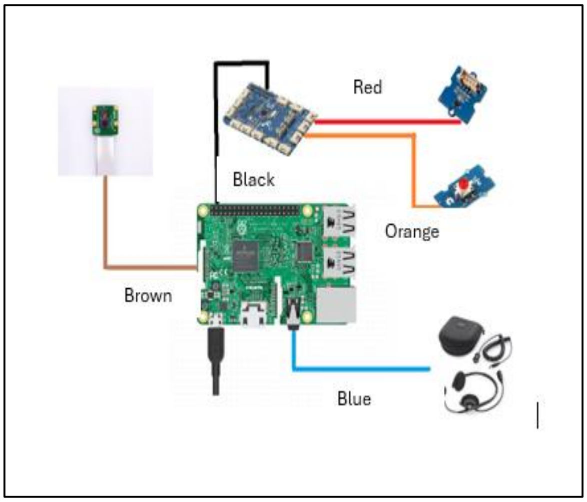


Fig. 6. Schematic layout of the proposed HUD helmet system.

C. Implementation Set-up

A Raspberry Pi, an Android app, and a database were employed as development platforms. Python and a JavaScript file for the LCI-HUD system were used to execute GeanyIED. Using the Pixel 2 as a reliable test bench, Android Studio emulates Android 8.0. Using a mobile app, Google Firebase stores authentication and databases. The users/rider need to register before using the application.

Module design is done in the Geany IDE. The framework of the HUD helmet system is written in Python using the Geany IDE. All the essential tools required for debugging and running code are included. Developers may use Android Studio to create phone apps that function across a variety of devices and run on the Android operating system. Using Firebase DB, the primary software feature, apps linked to the Raspberry Pi (HUD Helmet) can update the database in real time, enabling the Internet of Things. The system begins to communicate with and store information about the user as soon as they register for the first time on the mobile application. Following that, users run the system via Bluetooth. For security purposes, only the user has access to a unique authentication token.

The main objective of the system's implementation and appropriate functioning are its functional prerequisites. The non-functional need focused on the five main elements, along with other demands that were related to the behavior of the system's interactions, or the time required to finish a specific task. Usability, security, accessibility, user engagement, and responsiveness are the non-functional requirements for the suggested HUD Helmet System. The responsiveness criterion makes ensuring that user inputs are processed fast and effectively, and that real-time feedback is provided. The criterion for accessibility makes sure that users with impairments or disabilities can navigate the system with ease. Providing a smooth and simple user experience that makes it easy for users to explore and interact with the system's capabilities is the main goal of the user interaction requirement. The system is safeguarded against unauthorized access and data breaches thanks to the security requirement. Making the system simple to use and intuitive to understand is the main goal of the usability requirement.

III. Results and Discussion

Fig. 7 illustrates the three "modules" that the user of the LCI-HUD App can access on the first page after setup which includes the clock, weather, and rear-view camera. Since it is always necessary for the user to see the blind spot, the rear-view camera module remains stationary. By doing this, the user can continuously monitor the stream from the rear-view camera without ever leaving the first page. The clock and weather modules can also be customized by the user to show information that they want, like the location of weather updates or the format of the time.

The user can choose, edit, or remove data for numerous HUD helmet devices on the left page, as shown in Fig. 8. Users can flip between pages and change the brightness on the right page of the HUD helmet. The right page of the HUD helmet allows for simple page switching, offering a fluid and adaptable experience. Additionally, the brightness control feature guarantees the best possible visibility under different lighting circumstances.

When the user clicks on the red button as shown in Fig. 7. That will be simulated installed in the handler bar of the motorbike this will scroll to the next page as shown in Fig. 9 which at the moment showing the Google map route and music. User are able to expand on this by customizing it for which other widgets that they wish to view at a glance from the LCI-HUD App.

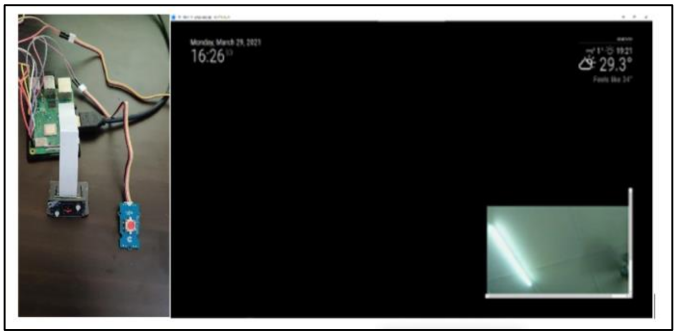
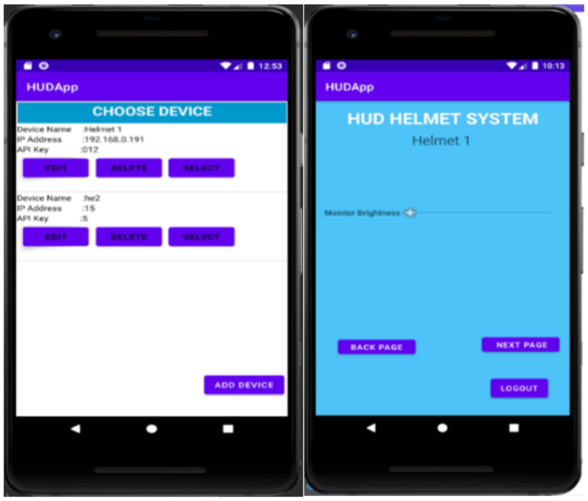


Fig. 7. LCI-HUD System.



HUDApp

CHOOSE DEVICE

Device Name :Helmet 1
IP Address :192.168.0.191
API Key :012

Device Name :he2
IP Address :15
API Key :5

HUD HELMET SYSTEM

Helmet 1

Mointor Brightness

Fig. 8. LCI-HUD App.

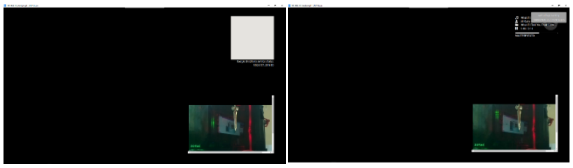


Fig. 9. LCI-HUD Different pages.

IV. Conclusion and Future Enhancement

The need for IoT devices that bikers could use daily motivated the development of the LCI-HUD. In the Asian market, HUD helmets are extremely expensive and of inferior quality. Additionally, helmets that come with HUDs rarely offer much functionality. The HUD Helmet System application, which has a working prototype included in the LCI-HUD, integrates the Raspberry Pi (Python and JavaScript) with Android Studio. The LCI-HUD was intended to be modular since it is easy to swap out any section with an improved version in the future that incorporates third-party APIs to enable user customization. The LCI-HUD is easily upgraded or replaced by users, due to its modular design, which keeps it up to date with new technologies. Furthermore, the incorporation of third-party APIs creates countless opportunities for user customization and feature growth beyond what is initially provided. When comparing the software features and safety offerings in cars equipped with Google's Android Auto and Apple's CarPlay, it becomes evident that there is significant potential for future growth in integrating similar technologies for motorcycles. Currently, there is a noticeable gap in support for these features and safety measures tailored specifically for bikers. Overall, the LCI-HUD's modular design and integration with third-party APIs make it a versatile and future-proof solution for users looking to customize their experience. Using the ability to easily upgrade and replace components, users can stay ahead of the curve in technological advancements and this reduces e-waste as it promotes backward compatibility and enhancing existing helmet to retrofit with LCI-HUD.

References

[1] Witsaman, D. (2016). Motorcycle Helmet Crash Detection and Prevention System.

[2] He, Z., Xiong, J., Ng, T., Fan, B., & Shoemaker, C. (2017). Managing competitive municipal solid waste treatment systems: An agent-based approach. European Journal of Operational Research, 263(3), 1063-1077.

[3] Jacobson, P.D., and Gostin, L.O., 2010. Reducing distracted driving: regulation and education to avert traffic injuries and fatalities JAMA, 303(14), pp. 1419–1420.

[4] Chang, W.J., and Chen, L.B., 2019. Design and implementation of an intelligent motorcycle helmet for large vehicle approach intimation. IEEE Sensors Journal, 19(10), pp. 3882–3892.

[5] Ahmed, S. U., Uddin, R., & Affan, M. (2020, September). Intelligent gadget for accident prevention: smart helmet. In 2020 International Conference on Computing and Information Technology (ICCIT-1441) (pp. 1-4). IEEE.

[6] Choi, Y., & Kim, Y. (2021). Applications of smart helmet in applied sciences: a systematic review. Applied Sciences, 11(11), 5039.

[7] Grahn, H., and Kujala, T., 2018, October. Visual Distraction Effects between In-Vehicle Tasks with a Smartphone and a Motorcycle Helmet-Mounted Head-Up Display In Proceedings of the 22nd International Academic Mindtrek Conference (pp. 153–162).

[8] Kircher, K., & Ahlstrom, C. (2017). Minimum required attention: A human-centered approach to driver inattention. Human factors, 59(3), 471-484.

[9] Häuslschmid, R., Fritzsche, B., and Butz, A., 2018, March. Can a helmet-mounted display make motorcycling safer? 23rd International Conference on Intelligent User Interfaces (pp. 467–476).

[10] Skirnewskaja, J. (2024). High-Definition Holographic Head-Up Displays (Doctoral dissertation).

[11] Okey, O. D., Maidin, S. S., Adasme, P., Lopes Rosa, R., Saadi, M., Carrillo Melgarejo, D., & Zegarra Rodríguez, D. (2022). BoostedEnML: Efficient technique for detecting cyberattacks in IoT systems using boosted ensemble machine learning. Sensors, 22(19), 7409.

[12] Gokhale, P., Bhat, O., & Bhat, S. (2018). Introduction to IOT. International Advanced Research Journal in Science, Engineering and Technology, 5(1), 41-44.

[13] Mohammad, M., Pagkale, P. J., Abd Rahman, N. F., & Shariff, M. S. M. (2022). Hydrological Safety of Vaturu Dam by Evaluating Spillway Adequacy. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 21, 349-355.

A Robust Wrapper-Based Feature Selection Technique Using Real-Valued Triangulation Topology Aggregation Optimizer

Li Pan[1], Wy-Liang Cheng[2]*, Sew Sun Tiang[3], Kim Soon Chong[4],
Chin Hong Wong[5], Abhishek Sharma[6], Touseef Sadiq[7], Aasam Karim[8], Wei Hong Lim[9*]
Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, 56000, Malaysia[1, 2, 3, 4, 9]
Zhengzhou Institute of Engineering and Technology, Zhengzhou, 450044, China[1]
Maynooth International Engineering College, Maynooth University, Maynooth, Co Kildare, Ireland[5]
Maynooth International Engineering College, Fuzhou University, Fujian University, 350116, China[5]
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, 248002, India[6]
Centre for Artificial Intelligence Research (CAIR)-Department of Information and Communication Technology,
University of Agder, Grimstad, Norway[7]
Instytut Informatyki Uniwersytet Opolski 45-040 Opole Poland[8]

***Abstract*—Feature selection is a critical preprocessing technique used to remove irrelevant and redundant features from datasets while maintaining or improving the accuracy of machine learning models. Recent advancements in this area have primarily focused on wrapper-based feature selection methods, which leverage metaheuristic search algorithms (MSAs) to identify optimal feature subsets. In this paper, we propose a novel wrapper-based feature selection method utilizing the Triangulation Topology Aggregation Optimizer (TTAO), a newly developed algorithm inspired by the geometric properties of triangular topology and similarity. To adapt the TTAO for binary feature selection tasks, we introduce a conversion mechanism that transforms continuous decision variables into binary space, allowing the TTAO—which is inherently designed for real-valued problems—to function efficiently in binary domains. TTAO incorporates two distinct search strategies, generic aggregation and local aggregation, to maintain an effective balance between global exploration and local exploitation. Through extensive experimental evaluations on a wide range of benchmark datasets, TTAO demonstrates superior performance over conventional MSAs in feature selection tasks. The results highlight TTAO's capability to enhance model accuracy and computational efficiency, positioning it as a promising tool to advance feature selection and support industrial innovation in data-driven tasks.**

Keywords—Classification; exploration; exploitation; feature selection; metaheuristic search algorithm; machine learning; optimization; triangulation topology aggregation optimizer

I. Introduction

Recent advancements in data-driven methodologies such as machine learning approaches have demonstrated significant benefits in addressing diverse, complex real-world challenges. Nonetheless, the escalating complexity of datasets from various domains and sources increasingly burdens machine learning models with inefficiencies and elevated computational costs [1]. This burden is often exacerbated by the presence of excessive, irrelevant, and redundant features, particularly when datasets are laden with noise, inconsistencies, and non-contributory information. Such datasets do not enhance model performance and may even compromise system approximation accuracy due to overfitting. Additionally, the presence of a large number of features necessitates the use of more complex machine learning models. These models require extensive data to optimize learning parameters, which can degrade their ability to generalize effectively [1].

To address these challenges, it becomes crucial to select a relevant subset of features while eliminating redundancies from the original datasets. This process not only enhances the efficiency, accuracy, and generalization capability of machine learning models but also serves as a vital preprocessing strategy. Feature selection effectively improves model performance by minimizing redundant input during training. It aims to optimize model efficiency by identifying and employing an optimal subset of features, thus alleviating the detrimental effects associated with the “curse of dimensionality”. This is particularly important when dealing with input datasets that contain an excessively large number of primitive features. Feature selection is instrumental in addressing a wide range of real-world machine-learning challenges, such as food fraud detection, automatic modulation recognition, predictive maintenance, robot path planning, kitchen waste segregation etc. [2-8].

Feature selection techniques are primarily divided into three categories: filter, wrapper, and embedded methods [9]. Filter methods determine feature subsets using statistical techniques that evaluate data dependencies. These methods assign rankings based on criteria such as inter-feature distances, correlations, and consistency indices. Widely utilized filter methods include the correlation coefficient, F-score, and Gini index [10]. Although filter methods are computationally efficient due to their classifier independence, they often do not reflect true feature relevance within specific models, potentially leading to reduced predictive accuracy. Conversely, wrapper methods integrate with a specific classifier, employing classification accuracy to assess feature subset quality [11]. While these methods typically enhance classifier performance, they also increase the risk of overfitting and require extensive computational resources due to repeated classifier executions to ascertain the optimal subset. Embedded methods, merging the benefits of filter and wrapper approaches, interact directly with the classifier while managing dependencies more efficiently to reduce computational demands. Although embedded methods offer a compromise in computational load, they remain more resource-intensive than filter methods.

The feature selection technique proposed in this paper is classified as wrapper-based, typically involves three components [9]: classifiers employed, evaluation criteria for feature selection, and the search algorithms used to derive feature subsets from raw data. Conventional search strategies [11] such as backward elimination, forward selection, greedy search, and complete search often exhibit significant limitations within the wrapper-based framework. These limitations include poor global search abilities, entrapment in local optima, and high computational costs. To address these deficiencies, this study advocates the use of metaheuristic search algorithms (MSAs), which offer superior global search strength, stochastic behavior, simple implementation, and do not rely on gradient information, making them well-suited to address complex optimization challenges [12-16]. A review of how MSAs effectively overcome feature selection challenges is provided [9].

MSAs represent a varied collection of optimization techniques, categorized by their foundational inspirations and search mechanisms [17]. The first category of MSAs is the evolutionary algorithms that are influenced by Darwin's theory of evolution and natural selection. Swarm intelligence algorithms are the second category of MSAs and they are inspired by collective animal behaviors such as flocking and foraging. Human-based algorithms are the third category of MSAs and they mimic aspects of human cognition including learning and social interactions, whereas the last category of MSAS are physics-based algorithms that apply principles from physical sciences and mathematics. Although numerous MSAs have been developed in response to the No-Free-Lunch (NFL) theorem, which asserts that no single algorithm can optimally solve all types of problems, their validation has predominantly been confined to mathematical benchmarks.

While significant theoretical advancements have been made in the development of MSAs, their performance evaluations remain largely confined to theoretical benchmarks. This narrow focus limits our understanding of their practical effectiveness in solving real-world problems, emphasizing the need for more empirical studies. In particular, the practical application of many recently developed MSAs in addressing real-world optimization challenges, such as feature selection, remains insufficiently explored. Moreover, most MSAs have not been rigorously validated in complex, high-dimensional feature selection tasks involving binary decision variables. This gap highlights the pressing need for empirical research that assesses the performance of novel MSAs in feature selection tasks, extending beyond traditional continuous-variable optimization problems.

This paper introduces an advanced wrapper-based feature selection method leveraging the unique search mechanisms of the Triangulation Topology Aggregation Optimizer (TTAO), a novel physics-based MSA proposed by Zhao et al. in 2024 [18]. Inspired by the geometric properties of triangular topology and the principle of triangular similarity, TTAO utilizes the consistent shape but variable sizes of similar triangles to generate diverse triangular topological units that serve as dynamic evolutionary entities throughout the optimization process. This technique aims to enhance the performance of machine learning models by effectively eliminating irrelevant features from datasets. TTAO incorporates two primary aggregation strategies: generic aggregation and local aggregation. Generic aggregation enhances exploratory search by promoting information exchange across different triangular topological units, whereas local aggregation focuses on exploitation, refining the search within individual units. Although initially applied in limited real-world contexts such as transmission expansion planning [19], productivity prediction [20], and controller parameter adjustment [21], where decision variables are real-valued, the application of TTAO to feature selection tasks involving binary decision variables is an unexplored area of research. This study aims to fill this gap by demonstrating how TTAO can be adapted to binary feature selection, presenting a novel conversion mechanism that enables its application in this domain. By expanding TTAO’s utility to feature selection tasks with binary decision variables, this paper contributes to addressing the broader challenge of validating MSAs in real-world optimization problems.

The technical contributions and novelty of this study are summarized as follows:

- We propose an advanced wrapper-based feature selection technique that utilizes the unique search mechanisms of the TTAO to identify optimal feature subsets. This approach aims to achieve high classification accuracy while maintaining low model complexity.
- To our knowledge, this is the first application of TTAO to address feature selection problems involving binary decision variables, which present more complex optimization challenges compared to those with continuous variables.
- A novel conversion mechanism is introduced, transforming continuous decision variables into binary ones, thus adapting the inherently real-valued TTAO for use in binary solution spaces.
- We provide a comprehensive evaluation of TTAO’s effectiveness as a wrapper-based method for feature selection, demonstrating its superior performance against other MSAs using diverse datasets from the UCI Machine Learning Repository.

The remainder of this paper is organized as follows: Section II reviews related work. Section III outlines the formulation of wrapper-based feature selection as an optimization problem and details the search mechanisms of TTAO. Section IV presents performance evaluations of various wrapper-based feature selection techniques. Section V concludes with a summary and future works.

II. Related Works of Using Different MSAs for Wrapper-Based Feature Selection

A wrapper-based feature selection technique incorporates three core components: the classifier types, the search algorithms for discovering optimal feature subsets, and the criteria for assessing the quality of these subsets. MSAs are often favored for wrapper-based feature selection due to their robust global search capabilities and straightforward implementation, as documented in study [9]. These MSAs are particularly effective in identifying feature subsets that optimize classification accuracy while minimizing the complexity of the machine learning model.

Recent developments in MSAs have significantly enhanced the robustness of feature selection methodologies. Various novel MSAs such as the Flow Direction Algorithm [22], African Vultures Optimization Algorithm [23], Sperm Swarm Optimization [24], Grasshopper Optimization Algorithm [25], Artificial Butterfly Optimization [26], have been employed to tackle feature selection challenges. Researchers also continue to refine these algorithms, creating more efficient versions tailored to specific problem characteristics. For example, Zekeri and Hokmabadi [11] introduced a real-value Grasshopper Optimization Algorithm (GOFS), utilizing a mathematical model that leverages repulsion and attraction forces between grasshoppers to effectively navigate the feature space. They enhanced GOFS with an adaptive parameter that modifies the influence zones to improve feature exploration and exploitation. Additionally, they implemented a feature probability factor to eliminate redundant features each iteration. Mostafa et al. [27] developed a Modified Chameleon Swarm Algorithm (mCSA), incorporating a transfer operator and a randomization Levy flight control parameter to fine-tune search behaviors. They also hybridized mCSA with the consumption operator from Artificial Ecosystem-based Optimization to augment its global search capabilities.

Zhang et al. [10] developed a novel wrapper-based feature selection method utilizing the Return-Cost-Based Binary Firefly Algorithm (Rc-BBFA), enhanced with three key modifications to address premature convergence. This version replaces traditional distance-based attractiveness with a return-cost metric to gauge each firefly's appeal. Additionally, a Pareto dominance strategy selects the most attractive firefly based on cost and return values. A new binary movement operator, driven by return-cost attractiveness and supplemented by an adaptive jump, updates each firefly's position within Rc-BBFA. Ma et al. [28] introduced the Multi-Strategy Binary Hunger Games Search (MS-bHGS) to tackle feature selection across 20 benchmark datasets. MS-bHGS incorporates chaotic maps, a vertical crossover scheme, and a greedy selection strategy, enhancing the balancing of exploration and exploitation. Wu et al. [29] enhanced a wrapper-based feature selection method using the Sparrow Search Algorithm, augmented by Quantum Computation and Multi-Strategy Enhancement (QMESSA). This approach integrates an improved circle chaotic map with a quantum gate mutation mechanism to diversify the initial population. Adaptive T-distribution and a novel position update formula were also embedded in QMESSA to boost its convergence speed. capabilities.

Zhong et al. [30] introduced the Self-Adaptive Quantum Equilibrium Optimizer with Artificial Bee Colony (SQEOABC) for feature selection, incorporating quantum theory and a self-adaptive mechanism to improve its convergence. Additionally, SQEOABC utilizes updating mechanisms from the Artificial Bee Colony to enhance the selection of effective feature subsets. Khafaga et al. [31] proposed a novel wrapper-based feature selection method using the Adaptive Squirrel Search Optimization Algorithm (ASSOA), paired with a KNN classifier. This method was applied to ten datasets from the UCI Machine Learning Repository. ASSOA was enhanced with new relocation equations and various movements (vertical, horizontal, exponential, and diagonal) to improve its search capabilities. Furthermore, various feature selection techniques were advanced by combining the Dipper Throated Optimization Algorithm with the Grey-World Optimizer [32] and Sine Cosine Algorithm [33]. These hybrid methods were tailored to identify superior feature subsets, contributing to higher accuracy and reduced model complexity in handling publicly available datasets.

Image Analysis Society (MIAS), the selected features were evaluated using the XGBoost classifier. In a follow-up study, they developed an adaptive binary TLBO with an ensemble classifier combining XGBoost and Random Forest, aimed at the early detection of breast cancer using mammograms from MIAS and the Digital Database for Screening Mammography (DDSM) [23].

III. Wrapper-Based Feature Selection Using TTAO

A. *Solution Representation of TTAO in Feature Selection*

In the context of feature selection, consider a dataset where $|F_o|$ denotes the total number of input features. Within the framework of TTAO, each search agent or vertex of the n-th triangular topological unit is defined by a position vector $X_n = [X_{n,1}, \ldots, X_{n,d}, \ldots, X_{n,D}]$, with D equating to $|F_o|$, representing the dimensionality of the problem. Each dimensional index d corresponds directly to a feature index l. Initially, the decision variables for each search agent are continuous. However, the binary nature required for feature selection dictates that these variables must be converted to binary values – 0 or 1.

To facilitates this conversion, the proposed wrapper-based feature selection technique based on TTAO implements a threshold parameter γ. This parameter is used to transform continuous decision variables into binary decisions by evaluating each real-valued decision variable $X_{n,d}$ against γ:

$$S_{n,l} = \begin{cases} 0, & \text{if } X_{n,d} < \gamma \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Here, the binary value $S_{n,l}$ determines the inclusion status of each *l*-th feature, where a value of 1 indicates inclusion and 0 indicates exclusion. For example, a status of $S_n = [0,1,1,1,0]$ implies that features at indices *l* = 2 to 4 are selected, while those at indices *l* = 1 and 5 are excluded. This mechanism effectively transforms continuous input values encoded in the search agent into the discrete decisions crucial for effective feature selection.

B. Fitness Evaluation of TTAO Search Agent in Feature Selection

Feature selection plays a crucial role in machine learning by facilitating the identification of an optimal subset of features. This subset not only enhances classification accuracy but also reduces the numbers of utilized features, addressing a twofold challenge: lowering the classifier's error rate and minimizing the ratio of selected features to the total available.

Define ξ_{Error} as the classifier's error rate and $|F_s|$ as the count of features selected for the subset, with $|F_s| \leq |F_o|$. The fitness value, which assess the quality for each search agent of the *n*-th triangular topological unit via the feature status vector $S_n = [S_{n,1}, \dots, S_{n,l}, \dots, S_{n,|F_o|}]$, is given by:

$$F(X_n) = \omega \times \xi_{Error} + \mu \times \frac{|F_s|}{|F_o|} \tag{2}$$

Here, ω is a coefficient ranging from 0 to 1, and μ is defined as $1 - \omega$. These parameters are designed to weigh the impacts of classification error and feature proportionality, respectively. The optimal feature subset minimizes the fitness function outlined in Eq. (2), achieving a balance between high classification accuracy and reduced feature set complexity, thereby simplifying and enhancing the efficacy of machine learning models.

In the wrapper-based feature selection framework using TTAO, the fitness evaluation process, denoted as Algorithm 1, employs the KNN classifier to measure each n-th search agent's performance based on the feature selection status S_n. Feature normalization is applied to scale the selected features between 0 and 1, followed by performance evaluation using K-fold cross-validation with the KNN classifier. A lower $F(X_n)$ value signifies superior fitness, indicative of higher classification accuracy and a smaller number of selected features.

Algorithm 1: Fitness Evaluation Process of Wrapper-Based Feature Selection Using TTAO

```
Inputs: $X_n, F_o, D, \gamma$
    Convert $X_n$ into $S_n$ using Eq. (1);
    Determine $|F_s|$ from $S_n$ and train KNN classifier to get $\xi_{Error}$;
    Calculate $F(X_n)$ using Eq. (2) based on $|F_s|$ $and$ $\xi_{Error}$;
Outputs: $F(X_n)$
```

C. Mechanisms of TTAO to Identify Optimal Feature Subsets

1) Conceptual ideas of TTAO: The search mechanisms of TTAO draw inspiration from the fundamental properties of triangular topology in mathematics. The triangle, recognized as the most basic yet stable shape in planar geometry, serves as a cornerstone in both finite and infinite dimensional spaces. It functions as a graph within its two-dimensional subspace. Due to its inherent simplicity and robustness, the triangular topology is extensively employed as a structural unit in model representation and analysis across a variety of real-world applications. These applications span multiple disciplines, including computational geometry, structural engineering, digital image processing, etc.

The concept of triangular similarity is pivotal in geometry and plays a key role in the search mechanisms of TTAO. The principles of triangular similarity are covered in four theorems:

- Theorem 1: A new triangle formed by drawing a line parallel to one side of an original triangle and intersecting the extensions of the other two sides is similar to the original.
- Theorem 2: Two triangles are similar if their corresponding sides and angles are proportional.
- Theorem 3: A triangle is similar to another if the ratios of their corresponding sides are equal.
- Theorem 4: Triangles that have identical corresponding angles are similar.

TTAO employs these theorems of triangular similarity to direct its search strategy. Throughout its iterative search process, the algorithm continuously generates new vertices in the solution space to construct triangles of varying sizes, each considered an evolutionary unit with three external vertices and one internal random vertex. Additionally, the TTAO utilizes the concept of aggregation to merge vertices with superior traits, enhancing the information exchange within and across different topological units. All triangles within the TTAO framework are equilateral, maintaining geometric consistency by adhering to the second theorem of similarity. The optimization process of TTAO consists of two primary stages: aggregation between and within units, streamlining the exploration and exploitation phases.

2) Initialization phase of TTAO: The initialization phase of TTAO involves randomly generating a diverse set of potential solutions across the solution space. Let N and D represent the population size and problem dimensionality of TTAO, respectively. Each vertex within a triangular topological unit is treated as a search agent or potential solution. Using the floor rounding operator $\lfloor \cdot \rfloor$, the population set of N search agents is organized into $\lfloor N/3 \rfloor$ triangular topological units. Any additional search agents, arising when N is not divisible by 3, are randomly generated within the solution space.

The lower and upper boundary limits of decision variables are denoted as $X^L = [x_1^l, \dots, x_d^l, \dots, x_D^l]$ and $X^U = [x_1^u, \dots, x_d^u, \dots, x_D^u]$, respectively. Let r_0 be a random number between 0 and 1. For each *n*-th triangular topological unit, where $n = 1, \dots, \lfloor N/3 \rfloor$, the position of the first search agent (vertex) is randomly determined within the feasible regions of solution space as follows:

$$X_{n,1} = X^L + r_0(X^U - X^L) \tag{3}$$

3) Construction of triangular topological unit: In addressing multi-dimensional optimization challenges, TTAO constructs equilateral triangles within each two-dimensional projection of a higher-dimensional space. The TTAO leverages transformations between polar and Cartesian coordinate systems to establish the vertices of each triangular topological unit.

For every *n*-th triangular topological unit, a direction vector, denoted as $lf(\cdot)$, is calculated and applied to the first vertex ($X_{n,1}$) to determine the second vertex ($X_{n,2}$) as follows:

$$X_{n,2} = X_{n,1} + lf(\theta) \tag{4}$$

The third vertex $(X_{n,3})$ is then generated by rotating the direction vector $lf(\cdot)$ by $\pi/3$ radians anticlockwise:

$$X_{n,3} = X_{n,1} + lf(\theta + \pi/3) \tag{5}$$

Here, l signifies the length of the edges of the triangular topology unit, given by:

$$l = 9e^{-\frac{t}{T^{max}}} \tag{6}$$

where t is the current iteration numbers, and T^{max} represents the maximum iteration numbers. According to Eq. (6), l decreases as the number of fitness evaluations increases. This adaptive strategy enables broader exploratory moves in the initial stages and more focused exploitation in the latter phases to refine the search in promising regions. The exponential decay ensures l remains positive, preventing excessive exploitation and potential premature convergence.

Moreover, the vectors $f(\theta)$ and $f(\theta + \pi/3)$, directing the edges from the first vertex, are defined respectively as:

$$f(\theta) \ = [\cos\theta_1, \dots., \cos\theta_d, \dots, \cos\theta_D] \tag{7}$$

$$f(\theta + \pi/3) = [\cos(\theta_1 + \pi/3), \dots.,$$

$$\cos(\theta_d + \pi/3), \dots, \cos(\theta_D + \pi/3)] \tag{8}$$

where θ_d for $d = 1, \dots, D$ is a randomly generated angle ranging from 0 to π.

Within each n-th triangular topological unit, a fourth vertex $X_{n,4}$ is derived through an internal aggregation process using a linear combination of the first three vertices, weighted by randomly generated coefficients:

$$X_{n,4} = r_1X_{n,1} + r_2X_{n,2} + r_3X_{n,3} \tag{9}$$

where r_1, r_2 and r_3 are randomly numbers between 0 to 1, ensuring $r_1 + r_2 + r_3 = 1$.

In each iteration, a new triangular topological unit is generated from a vertex and two sides of equal lengths l, which dynamically change throughout optimization process. Within each n-th triangular topological unit, the vertex exhibiting the best fitness during the current iteration is designated as the lead vertex. This lead vertex plays a crucial role in guiding the search process of the other vertices within the same unit. As detailed in subsequent sections, vertices within and across different triangular topological units employ two pivotal search mechanisms: generic aggregation and local aggregation. These mechanisms enable exploration and exploitation, respectively.

4) Generic aggregation of TTAO: Generic aggregation facilitates exploration by enabling the information exchange between the best search agent (vertex) in each triangular topological unit and the best vertex from a randomly selected unit. This mechanism draws inspiration from the crossover operator in genetic algorithm, which creates a new offspring solution by merging genetic information from two parent solutions.

Let $X_{n,best}^t$ denote the best vertex of the n-th triangular topological unit at iteration t, and $X_{n_{rand},best}^t$ represents the best vertex from a randomly selected unit at the same iteration. For each n-th triangular topological unit, a new vertex $X_{n,new1}^{t+1}$ is generated through generic aggregation by linearly combining the dimensional variables of these two superior vertices with different weights:

$$X_{n,new1}^{t+1} = r_4X_{n,best}^t \ + (1 - r_4)X_{n_{rand},best}^t \tag{10}$$

where r_4 is a random number between 0 to 1.

The fitness value of the newly generated vertex $X_{n,new1}^{t+1}$ is evaluated as $F(X_{n,new1}^{t+1})$, and compared against the fitness values of the current optimal and suboptimal vertices in the n-th triangular topological unit, represented as $F(X_{n,best}^t)$ and $F(X_{n,sbest}^t)$, respectively. Here, the suboptimal vertex $X_{n,sbest}^t$ is defined as the search agent with the second-best fitness in the n-th unit. For minimization problems, updates to the optimal and suboptimal vertices of n-th triangular topological unit for the subsequent iteration $(t + 1)$ are made according to the conditions below:

$$X_{n,best}^{t+1} = \begin{cases} X_{n,new1}^{t+1}, & \text{if } F\left(X_{n,new1}^{t+1}\right) < F(X_{n,best}^t) \\ X_{n,best}^t, & \text{otherwise} \end{cases} \tag{11}$$

$$X_{n,sbest}^{t+1} = \begin{cases} X_{n,new1}^{t+1}, & \text{if } F\left(X_{n,new1}^{t+1}\right) < F(X_{n,sbest}^t) \\ X_{n,sbest}^t, & \text{otherwise} \end{cases} \tag{12}$$

5) Local aggregation of TTAO: Local aggregation within the TTAO is pivotal for exploitation, refining searches within promising areas previously identified by the generic aggregation's exploratory processes. This strategy operates within each triangular topological unit, optimizing based on the best available internal information to enhance solution quality. Following generic aggregation, a temporary triangular topological unit is formed among the updated optimal or suboptimal vertex and two other vertices with relatively good fitness. Notably, this temporary unit may not necessarily form an equilateral triangle.

Within each n-th triangular topological unit, the optimal vertex's position is locally perturbed to refine the vicinity around the best current solution, based on the differences between the optimal and suboptimal vertices, thus ensuring the new search direction leverages the promising information. The new vertex generated through local aggregation is given by:

$$X_{n,new2}^{t+1} = X_{n,best}^{t+1} \ + \alpha X_{n,best}^{t+1} \tag{13}$$

where α is a decreasing parameter regulating the local aggregation's scope, defined as:

$$\alpha = \ln\left(\frac{e - e^3}{T^{max}{}_{-1}}t + e^3 - \frac{e - e^3}{T^{max}{}_{-1}}\right) \tag{14}$$

The parameter α progressively narrows the search area across iterations to emphasize exploitation in algorithm's later stages.

After local aggregation, it is crucial that the lead vertex of the triangular topological unit is the optimal within that unit. To assure convergence towards the most promising directions, the fitness of the newly aggregated vertex $X_{n,new2}^{t+1}$, denoted as $F(X_{n,new2}^{t+1})$, is compared against the current optimal vertex's fitness $F(X_{n,new1}^{t+1})$. For minimization problems, the updates of optimal vertex during local aggregation is determined as:

$$X_{n,best}^{t+1} = \begin{cases} X_{n,new2}^{t+1}, & \text{if } F\left(X_{n,new2}^{t+1}\right) < F(X_{n,best}^{t+1}) \\ X_{n,best}^{t+1}, & \text{otherwise} \end{cases} \quad (15)$$

Following this update, new similar triangular topological units are constructed for the subsequent iteration based on these updated optimal vertices, employing Eqs. (4) to (9) to refine the search space further and focus on previously promising areas.

6) Optimization workflow of TTAO for wrapper-based feature selection: Algorithm 2 delineates the workflow of the proposed TTAO-based wrapper feature selection technique. The process commences by loading the dataset and establishing dimensionality D equal to the number of input features, $|F_o|$. Initial setting of TTAO include resetting the iteration counter *t* and determining the number of triangular topological units as $\lfloor N/3 \rfloor$.

To deploy TTAO in searching for optimal feature subsets, initial positions for the first vertices, $X_{n,1}$ for $n = 1, \dots, \lfloor N/3 \rfloor$, of all triangular topological units are randomly generated as per

Algorithm 2: Proposed Wrapper-Based Feature Selection Using TTAO

```
Inputs: $N, \gamma, T^{Max}, F_o$
Load dataset containing $|F_o|$ input features and set the total dimensional size as $D \leftarrow |F_o|$;
Initialize $t \leftarrow 0$, number of triangular topological unit $\leftarrow \lfloor N/3 \rfloor$;
for n = 1 to $\lfloor N/3 \rfloor$ do   /*Random initialization of the first vertices*/
    Randomly generate the first vertex $X_{n,1}$ of each n-th triangular topology unit using Eq. (3);
end for
while $t \leq T^{Max}$ do   /*Iterative search process*/
    Update the value of parameter l using Eq. (6);
    /*Construction of each n-th triangular topological unit*/
    for n = 1 to $\lfloor N/3 \rfloor$ do
        Determine $f(\theta)$ and $f(\theta + \pi/3)$ using Eqs. (7) and (8), respectively;
        Calculate second vertex $X_{n,2}$ of each n-th triangular topology unit using Eq. (4);
        Calculate third vertex $X_{n,3}$ of each n-th triangular topology unit using Eq. (5);
        Boundary checking for $X_{n,2}$ and $X_{n,3}$ to ensure solution feasibility;
        Calculate fourth internal vertex $X_{n,4}$ of each n-th triangular topology unit using Eq. (9);
        Boundary checking for $X_{n,4}$ to ensure solution feasibility;
        Fitness evaluation of all vertices (i.e., $X_{n,1}$ to $X_{n,4}$) using Algorithm 1;
        Identify the vertices with best and second-best fitness as $X_{n,best}^{t}$ and $X_{n,sbest}^{t}$, respectively.
    end for
    /*Generic aggregation*/
    for n = 1 to $\lfloor N/3 \rfloor$ do
        Calculate new vertex $X_{n,new1}^{t+1}$ of each n-th triangular topology unit using Eq. (10);
        Boundary checking for $X_{n,new1}^{t+1}$ to ensure solution feasibility;
        Fitness evaluation of $X_{n,new1}^{t+1}$ using Algorithm 1;
        Update $X_{n,best}^{t+1}$ and $X_{n,sbest}^{t+1}$ along with their fitness using Eqs. (11) and (12), respectively;
    end for
    /*Local aggregation*/
    for n = 1 to $\lfloor N/3 \rfloor$ do
        Update the value of parameter $\alpha$ using Eq. (14)
        Calculate new vertex $X_{n,new2}^{t+1}$ of each n-th triangular topology unit using Eq. (13);
        Boundary checking for $X_{n,new2}^{t+1}$ to ensure solution feasibility;
        Fitness evaluation of $X_{n,new2}^{t+1}$ using Algorithm 1;
        Update $X_{n,best}^{t+1}$ along with its fitness using Eq. (15);
    end for
    /*To check if the population size N is divisible by 3*/
    if $N - \lfloor N/3 \rfloor \neq 0$ then
        $N^{Remain} = N - \lfloor N/3 \rfloor$;
        for i = 1 to $N^{Remain}$ do
            Randomly generate the i-th remaining search agent $X_i^{Remain}$ using Eq. (3);
            Fitness evaluation of $X_i^{Remain}$ using Algorithm 1;
        end for
        Compare the fitness value of $X_{n,best}^{t+1}$ for n = 1 to $\lfloor N/3 \rfloor$ and $X_i^{Remain}$ for i = 1 to $N^{Remain}$;
        Extract the top $\lfloor N/3 \rfloor$ search agents with better fitness to be lead vertices in next iteration;
    end if
    Record the best solution $X^{Best}$ and its fitness $F(X^{Best})$ found in each iteration;
end while
Outputs: $X^{Best}$ and $S^{Best}$
```

Eq. (1). During iterative searching, subsequent vertices of each n-th triangular topological unit ($X_{n,2}$, $X_{n,3}$ and $X_{n,4}$) are constructed using Eqs. (4), (5) and (9), respectively, to form equilateral triangle. Boundary conditions for $X_{n,2}$, $X_{n,3}$ and $X_{n,4}$ are checked to maintain solution feasibility. In the fitness evaluation process using Algorithm 1, continuous decision variables in each vertex of the n-th triangular topological unit are converted into binary selection status vectors (i.e., $S_{n,1}$, $S_{n,2}$, $S_{n,3}$ and $S_{n,4}$) using Eq. (1), and their respective fitness values ($F(X_{n,1})$, $F(X_{n,2})$, $F(X_{n,3})$, $F(X_{n,4})$) are calculated using Eq. (2). Given these fitness values, the optimal and suboptimal vertices for each n-th unit at iteration t are identified as $X_{n,best}^{t}$ and $X_{n,sbest}^{t}$, respectively.

During the generic aggregation, a new vertex $X_{n,new1}^{t+1}$ is formulated for each *n*-th triangular topological unit via Eq. (10). The updates of position vector and fitness of $X_{n,best}^{t}$ and $X_{n,sbest}^{t}$ are performed using Eq. (11) and (12) if $X_{n,new1}^{t+1}$ demonstrates superior fitness. Similarly, during local aggregation, another new vertex $X_{n,new2}^{t+1}$ is generated for each *n*-th triangular topological unit using Eq. (13), with updates to $X_{n,best}^{t+1}$ conducted via Eq. (15) if $X_{n,new2}^{t+1}$ exhibits enhanced fitness. For population where *N* is not divisible by three, remaining search agents are randomly generated within the solution space per Eq. (3) and evaluated using Eq. (2). After completing both generic and local aggregations, the fitness values of the optimal vertices across all $\lfloor N/3 \rfloor$ triangular topological units and the randomly generated search agents are compared, selecting only the best $\lfloor N/3 \rfloor$ agents as lead vertices for the subsequent iteration.

This iterative search process persists until reaching the predetermined termination criterion, typically the maximum iteration number T^{max}. Upon termination, the decision variables in the best solution X^{Best} are translated into a binary feature subset S^{Best}, employed to train machine learning models that are both more accurate and less complex.

IV. Results

A. Datasets Used and Simulation Settings

This performance evaluation study employs ten benchmark datasets from the UCI Machine Learning Repository to assess the efficacy of the proposed wrapper-based feature selection technique utilizing the TTAO. These datasets were chosen based on their diverse characteristics, including varying numbers of input features, instances, and output classes, which represent a wide range of feature selection challenges. The datasets cover different problem domains such as medical diagnosis, survival analysis, signal processing, etc., providing a comprehensive assessment of the algorithm's versatility and robustness.

The number of input features influences the dimensionality of the problem, which is crucial in testing the capability of the algorithm to handle high-dimensional spaces. For instance, datasets like Multiple Features (649 features) and Arrhythmia (279 features) represent high-dimensional feature selection challenges, whereas datasets such as Diabetes and Haberman’s Survival provide lower-dimensional tasks. The diversity in the number of instances, ranging from small datasets like Lung Cancer (27 instances) to larger datasets such as Maternal Health Risk (1014 instances)**,** ensures that the algorithm's performance is evaluated under varying data sizes. Additionally, the datasets include both binary and multiclass classification problems, further demonstrating the algorithm's adaptability to different problem types. Table I presents the detailed characteristics of the ten datasets, including the number of instances, features, and output classes. This diversity in datasets allows for a thorough evaluation of TTAO's performance in feature selection tasks across various real-world applications.

TABLE I. THE CHARACTERISTICS OF 10 BENCHMARK DATASETS USED IN PERFORMANCE EVALUATION

No	Dataset	No. of Instances	No. of Features	No. of Classes
DS1	Lung Cancer Data Set	27	56	10
DS2	Multiple Features Data Set	2000	649	2
DS3	Ionosphere Data Set	351	34	2
DS4	Arrhythmia Data Set	452	279	13
DS5	Echocardiogram Data Set	61	8	2
DS6	Haberman's Survival Data Set	306	3	2
DS7	Diabetes Data Set	768	8	2
DS8	Wine Data Set	178	13	3
DS9	Maternal Health Risk Data Set Data Set	1014	6	3
DS10	Zoo Data Set	101	16	7

This study evaluates the performance of the TTAO in feature selection tasks relative to seven other state-of-the-art MSAs: Bezier Search Differential Evolution (BeSD) [34], Coronavirus Herd Immunity Optimization Algorithm (CHIO) [35], Chaotic Oppositional based Hybridized Differential Evolution with Particle Swarm Optimization (CO-HDEPSO) [36], Differential Squirrel Search Algorithm (DSSA) [37], Flow Direction Algorithm (FDA) [38], Generalized Normal Distribution Optimization (GNDO) [39], and Oppositional and Social Learning with Enhanced Operator with Particle Swarm Optimization (ODSFMFO) [40]. Optimal parameters for these algorithms are adopted as per the specifications in their respective foundational publications.

To facilitate the conversion of real-valued decision variables within the TTAO and other MSAs to binary values for feature selection, the threshold parameter γ is set at 0.5. A KNN classifier with *k* = 5 is utilized to evaluate classification accuracy based on the selected feature subsets. Each dataset is split into two segments, with 80% of the instances designated as the training set and the remaining 20% as the testing set. The population size and the maximum iteration number for all MSAs are standardized at $N = 20$ and $T^{Max} = 200$, respectively. Given the stochastic nature of MSAs, each algorithm undergoes 30 simulation runs to ensure robustness in addressing the feature selection challenges across different datasets.

B. Performance Comparisons on Average Classification Accuracies

Table II presents the average classification accuracy achieved by the TTAO and seven other MSAs, each employed as wrapper-based feature selection techniques. The results reflect the mean values across 30 independent runs for each dataset, labeled as DS1 through DS10. Classification accuracy serves as a crucial validation measure, as it directly indicates how effectively the selected feature subsets enable the KNN classifier to distinguish between instances with high precision. The MSAs yielding higher average classification accuracies demonstrate superior capability in feature subset selection, contributing to improved overall performance. In this context, the MSAs that achieve the highest and second-highest accuracies are highlighted in bold and underlined, respectively, to facilitate a clear comparison. Furthermore, these accuracy results are juxtaposed with those from existing related studies to benchmark the efficacy of TTAO and verify its potential superiority in various classification tasks. This thorough comparison provides a more robust understanding of the advancements TTAO offers over previously proposed methods.

Table II shows that BeSD, ODSFMFO, and DSSA generally exhibit subpar performance when used as wrapper-based feature selection techniques across the ten benchmark datasets. The performance deficits are particularly pronounced in datasets with a large number of features or output classes. Specifically, BeSD recorded the lowest classification accuracies in three datasets (DS1, DS2, and DS6) and the second lowest in another (DS4). ODSFMFO displayed the poorest performance in two datasets (DS3 and DS4) and was second poorest in another two (DS8 and DS9). DSSA consistently ranked as having the second-worst average classification accuracy in four datasets (DS1, DS2, DS6, and DS7). Conversely, CO-HDEPSO and FDA demonstrated moderate performance, with their classification accuracies neither exceptionally high nor low across the majority of the datasets.

The wrapper-based feature selection techniques utilizing the three MSAs, including TTAO, have exhibited exemplary performance across the ten evaluated datasets. Specifically, CHIO recorded the highest average classification accuracy in two datasets (DS1 and DS5) and the second highest in four others (DS2 to DS4 and DS7). GNDO showed robust performance, achieving the highest classification accuracy in three datasets (DS2, DS5, and DS8) and the second highest in two (DS1 and DS9). However, both CHIO and GNDO demonstrated limitations in certain datasets, such as DS8 for CHIO and DS2 and DS10 for GNDO, indicating a need for improved robustness in diverse feature selection scenarios. TTAO emerged as the most effective MSA, securing the highest accuracy in eight datasets (DS1, DS3 to DS8, and DS10) and the second highest in DS9. Its superior performance, especially in datasets with a large number of features or classes (DS1, DS4, and DS10), underscores TTAO's capability to adeptly handle complex feature selection tasks prevalent in real-world applications.

C. *Performance Comparisons on Average Numbers of Selected Features*

While high classification accuracy is essential for effectively solving the given datasets, minimizing the size of the selected feature subset is equally important to prevent unnecessary complexity in the resulting machine learning models. Reducing feature subsets without compromising accuracy leads to simpler, more interpretable, and computationally efficient models, a critical goal in real-world applications. Achieving an optimal balance between classification accuracy and model simplicity is thus a fundamental validation criterion for evaluating feature selection techniques. To assess this trade-off, the average number of features selected by the KNN classifier is used as an additional performance metric in this study.

TABLE II. COMPARISON OF AVERAGE CLASSIFICATION ACCURACIES OF ALL MSAS FOR FEATURE SELECTION

No	BeSD	CHIO	CO-HDEPSO	DSSA	FDA	GNDO	ODSFMFO	TTAO
DS1	5.670E-01	**1.000E+00**	**1.000E+00**	6.130E-01	8.920E-01	*9.130E-01*	8.600E-01	**1.000E+00**
DS2	9.680E-01	*9.800E-01*	*9.800E-01*	9.710E-01	9.780E-01	**9.810E-01**	9.740E-01	9.750E-01
DS3	9.310E-01	*9.510E-01*	9.380E-01	9.290E-01	9.440E-01	9.270E-01	9.000E-01	**9.840E-01**
DS4	6.210E-01	*7.360E-01*	7.040E-01	6.620E-01	6.730E-01	6.730E-01	6.050E-01	**7.440E-01**
DS5	**1.000E+00**	**1.000E+00**	**1.000E+00**	**1.000E+00**	*9.790E-01*	**1.000E+00**	**1.000E+00**	**1.000E+00**
DS6	6.440E-01	8.280E-01	7.950E-01	7.540E-01	*8.300E-01*	8.010E-01	7.870E-01	**8.360E-01**
DS7	*7.810E-01*	*7.810E-01*	7.570E-01	7.520E-01	7.420E-01	7.700E-01	7.650E-01	**8.220E-01**
DS8	9.930E-01	9.310E-01	*9.990E-01*	9.560E-01	9.540E-01	**1.000E+00**	9.460E-01	**1.000E+00**
DS9	**7.680E-01**	7.370E-01	7.460E-01	7.430E-01	7.500E-01	*7.670E-01*	7.270E-01	*7.670E-01*
DS10	*9.970E-01*	9.550E-01	9.950E-01	9.820E-01	9.950E-01	8.900E-01	9.870E-01	**1.000E+00**

TABLE III. COMPARISON OF AVERAGE NUMBERS OF SELECTED FEATURES BY ALL MSAS FOR FEATURE SELECTION

No	BeSD	CHIO	CO-HDEPSO	DSSA	FDA	GNDO	ODSFMFO	TTAO
DS1	26.30	20.33	10.90	*9.33*	15.03	14.23	28.93	**5.23**
DS2	325.23	*289.13*	305.93	441.23	309.70	302.17	446.93	**243.30**
DS3	16.80	13.80	*7.83*	12.47	10.27	10.63	16.87	**4.33**
DS4	135.80	112.17	119.13	*83.70*	131.47	126.97	155.87	**49.93**
DS5	3.83	*2.60*	**1.00**	7.10	**1.00**	**1.00**	3.63	**1.00**
DS6	6.67	6.13	4.57	*2.03*	5.33	6.53	2.30	**2.00**
DS7	4.67	4.07	5.03	**2.27**	4.97	4.80	6.20	*3.87*
DS8	6.23	5.60	4.03	5.37	3.67	**2.97**	9.00	*3.00*
DS9	3.70	3.57	3.30	*3.17*	**3.00**	4.00	5.00	**3.00**
DS10	8.53	6.27	*4.97*	12.13	5.00	6.80	9.97	**3.07**

Table III details the average number of features selected by all MSAs, implemented as wrapper-based feature selection techniques across 30 simulation runs for each dataset. The MSAs achieving the smallest and second-smallest feature subset sizes for each dataset are highlighted in bold and underlined text, respectively. In addition to their poor performance in terms of average classification accuracy, the results also reveal that ODSFMFO and BeSD are notably ineffective in minimizing the number of selected features, often yielding the largest and second-largest feature subsets across most datasets. Specifically, ODSFMFO consistently produced the largest feature subsets in seven datasets (DS1 to DS4, DS7 to DS9) and the second largest in one dataset (DS10). Meanwhile, BeSD was frequently associated with the second-largest feature subsets in six datasets (DS1, DS3 to DS6, and DS8). In contrast, CO-HDEPSO and FDA exhibited moderate performance, maintaining an average number of selected features that was neither particularly high nor low across most datasets.

Moreover, certain MSAs demonstrate inconsistent performance across both evaluation metrics, highlighting their inability to effectively balance the trade-off between model accuracy and complexity. For instance, while CHIO and GNDO achieve competitive average classification accuracies across most datasets, they fall short in consistently identifying smaller feature subsets that could reduce machine learning model complexity. Conversely, DSSA successfully identified the smallest feature subset size for one dataset (DS7) and the second smallest for four others (DS1, DS4, DS6, and DS9). However, DSSA ranks among the poorest in terms of average classification accuracy, as detailed in Table II. Additionally, DSSA was found to select the largest feature subsets for two datasets (DS5 and DS10) and the largest subset for another (DS2), indicating its potential inconsistency in handling datasets with diverse characteristics.

Contrary to the other MSAs evaluated, TTAO has exhibited superior performance by consistently identifying the smallest average number of selected features in eight datasets (DS1 to DS6, DS9, and DS10) and the second smallest in two others (DS7 and DS8). This emphasizes the effectiveness of TTAO's inherent search mechanism in optimally selecting feature subsets across datasets with diverse characteristics, thereby reducing the complexity of the machine learning models. The results presented in Tables II and III affirm TTAO's excellence in harmonizing accuracy with model simplicity, effectively addressing the challenges associated with feature selection.

D. Discussion

A key strength of TTAO lies in its ability to achieve an optimal trade-off between classification accuracy and feature subset size, primarily due to the effective balance it strikes between exploration and exploitation. The presence of both generic and local aggregation mechanisms enables TTAO to explore the search space while refining promising solutions, thus ensuring a well-balanced search process. In feature selection, high classification accuracy alone is insufficient if the feature subset is excessively large, as it can lead to overly complex models that are difficult to interpret and computationally expensive. TTAO addresses this issue by selecting smaller feature subsets while maintaining high accuracy, making it valuable in applications where simplicity and efficiency are critical. This balance is crucial for developing robust machine learning models that generalize well to unseen data. TTAO's consistent ability to reduce feature subset size across diverse datasets without sacrificing accuracy demonstrates the efficacy of its search strategies, allowing it to perform effectively even in high-dimensional spaces or datasets with multiple output classes, where many other algorithms tend to struggle.

Another practical advantages of TTAO over other MSAs is its reduced reliance on extensive parameter tuning. Many MSAs require fine-tuning of algorithm-specific parameters to balance exploration (global search) and exploitation (local search) effectively. In contrast, TTAO's performance depends primarily on the population size N and a few stochastically generated random variables (i.e, r_0, r_1, $r_{1'}$ r_1 and r_4), all of which require minimal adjustment. This reduction in parameter dependency simplifies the application of TTAO to different problem domains. By ensuring that the algorithm performs well without requiring extensive experimentation to find the optimal parameter settings, TTAO is highly adaptable and user-friendly. This makes it particularly attractive for real-world applications where tuning complex algorithmic parameters may not be feasible due to time constraints or lack of domain expertise.

The consistent performance of TTAO across a wide range of datasets indicates its versatility in tackling real-world feature selection problems. Its ability to handle datasets with high-dimensional features and varying class distributions highlights its robustness and generalizability. Furthermore, the efficiency of TTAO to reduce feature subsets without compromising accuracy can have significant practical implications. For example, in industries where computational resources are limited or where model interpretability is crucial (e.g., such as healthcare, finance, or sensor-based monitoring systems), TTAO's approach can lead to more efficient models with fewer features, ultimately reducing training time, memory requirements, and the risk of overfitting.

V. Conclusion

This paper introduces a novel wrapper-based feature selection technique utilizing the Triangulation Topology Aggregation Optimizer (TTAO), which is inspired by the geometric properties of triangular topology and principles of triangular similarity. Unlike its prior applications to real-valued decision variable problems, this study explores TTAO's adaptability to challenging real-world optimization problems with binary solution spaces. To facilitate this adaptation, a conversion mechanism is employed to transform continuous decision variables into binary ones, thus enabling the inherently real-valued TTAO for use in binary domains. TTAO generates diverse triangular topological units of consistent shape but varying sizes, serving as dynamic evolutionary entities throughout the optimization process. It incorporates two primary search strategies, generic and local aggregation, designed to balance exploration and exploitation effectively. Extensive simulations compare TTAO's performance in feature selection against seven other metaheuristic search algorithms (MSAs). The results indicate varied performances among the MSAs, with some underperform across both matrices (i.e., classification accuracy and feature subset size), while others fail to achieve a satisfactory balance. In contrast, the TTAO-based wrapper method excels, demonstrating an outstanding ability to achieve superior classification accuracy while minimizing feature subset size, thereby solving datasets with varied characteristics effectively.

While TTAO has shown excellent performance across the datasets used in this study, its scalability to extremely large datasets (in terms of both the number of features and instances) remains untested. Future research is needed to assess its computational efficiency and performance under more demanding, large-scale conditions. Additionally, the current study has focused on datasets with relatively balanced class distributions. TTAO's performance in highly imbalanced datasets, where classification bias might occur, has not been thoroughly explored and may require algorithmic adjustments to address such challenges. One promising direction for future research is to explore hybridization between TTAO and other MSAs to further improve performance. Combining the strengths of different algorithms could enhance the ability to balance exploration and exploitation, especially for more complex and dynamic datasets. Another potential extension of this work is applying TTAO in a multi-objective optimization framework, allowing the simultaneous optimization of multiple criteria (e.g., accuracy, computational cost, interpretability) to provide more comprehensive solutions for real-world feature selection tasks.

Acknowledgment

This study was funded by the Malaysian Ministry of Higher Education through the Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/UCSI/02/1) and UCSI University Research Excellence & Innovation Grant (REIG-FETBE-2022/038).

References

[1] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," Neural Computing and Applications, 2014, vol. 24, no. 1, pp. 175-186.

[2] T. Berghout and M. Benbouzid, "EL-NAHL: Exploring labels autoencoding in augmented hidden layers of feedforward neural networks for cybersecurity in smart grids," Reliability Engineering & System Safety, 2022, vol. 226, p. 108680.

[3] S. Mahmud Iwan *et al.*, "Spectroscopy data calibration using stacked ensemble machine learning," IIUM Engineering Journal, 2024, vol. 25, no. 1, pp. 208 - 224.

[4] W. Li, M. I. Solihin, and H. A. Nugroho, "RCA: YOLOv8-based surface defects detection on the inner wall of cylindrical high-precision parts," Arabian Journal for Science and Engineering, 2024.

[5] B. Jdid, W. H. Lim, I. Dayoub, K. Hassan, and M. R. B. M. Juhari, "Robust automatic modulation recognition through joint contribution of hand-crafted and contextual features," IEEE Access, 2021, vol. 9, pp. 104530-104546.

[6] Z. H. Ang, C. K. Ang, W. H. Lim, L. J. Yu, and M. I. Solihin, "Development of an artificial intelligent approach in adapting the characteristic of polynomial trajectory planning for robot manipulator," International Journal of Mechanical Engineering and Robotics Research, 2020.

[7] M. I. Solihin;, Y. Shameem;, T. Htut;, C. K. Ang;, and M. b. Hidayab', "Non-invasive blood glucose estimation using handheld near infra-red device," International Journal of Recent Technology and Engineering, 2019, vol. 8, no. 3, pp. 16-19.

[8] T. B. Hong et al., "Intelligent kitchen waste composting system via deep learning and Internet-of-Things (IoT)," Waste and Biomass Valorization, 2024, vol. 15, no. 5, pp. 3133-3146.

[9] O. O. Akinola, A. E. Ezugwu, J. O. Agushaka, R. A. Zitar, and L. Abualigah, "Multiclass feature selection with metaheuristic optimization algorithms: a review," Neural Computing and Applications, 2022, vol. 34, no. 22, pp. 19751-19790.

[10] Y. Zhang, X.-f. Song, and D.-w. Gong, "A return-cost-based binary firefly algorithm for feature selection," Information Sciences, 2017, vol. 418-419, pp. 561-574.

[11] A. Zakeri and A. Hokmabadi, "Efficient feature selection method using real-valued grasshopper optimization algorithm," Expert Systems with Applications, 2019, vol. 119, pp. 61-72.

[12] A. Machmudah et al., "Design optimization of a gas turbine engine for marine applications: off-design performance and control system considerations," Entropy, 2022, vol. 24, no. 12, p. 1729.

[13] L. Yao and W. H. Lim, "Optimal purchase strategy for demand bidding," IEEE Transactions on Power Systems, 2018, vol. 33, no. 3, pp. 2754-2762.

[14] Z. Danin, A. Sharma, M. Averbukh, and A. Meher, "Improved moth flame optimization approach for parameter estimation of induction motor," Energies, 2022, vol. 15, no. 23, p. 8834.

[15] L. Yao, W. H. Lim, S. S. Tiang, T. H. Tan, C. H. Wong, and J. Y. Pang, "Demand bidding optimization for an aggregator with a genetic algorithm," Energies, 2018, vol. 11, no. 10, p. 2498.

[16] A. Machmudah;, E. A. Bakar;, R. Rajendran;, W. H. Nugroho;, M. I. Solihin;, and A. Ghofur, "Control system optimisation of biodiesel-based gas turbine for ship propulsion," IAES International Journal of Artificial Intelligence, 2024, vol. 13, no. 2, pp. 1990-2000.

[17] M. F. Ahmad, N. A. M. Isa, W. H. Lim, and K. M. Ang, "Differential evolution: A recent review based on state-of-the-art works," Alexandria Engineering Journal, 2022, vol. 61, no. 5, pp. 3831-3872.

[18] S. Zhao, T. Zhang, L. Cai, and R. Yang, "Triangulation topology aggregation optimizer: A novel mathematics-based meta-heuristic algorithm for continuous optimization and engineering applications," Expert Systems with Applications, 2024, vol. 238, p. 121744.

[19] A. Almalaq, K. Alqunun, R. Abbassi, Z. M. Ali, M. M. Refaat, and S. H. E. Abdel Aleem, "Integrated transmission expansion planning incorporating fault current limiting devices and thyristor-controlled series compensation using meta-heuristic optimization techniques," Scientific Reports, 2024, vol. 14, no. 1, p. 13046.

[20] M. A. Elaziz, F. A. Essa, H. A. Khalil, M. S. El-Sebaey, M. Khedr, and A. Elsheikh, "Productivity prediction of a spherical distiller using a machine learning model and triangulation topology aggregation optimizer," Desalination, 2024, vol. 585, p. 117744.

[21] M. A. Zeidan, M. R. Hammad, A. I. Megahed, K. M. AboRas, A. Alkuhayli, and N. Gowtham, "Enhancement of a hybrid electric shipboard microgrid's frequency stability with triangulation topology aggregation optimizer-based 3DOF-PID-TI controller," IEEE Access, 2024, vol. 12, pp. 66625-66645.

[22] W.-L. Cheng et al., "Flow direction algorithm for feature selection," Singapore, 2023: Springer Nature Singapore, in Advances in Intelligent Manufacturing and Mechatronics, pp. 187-198.

[23] W.-L. Cheng et al., "Feature selection of medical dataset using african vltures optimization algorithm," Singapore, 2023: Springer Nature Singapore, in Advances in Intelligent Manufacturing and Mechatronics, pp. 175-185.

[24] W.-L. Cheng et al., "Wrapper-based feature selection using sperm swarm optimization: a comparative study," Singapore, 2024: Springer Nature Singapore, in Advances in Intelligent Manufacturing and Robotics, pp. 343-353.

[25] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Al-Zoubi, and S. Mirjalili, "Binary grasshopper optimisation algorithm approaches for feature selection problems," Expert Systems with Applications, 2019, vol. 117, pp. 267-286.

[26] D. Rodrigues, V. H. C. de Albuquerque, and J. P. Papa, "A multi-objective artificial butterfly optimization approach for feature selection," Applied Soft Computing, 2020, vol. 94, p. 106442.

[27] R. R. Mostafa, A. A. Ewees, R. M. Ghoniem, L. Abualigah, and F. A. Hashim, "Boosting chameleon swarm algorithm with consumption AEO operator for global optimization and feature selection," Knowledge-Based Systems, 2022, vol. 246, p. 108743.

[28] B. J. Ma, S. Liu, and A. A. Heidari, "Multi-strategy ensemble binary hunger games search for feature selection," Knowledge-Based Systems, 2022, vol. 248, p. 108787.

[29] R. Wu et al., "An improved sparrow search algorithm based on quantum computations and multi-strategy enhancement," Expert Systems with Applications, 2023, vol. 215, p. 119421.

[30] C. Zhong, G. Li, Z. Meng, H. Li, and W. He, "A self-adaptive quantum equilibrium optimizer with artificial bee colony for feature selection," Computers in Biology and Medicine, 2023, vol. 153, p. 106520.

[31] D.-S. Khafaga, E.-S.-M. El-kenawy, F. Alrowais, S. Kumar, A. Ibrahim, and A.-A. Abdelhamid, "Novel optimized feature selection using metaheuristics applied to physical benchmark datasets," Computers, Materials \& Continua, 2023 vol. 74, no. 2, pp. 4027--4041.

[32] D.-S. Khafaga et al., "Hybrid dipper throated and grey wolf optimization for feature selection applied to life benchmark datasets," Computers, Materials \& Continua, 2023, vol. 74, no. 2, pp. 4531--4545.

[33] A. A. Abdelhamid et al., "Innovative feature selection method based on hybrid sine cosine and dipper throated optimization algorithms," IEEE Access, 2023, vol. 11, pp. 79750-79776.

[34] P. Civicioglu and E. Besdok, "Bezier Search Differential Evolution Algorithm for numerical function optimization: A comparative study with CRMLSP, MVO, WA, SHADE and LSHADE," Expert Systems with Applications, 2021, vol. 165, p. 113875.

[35] M. A. Al-Betar, Z. A. A. Alyasseri, M. A. Awadallah, and I. Abu Doush, "Coronavirus herd immunity optimizer (CHIO)," Neural Computing and Applications, 2021, vol. 33, no. 10, pp. 5011-5042.

[36] Z. C. Choi et al., "Hybridized metaheuristic search algorithm with modified initialization scheme for global optimization," Cham, 2021: Springer International Publishing, in Advances in Robotics, Automation and Data Analytics, pp. 172-182.

[37] B. Jena, M. K. Naik, A. Wunnava, and R. Panda, "A differential squirrel search algorithm," Singapore, 2021: Springer Singapore, in Advances in Intelligent Computing and Communication, pp. 143-152.

[38] H. Karami, M. V. Anaraki, S. Farzin, and S. Mirjalili, "Flow Direction Algorithm (FDA): a novel optimization approach for solving optimization problems," Computers & Industrial Engineering, 2021, vol. 156, p. 107224.

[39] Y. Zhang, Z. Jin, and S. Mirjalili, "Generalized normal distribution optimization and its applications in parameter extraction of photovoltaic models," Energy Conversion and Management, 2020, vol. 224, p. 113301.

[40] Z. Li, J. Zeng, Y. Chen, G. Ma, and G. Liu, "Death mechanism-based moth–flame optimization with improved flame generation mechanism for global optimization tasks," Expert Systems with Applications, 2021, vol. 183, p. 115436.

Mixed Integer Programming Model Based on Data Algorithms in Sustainable Supply Chain Management

Shaobin Dong[1]*, Aihua Li[2]
Faculty of Business, Huaiyin Institute of Technology, Huai'an, 223001, China[1]
Faculty of Electronic and Information Engineering, Huaiyin Institute of Technology, Huai'an, 223001, China[2]

Abstract*—With the deepening of globalization and increasing demands for environmental sustainability, modern supply chains are faced with increasingly complex management challenges. To reduce management costs and enhance efficiency, an experimental approach is proposed based on a Mixed Integer Programming Model, integrating heuristic algorithms with adaptive genetic algorithms. The objective is to improve both the efficiency and sustainability of supply chain management. Initially, the selection of suppliers within the supply chain is analyzed. Subsequently, heuristic algorithms and genetic algorithms are jointly employed to design, generate, and optimize initial solutions. Results indicate that during initial runs on training and validation sets, the fitness values of the research method reached as high as 99.67 and 96.77 at the 22nd and 68th iterations, respectively. Moreover, on the training set with a dataset size of 112, the accuracy of the research method was 98.56%, significantly outperforming other algorithms. With the system running five times, the time consumed for supplier selection and successful order allocation was merely 0.654s and 0.643s, respectively. In practical application analysis, when the system iterated 99 times, the research method incurred the minimum total cost of 962,700 yuan. These findings demonstrate that the research method effectively minimizes supply chain management costs while maximizing efficiency, offering practical strategies for optimizing and sustainably developing supply chain management.

***Keywords*—Mixed Integer Programming Model; sustainable; supply chain management; heuristic algorithm; adaptive genetic algorithm**

I. Introduction

As environmental issues and social responsibilities become increasingly prominent, enterprises must consider not only economic benefits but also the long-term impacts of their operations on the environment and society. Sustainable supply chain management has become an integral part of corporate strategies, demanding that companies maintain supply chain efficiency and cost advantages while also considering environmental protection and social responsibility [1-2]. The complexity of supply chain management arises primarily from multi-stage coordination and the handling of numerous decision variables [3]. Against this backdrop, traditional supply chain management approaches often prove inadequate when faced with growing market demand fluctuations and environmental sustainability requirements, necessitating reevaluation and optimization [4]. Therefore, leveraging modern technologies and algorithms to enhance the flexibility and adaptability of supply chains while achieving environmental and social sustainability goals has become a pressing issue. Data algorithms, particularly Mixed Integer Programming Models, offer new solutions for supply chain management due to their efficiency and precision in addressing complex decision-making problems [5]. These models effectively integrate and coordinate various elements within the supply chain, such as supplier selection, product manufacturing, and finished product consumption, while considering factors like time, resources, and environmental impact. However, relying solely on Mixed Integer Programming Models is insufficient to address all complexities of supply chain issues, especially in large-scale problem-solving and real-time decision optimization. To better address these challenges, the experiment incorporates heuristic algorithms and adaptive genetic algorithms (GA) to jointly enhance the Mixed Integer Programming Model, aiming for a more comprehensive and profound resolution of optimization issues in supply chain management. The aspiration is to provide new perspectives and methods for sustainable supply chain management, assisting enterprises in pursuing efficiency while also achieving environmental protection and social responsibility objectives.

The article is mainly divided into five sections. Section II is a literature review, which mainly analyzes and summarizes the relevant research at home and abroad. Section III is the research method, which analyzes the selection problem of the supply chain in supply chain management, and then uses an interactive product order and supplier selection sorting method to generate the required initial solution. Finally, the adaptive genetic algorithm is used to optimize the supplier selection solution. Section IV is the research results, mainly analyzing the performance and application effects of the improved mixed integer programming method constructed. Section V is the conclusion, which mainly summarizes the content of the entire article and proposes current shortcomings and future research directions.

II. Related Work

In recent years, sustainable supply chain management has become a focal point of attention for both businesses and scholars. Sustainable supply chains not only emphasize traditional costs and efficiency but also involve balancing environmental protection, social responsibility, and economic benefits. Numerous scholars have undertaken summarizations of the design and management of supply chain networks. Researchers, such as Dwivedi et al., addressed the significant complexities between grain supply chain management at different levels and carbon emissions. They combined GAs with quantum GAs and metaheuristic algorithms to analyze the allocation of vehicles and the selection of order sets. The results indicated significantly reduced computation time and enhanced operational efficiency [6]. Sadeghi and the team proposed a novel closed-loop supply chain network approach based on a mixed-integer linear programming model to reduce transportation and management costs in the supply chain. The model covered both fleet transportation routes and locations and was validated through the production of Iranian automotive components, gaining full recognition from management personnel [7]. Manupati et al. introduced a monitoring system based on blockchain technology to comprehensively monitor changes in supply chain performance. Considering carbon emissions constraints, they treated product production, sales, and inventory levels as control factors. Compared to non-dominated sorting GAs, their method demonstrated better feasibility [8]. Isaloo and Paydar presented a supply chain network design based on a dual-objective mathematical programming model to enhance the flow of the entire sustainable supply chain. Through sensitivity analysis of weights, they found the optimal objective solution, validating the superiority of their model [9]. Mogale et al. proposed a data-driven supply chain network approach based on a mixed-integer linear programming model to improve the convenience of grain procurement and transportation in developing countries. The results showed a significant enhancement in grain transportation efficiency and cost reduction [10].

Furthermore, many scholars have conducted analyses on the design and application of Mixed Integer Programming Models. Ahmadini and academic professionals aimed to reduce the environmental impact of pollutants emitted during the preservation and transportation of products. They proposed a green supply chain network based on a multi-project multi-objective inventory model. Simulation experiments verified the practical effectiveness of the constructed method, promoting the rapid development of the manufacturing industry [11]. Researchers, including Zhang Y, addressed the prolonged time consumption in logistics distribution within the supply chain network by proposing a method based on mixed-integer nonlinear programming. The results indicated that the constructed model effectively resolves nonlinear layout optimization issues in logistics transportation [12]. Beiki H and others, in order to tackle the sustainability of supplier selection and order allocation issues, introduced an integrated approach based on language entropy weight and multi-objective programming. The method demonstrated its effectiveness in improving the relationship between supply chain practitioners and suppliers, leading to the maximization of product profits [13]. Li C and colleagues aimed to increase the adoption rate of renewable energy generator units while reducing the complexity of operational decisions. They proposed an operational method based on the MILP-GTEP model. The study extensively explored the relationship between space and time, and the superiority of the experimentally constructed method was validated through a case study from the Texas Electricity Reliability Council, significantly reducing the difficulty of unit power generation [14]. In addressing the issue of maximizing profit extraction in open-pit mining production scheduling, Rivera Letelier O and fellow researchers presented a direct block scheduling and stage scheduling method based on mixed-integer programming modeling. Data revealed that, compared to traditional methods, the average gap during boundary cutting in the experimentally constructed method decreased from 1.52% to 0.71%, demonstrating its significant superiority [15].

In conclusion, with the advancement of globalization and technological innovation, supply chain management is facing unprecedented challenges and opportunities. Despite the availability of numerous technologies and tools to address supply chain network management issues, there are still challenges, such as balancing the computational complexity and solving efficiency of algorithms. In order to enhance the efficiency of enterprise supply chain management and reduce unnecessary time consumption, an experiment proposes a Mixed Integer Programming Model based on heuristic algorithms and adaptive GAs. This model is applied to the supply chain management of sustainable enterprises, with the expectation of promoting the development of sustainable supply chain management.

III. Heuristic Algorithm and GA Algorithm Mixed Integer Programming Model for Sustainable Supply Chain Management

Considering the complexities and variability in supply chain management, especially when facing sustainability challenges, traditional optimization methods often struggle to adapt. Therefore, the experiment introduces a Mixed Integer Programming Model based on a data-driven algorithm to enhance it. This model can not only handle problems involving a large number of decision variables and complex constraints but also flexibly adapt to the dynamic changes and diverse requirements in supply chain management.

A. *Supplier Selection in Supply Chain Management*

In order to provide a clearer description of supply chain management issues, the experiment focuses on supplier selection and product cost aspects, analyzing the management of the supply chain. The study, based on the characteristics of a three-tier supply chain with manufacturing enterprises at its core, makes assumptions about supplier selection in the three-tier supply chain management: first, within the production period of the same product order, the order will not exceed the maximum capacity of total production; second, all products in the order overlap; third, there are restrictions on the maximum capacity of the corresponding manufacturing enterprises, and so on. The operations of multi-supplier supply chain management are illustrated in Fig. 1.

The ultimate goal is to minimize the number of times a supplier is selected. The variability in supplier selection in supply chain management is defined by Eq. (1).

$$\begin{gathered} e_{jkq} \geq y_{jkq} - y_{jkq-1}, \\ \forall i \in \{1, \cdots, N\}, q \in \{1, \cdots, Q\}, j \in \{1, \cdots, M\}, k \in \{1, \cdots, K\} \end{gathered} \tag{1}$$

In Eq. (1), y_{jkq} and e_{jkq} represent binary variables. The calculation of supplier selection is then obtained, as shown in Eq. (2).

$$\begin{cases} T_j = \sum_{k=1}^{K} \sum_{q=1}^{Q} e_{jkq} \geq 0, \forall i \in \{1, \cdots, N\}, q \in \{1, \cdots, Q\}, j \in \{1, \cdots, M\}, k \in \{1, \cdots, K\} \\ T_{total} = \sum_{j=1}^{M} T_j \geq 0, \forall j \in \{1, \cdots, M\} \end{cases} \tag{2}$$

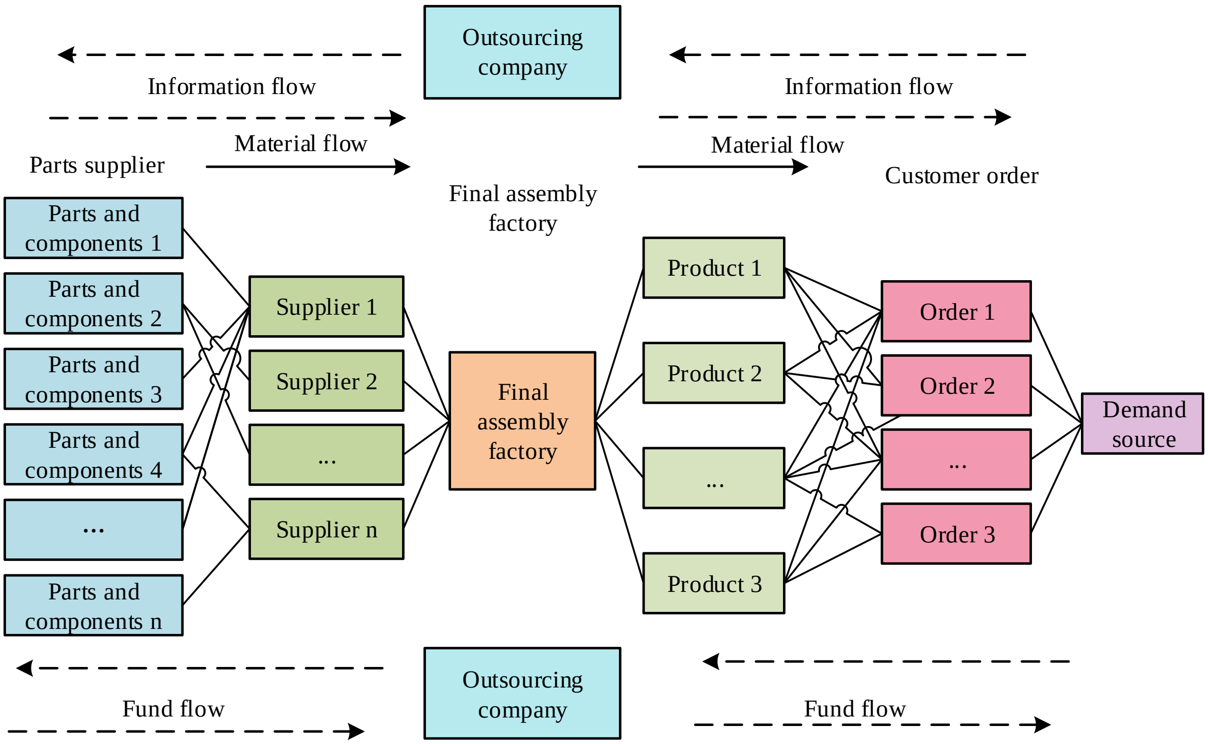


Fig. 1. Multi-supplier supply chain management operations.

In Eq. (2), T_j represents the number of times supplier j is selected, and T_{total} represents the total number of times all suppliers are selected. T_j must be above zero. In the process of supply chain management selection, these are positive integers. Considering the entire production stage, the optimization of the supplier selection frequency in the supply chain management process is achieved, and the objective function $\min_{total}$ of this mathematical model is obtained. However, selecting product suppliers to meet production orders is a highly complex problem that requires consideration of constraints on supplier selection [16]. To some extent, reducing the number of supplier selections must ensure that different products in the total production orders are supplied by at least one supplier, and the corresponding constraint is expressed in Eq. (3).

$$\begin{cases} \sum_{k=1}^{K}\sum_{j=1}^{M}\sum_{i=1}^{N} y_{ijk} \ge 1, \forall i \in \{1,\cdots,N\}, j \in \{1,\cdots,M\}, k \in \{1,\cdots,K\} \\ \sum_{j=1}^{M}\sum_{i=1}^{N} y_{ijk} \le 1, \forall i \in \{1,\cdots,N\}, j \in \{1,\cdots,M\}, k \in \{1,\cdots,K\} \\ \sum_{j=1}^{M} B_{is} \ge \sum_{i=1}^{N} A_{is}, \forall i \in \{1,\cdots,N\}, j \in \{1,\cdots,M\}, s \in \{1,\cdots,S\} \end{cases} \tag{3}$$

In Eq. (3), y_{ijk} and B_{is} are both binary variables. Their values are 1 when the supplier can complete the corresponding task. These constraints reflect the main attributes of supply chain management, optimizing the experimental process.

B. *Initial Solution Generation Method for Supplier Selection in Supply Chain Management Based on Heuristic Algorithms*

The experimental method employs an interactive approach to generate the required initial solutions by ordering product orders and selecting supplier rankings. Initially, orders are selected and sorted according to certain patterns in the interactive process of product orders and supplier ranking [17]. To assess the fulfillment rate of suppliers for products in the k th batch, assume the relationship between supplier j and products corresponds to B_{js}, and the relationship between order i and products corresponds to A_{is}. If the products in the orders cannot be produced by the corresponding suppliers, the decision variable $B_{js}=0$. Otherwise, if production by the supplier is possible, the corresponding pointer variable $B_{js}=1$. Simultaneously, if the number of component types in products from different orders is S_i, including product s, then the corresponding pointer variable $A_{is}=1$, otherwise, $A_{is}=0$. Based on the parameters and principles mentioned above, for a product supplier M, the calculation of the fulfillment rate for the corresponding supplier's product orders is given by Eq. (4).

$$f_{ijk} = \frac{\sum_{s=1}^{S} B_{js}A_{is}}{\sum_{s=1}^{S} A_{is}}, \\ \forall i \in \{1,\cdots,N\}, j \in \{1,\cdots,M\}, s \in \{1,\cdots,S\}\ldots \tag{4}$$

In Eq. (4), the multiplication of B_{js} and A_{is} represents the common product relationship between suppliers and orders. It can be observed that the maximum value of the fulfillment rate of products in the orders does not exceed 1. When $f_j = 1$, indicating a value of 1, all the required products for customers in the supply chain product orders can be provided by that supplier. The enterprise can freely choose suppliers in the process of management and sales. The experiment sets the supply rate of supplier j as f_{ijk} to represent the extent to which the product supplier can meet the demands of a set of customer orders. The calculation is cyclically performed on the order pool within the same batch [18-19]. By cumulatively adding the fulfillment rates of products related to suppliers in all orders, the fulfillment rate f_{jk} of the supplier providing the required products in the k th batch is obtained, as shown in Eq. (5).

$$f_{jk} = \sum_{i=1}^{N} f_{ijk} \tag{5}$$

In Eq. (5), f_j has a value range of [0, n]. After selecting the supplier with the highest product fulfilment rate, it is necessary to decide how to allocate the required quantity of these products to the corresponding suppliers. This requires an appropriate production sequence to reduce the number of supplier transports (i.e., product order selection strategy). To avoid orders being repeatedly assigned while ensuring a complete order is in a corresponding set of orders awaiting production, the supplier's set of orders awaiting production is designated as OA, with specific calculations outlined in Eq. (6).

$$OA = \left\{ O_i \middle| f_{ij} > 1, \forall j \in \{1, \cdots, M\}, i \in \{1, \cdots, N\} \right\} \tag{6}$$

In Eq. (6), OA represents the set of orders awaiting production. Additionally, due to various constraints and limitations on supplier production capacities, the production of ordered products by relevant suppliers may be significantly insufficient. To describe these constraints, the experiment introduces a new set of supplier production orders, assuming this order set is OW. The total demand value for orders in this set will not exceed the total production capacity of the factory, as detailed in Eq. (7).

$$OW = \left\{ Oi \middle| f_{ij} > 0, and \quad I_{is} \le L, \forall j \in \{1, \cdots, M\}, s \in \{1, \cdots, S\} \right\} \tag{7}$$

By comprehensively calculating the above equations, a specific method for arranging and managing supplier orders in the Kth batch is obtained. Furthermore, a feasible solution to the problem of selecting multiple suppliers in supply chain management is achieved by fixing the orders of suppliers one by one in priority order.

In Fig. 2, by analyzing the correspondence between suppliers and orders, the priority mechanism for selecting suppliers and orders is clarified. A heuristic algorithm for solving the multi supplier selection problem was developed based on this mechanism. The main process of supplier and order selection for each round of the algorithm is shown in the figure above. This heuristic algorithm can efficiently obtain an initial solution to the studied problem.

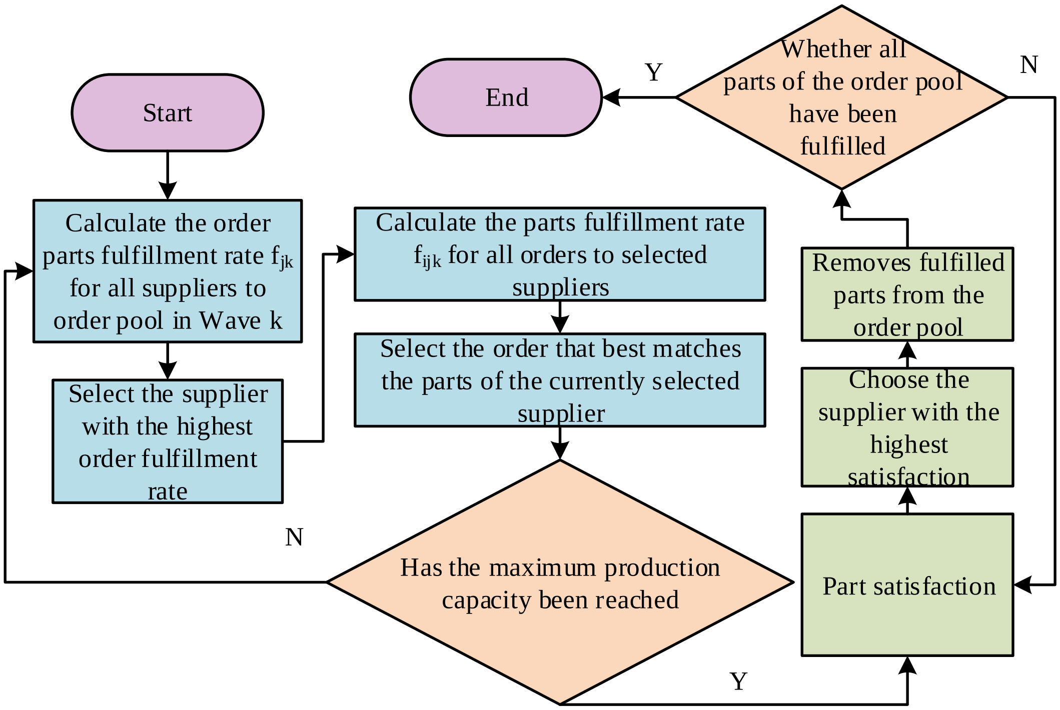

Start
Calculate the order parts fulfillment rate f_{jk} for all suppliers to order pool in Wave k
Select the supplier with the highest order fulfillment rate
Calculate the parts fulfillment rate f_{ijk} for all orders to selected suppliers
Select the order that best matches the parts of the currently selected supplier
Has the maximum production capacity been reached
N
Y
Part satisfaction
Choose the supplier with the highest satisfaction
Removes fulfilled parts from the order pool
Whether all parts of the order pool have been fulfilled
Y
N
End

Fig. 2. Supplier and order selection method in Wave k.

C. Optimization of Feasible Solutions for Supply Chain Supplier Selection Management Based on Adaptive GA

An initial feasible solution is obtained by decoupling the relationship between supplier management and order products. Subsequently, a local search is applied to the feasible solution to obtain an initial population. Iterative updates are performed using a Genetic Algorithm (GA) to generate various order production sequences. In each GA iteration, suppliers for each order group are sorted based on specific constraints, and this sorting is used as the fitness indicator for individuals in the population. After multiple iterations, a cost-optimized selection solution is finally obtained. The solution process of GA is illustrated in Fig. 3.

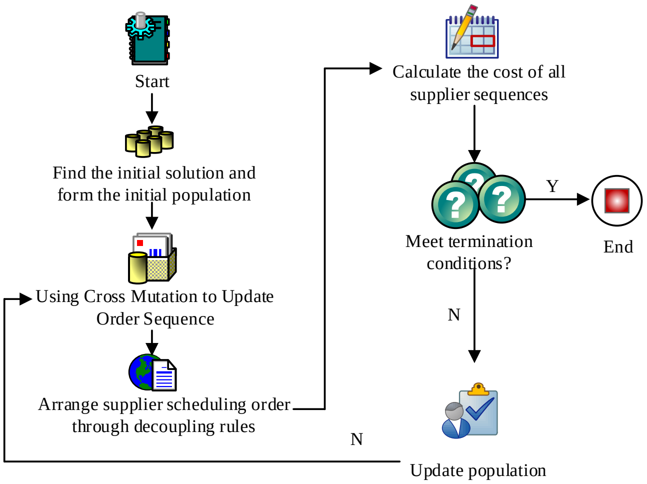

Start
Find the initial solution and form the initial population
Using Cross Mutation to Update Order Sequence
Arrange supplier scheduling order through decoupling rules
Calculate the cost of all supplier sequences
Meet termination conditions?
Y
End
N
Update population
N

Fig. 3. The whole process of initial solution optimization.

It is important to note that the management and selection problem involving multiple suppliers is, in fact, an NP-hard problem. Due to the high complexity of this problem, experiments result in numerous initial feasible solutions using heuristic algorithms. To preserve the characteristics of these initial feasible solutions to the maximum extent, the experiment introduces a simulated annealing algorithm to expand the search neighborhood of initial feasible solutions. The initial temperature is set as T_0, and the objective function values of the obtained initial feasible solutions are calculated. The generation process of new solutions is illustrated in Fig. 4.

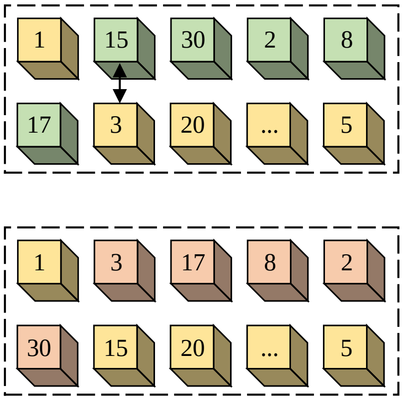


Fig. 4. Value generation method of new solution.

New solutions are evaluated using the Metropolis criterion in the simulated annealing algorithm. When the fitness of the new solution is higher than that of the original solution, the retention probability is 1. For lower fitness, acceptance is probabilistic. The mathematical formula for the Metropolis criterion is shown in Eq. (8).

$$p=\begin{cases}1 & if \quad E(x_{new})<E(x_{new})' \\ \exp\left(-\dfrac{E(x_{new})-E(x_{new})'}{T}\right) & if \quad E(x_{new})\geq E(x_{new})'\end{cases} \tag{8}$$

In Eq. (8), p represents the probability of accepting the new feasible solution. E represents the internal energy corresponding to each state. T represents the current temperature of the system. The experiment adopts an exponential cooling annealing strategy, and the decay function is shown in Eq. (9).

$$T_{n+1}=CT_n \tag{9}$$

In Eq. (9), T represents a constant less than 1. n represents the index of a state. Subsequently, the experiment employs a roulette wheel method to select outstanding individuals and uses a two-point crossover method for individual crossover operations. After randomly generating a crossover probability, two individuals are randomly selected from the population. If the random number is less than the crossover probability, a crossover operation is performed: determining two crossover points, exchanging and repairing the corresponding chromosome segments, and generating two new individuals. The process of crossover operation is illustrated in Fig. 5.

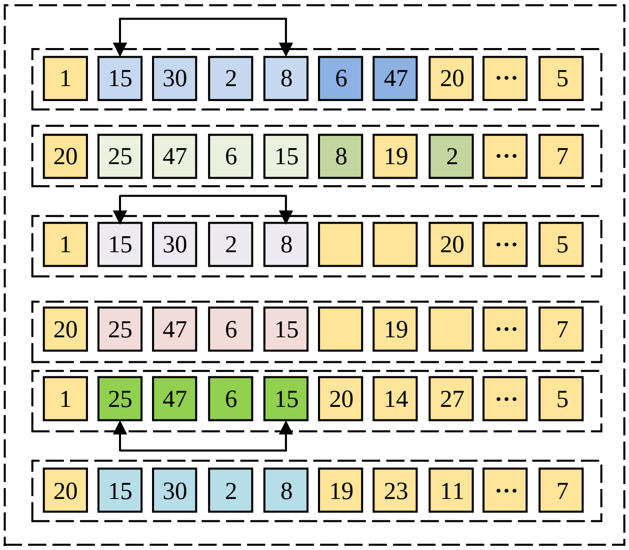


Fig. 5. The whole process of crossover operator calculation.

Similar to the crossover operation, the mutation method selected is two-point mutation. With a certain probability, two points are randomly generated, and natural numbers in the individual are exchanged to obtain a completely new order sequence. The calculation process of the mutation operator is illustrated in Fig. 6.

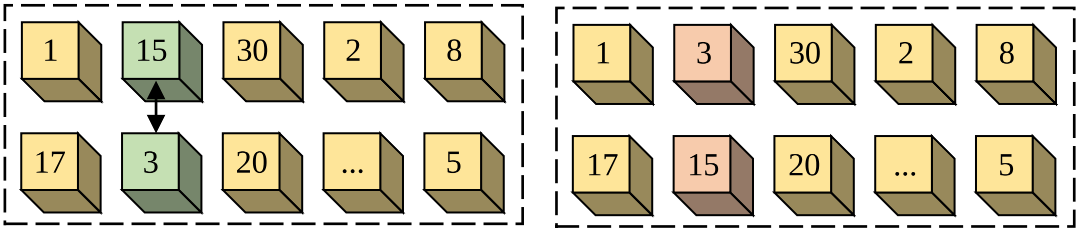


Fig. 6. The whole calculation process of mutation operator.

Based on the decoupling relationship, the order sequence is transformed into a supplier sequence, thereby determining the selection frequency of each supplier. The calculation method of the fitness function is shown in Eq. (10).

$$fitness\_F(x) = \frac{C}{T_{total}}, T_{total} > 0 \tag{10}$$

In Eq. (10), T_{total} represents the total cost of supplier selection. T_{total} represents a positive real number. After successfully obtaining the value of the fitness function, the experiment, to prevent the calculation process from falling into a local optimal solution, further establishes a Genetic Algorithm based on Invasive Weed Optimization (IWO) for adaptive large neighborhood search. In this algorithm, when the cumulative value of random numbers can disrupt the population, the larger the numerical value, the more severe the disruption. The relevant calculation is shown in Eq. (11).

$$\partial = \frac{\sum \lambda - C}{\sum 2\lambda} \tag{11}$$

In Eq. (11), λ represents the cumulative value of random numbers. C represents a constant. ∂ represents the proportion of individual disruption. It is observed that the range of ∂ is [0, 1/2). By controlling the numerical value of C, the degree of individual disruption is controlled. Based on the degree of disruption, the number of individuals to be destroyed in the population is calculated, and inappropriate individuals are randomly removed from the population. Additionally, based on the fitness of individuals, supplier management is ranked, and the optimal individual is selected, forming the final seed individual for the next round of optimization.

IV. Results

In order to validate the superior performance of the constructed method, three existing methods were chosen for comparison with the proposed approach in the context of sustainable pharmaceutical supply chain networks. These methods include the Improved Hybrid Multi-Objective Heuristic Algorithm (IHMOH) based on an improved mixed multi-objective heuristic algorithm, the Supply Chain Closed-Loop Management Method (MINLP) based on a mixed-integer nonlinear programming model, and the Two-Stage Integrated Method for Green Supply Chain Supplier Selection and Order Allocation (FAHP-MILP) based on fuzzy analytic hierarchy process and multi-objective mixed integer linear programming [20-22]. To ensure fairness and reasonableness in the experimental setup, all models shared identical simulation environment parameters, as detailed in Table I.

TABLE I. Settings of Related Parameters

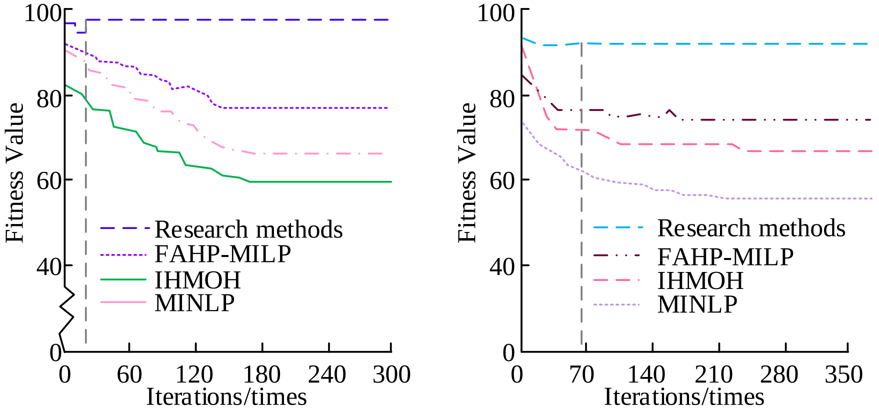
Project	Choice
Graphics card	NVIDIA RTX2080Ti
Graphics card memory	32G
Fresh products	50
Crossover rate	0.7
Mutation rate	0.1
Operating system	Ubuntu18.04
Network architecture	Pytorch v1.2.0
Optimizer	Adam
Programming software	Vscode2017 Anaconda3
CUDA	Cuda 11.0 with cudnn

The selected dataset for the experiment is from the supply chain management system of a well-known company in the United States, and a total of 10000 valid data points were obtained. Before applying the dataset to the model, the following preprocessing steps were performed on all data: data cleaning to remove incomplete or erroneous records; Standardization processing to eliminate the influence of different dimensions; Outlier detection and processing to ensure the reliability of the dataset. These steps are crucial for improving the predictive accuracy of the model. After data preprocessing to remove redundant data, 8000 valid data were obtained, and 80% of the total dataset was randomly selected as the validation set; another 20% is used as the training set. Firstly, the fitness values of the four algorithms were compared when performing tasks on the two datasets, as illustrated in Fig. 7.

Fig. 7(a) depicts the changes in fitness values of the four algorithms on the training set. As the number of system iterations increases, the fitness values of all four algorithms exhibit a rapid and fluctuating trend. At the commencement of the system operation, the fitness value of the research method undergoes a slight variation. Subsequently, at the 22nd iteration, the research method attains its maximum fitness value, maintaining a stable state thereafter with a numerical value as high as 99.67. In contrast, the fitness values of the FAHP-MILP, IHMOH, and MINLP algorithms start stabilizing only after a higher number of iterations, reaching stability at the 157th, 178th, and 176th iterations, respectively, with corresponding values of 81.23, 62.23, and 68.84. Fig. 7(b) illustrates the changes in fitness values of the four algorithms on the validation set. The research method achieves its optimal fitness value of 96.77 after 68 iterations, while the fitness values of the other three methods continue to decrease and remain consistently lower than that of the research method. These results indicate that the research method consistently maintains a higher fitness value, emphasizing its faster convergence speed and higher computational efficiency. Given the varying product categories supplied by different vendors and their differing accuracy in product selection during the management process, the experiment proceeds to compare the accuracy of vendor selection. Specific results are presented in Fig. 8.

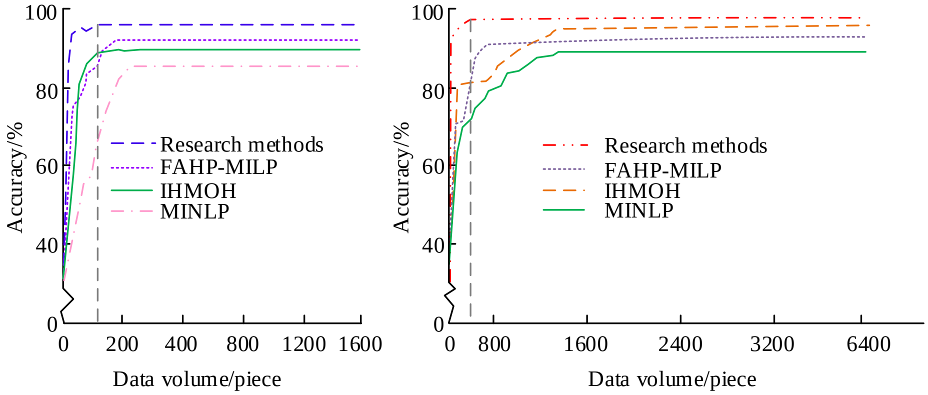
Fig. 8(a) displays the vendor accuracy obtained by different methods on the training set. It is observed that as the data volume increases, the accuracy of all four algorithms shows varying degrees of improvement. When the data volume is 112, the research method achieves the maximum accuracy at 98.56%, significantly higher than the accuracy of the other three methods. Fig. 8(b) shows the accuracy of different algorithms in selecting vendors on the validation set. When the accuracy of the experimentally constructed method reaches its maximum value at a data volume of 402, the corresponding accuracy is 98.33%. Additionally, at data volumes of 789, 1544, and 1502, the accuracy of vendor selection for the FAHP-MILP, IHMOH, and MINLP algorithms is 90.15%, 96.32%, and 89.36%, respectively. In summary, the research method exhibits the highest accuracy in vendor selection and can be applied in the development process of green enterprises. Subsequently, the four methods were applied to the training set for a comparative analysis of computation time. The entire experiment was conducted in five cycles, and the specific results for supplier selection time (Supplier option-T1) and successful order allocation time (Order allocation -T2) are presented in Fig. 9.

In Fig. 9 (a)–(d), different research methods and algorithms, namely the research method, FAHP-MILP algorithm, IHMOH algorithm, and MINLP algorithm, were employed. It can be observed that, in the five parallel experiments designed, the research method exhibited significantly less time consumption for adapting to suppliers compared to the other three algorithms. The minimum time for T1 was 0.654s, and for T2, it was 0.643s. This indicates that the research method, while running the supply chain system, has a more agile decision speed and lower time consumption in selecting suppliers. However, the system, despite having faster operational efficiency, also requires an analysis of the overall cost expenditure. To analyze the overall cost expenditure during the operation of the supply chain management system, the experiment proceeded by selecting the data source company as the product supplier and applying the four methods to two datasets. The analysis focused on the variation in supply chain management costs as the data volume increased, as shown in Fig. 10.



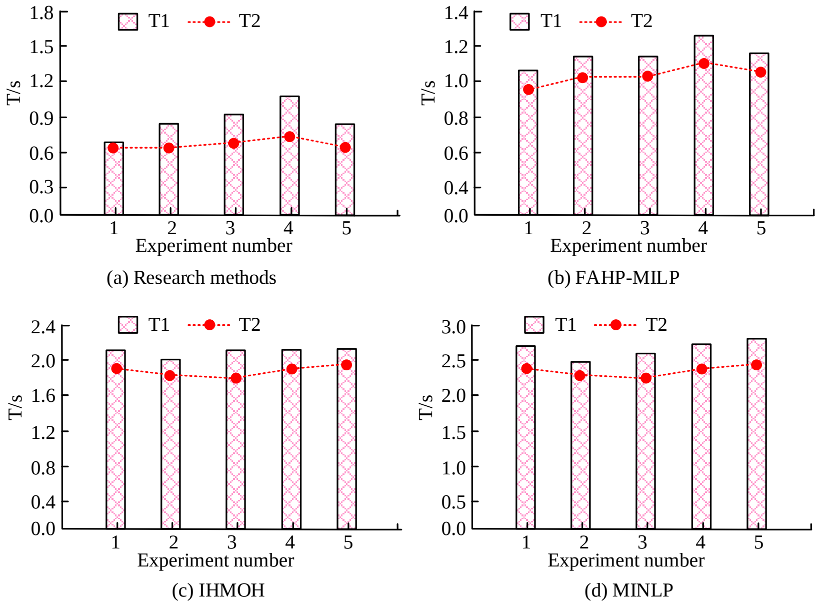
(a) Training set (b) Validation set

Fig. 7. Comparison of fitness values on the two data sets.



(a) Training set (b) Validation set

Fig. 8. Accuracy of supplier selection.



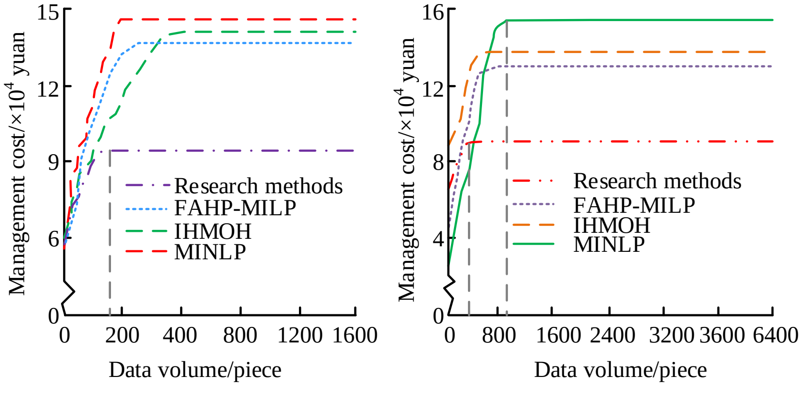
(a) Research methods

(b) FAHP-MILP

(c) IHMOH

(d) MINLP

Fig. 9. Comparison of the operation time of the four algorithms on the two data sets.



(a) Training set

(b) Validation set

Fig. 10. Changes in supply chain management costs under different algorithm operations.

Fig. 10 (a) depicts the variation in supply chain management costs on the training set. When the data volume reached 191, the research method exhibited the minimum management cost, valued at 97.6 thousand yuan. When the data volume reached 268, the FAHP-MILP algorithm's management cost reached its minimum, with a value as high as 135.8 thousand yuan. Fig. 10 (b) illustrates the variation in supply chain management costs when the four methods were applied to the validation set. As the data volume increased to 392, the research method had the minimum management cost at 92.3 thousand yuan. The supply chain management costs of the other three methods were significantly higher, especially when the data volume reached 908, where the MINLP algorithm incurred a very high cost of 158.9 thousand yuan. In summary, the research method demonstrated better cost-effectiveness in handling complex supply chain problems and effectively controlling supply chain management costs. Finally, the four algorithms were applied to the closed-loop supply chain network of a sustainable agricultural products enterprise. This enterprise, a large company in the United States, initiated the implementation of a green closed-loop supply chain network in response to national policies in August 2020. The study selected 50 examples of fresh products owned by the enterprise for closed-loop supply chain network optimization, and the total cost results are shown in Fig. 11.

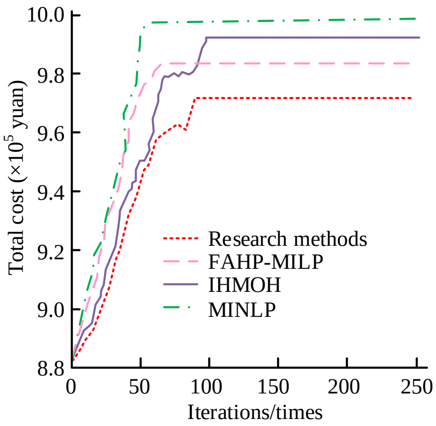


Fig. 11. Total cost under different algorithm operations.

From Fig. 11, it can be observed that the FAHP-MILP algorithm, IHMOH algorithm, and MINLP algorithm incurred higher total costs, with corresponding maximum costs of 983.8 thousand yuan, 975.7 thousand yuan, and 998.8 thousand yuan, respectively. Additionally, when the system reached the 99th iteration, the research method had the least total cost expenditure at 962.7 thousand yuan. Comparatively, although the research method's costs were initially comparable to the IHMOH algorithm, they gradually decreased with iterations and eventually became lower than other algorithms.

V. Discussion and Conclusion

A. Discussion

A mixed integer programming model based on data algorithms has been proposed, which combines heuristic algorithms and adaptive genetic algorithms to improve the efficiency and sustainability of sustainable supply chain management. Through comprehensive analysis and optimization of supplier selection issues in the supply chain, the proposed method has demonstrated significant advantages in multiple aspects. The hybrid model proposed by the research institute effectively integrates the fast convergence characteristics of heuristic algorithms and the global search capability of genetic algorithms, improving the efficiency of solving large-scale supply chain optimization problems. At the 22nd and 68th iterations, the fitness values reached as high as 99.67 and 96.77, respectively, demonstrating the fast convergence and stability of the algorithm. This result is similar to the research findings of scholar Charles D. [18]. In addition, the research method achieved a high accuracy of 98.56% on the training set, significantly better than other existing algorithms. This result is significantly better than Goodarzian F and Ebrahim et al. [20-22]. The proposed method can maintain low management costs and high operational efficiency on both small-scale and large-scale datasets, which has important guiding significance for supply chain management practices. By minimizing supply chain management costs and maximizing efficiency, companies can meet environmental protection and social responsibility requirements while pursuing economic benefits. This is particularly important in the current context of globalization and rapid technological development.

B. Conclusion

To optimize supply chain management and achieve the dual goals of cost efficiency and sustainability, this study proposes a supply chain management approach that integrates a Mixed Integer Programming Model with an improved GA. Firstly, by employing a supplier selection method based on heuristic algorithms, the study successfully generated initial feasible solutions. Subsequently, an optimization method based on adaptive GA was introduced to better adapt to and improve the complexity and dynamism of supply chain management. The data indicates that, on both the validation and training sets, the research method achieved maximum fitness values of 99.67 and 96.77, respectively, when the system iterated to the 68th and 22nd times. In contrast, the fitness values of other methods were significantly below 90.0. On the validation set, the accuracy of the research method reached a maximum of 98.33%. Through five parallel experiments designed, the minimum T1 time and minimum T2 value of the research method were 0.654s and 0.643s, respectively. In practical applications, when the data volume increased to 191, the research method demonstrated the minimum management cost of 97.6 thousand RMB; however, when the data volume reached 268, the FAHP-MILP algorithm exhibited the minimum management cost, reaching 135.8 thousand RMB. Additionally, when iterated to the 99th time, the research method incurred the least total cost, amounting to 962.7 thousand RMB. In summary, the research method not only reduced costs but also enhanced the flexibility and responsiveness of the supply chain. However, the experiment's choice of a single research enterprise, and factors such as supplier selection and product order allocation, which are influenced by various factors, have not been analyzed. Further application and research expansion are needed in the future.

References

[1] Lotfi R, Kargar B, Rajabzadeh M, Hesabi F, & Özceylan E. Hybrid fuzzy and data-driven robust optimization for resilience and sustainable health care supply chain with vendor-managed inventory approach. International Journal of Fuzzy Systems, 2022, 24(2): 1216-1231.

[2] Hua Wang. Research on the Influencing Factors of Block Chain Technology Adoption in Supply Chain Finance of Small and Medium-Sized Enterprises. Advanced Management Science, 2023, 12(1).

[3] Hebbi C, Mamatha H. Comprehensive Dataset Building and Recognition of Isolated Handwritten Kannada Characters Using Machine Learning Models. Artificial Intelligence and Applications, 2023, 1(3):179-190.

[4] Fakhrzad M B, Goodarzian F. A new multi-objective mathematical model for a Citrus supply chain network design: Metaheuristic algorithms. Journal of Optimization in Industrial Engineering, 2021, 14(2): 111-128.

[5] Brahami M A, Dahane M, Souier M, Sahnoun M H. Sustainable capacitated facility location/network design problem: a non-dominated sorting genetic algorithm based multiobjective approach. Annals of Operations Research, 2022, 311(2): 821-852.

[6] Dwivedi A, Jha A, Prajapati D, Sreenu N, & Pratap S. Meta-heuristic algorithms for solving the sustainable agro-food grain supply chain network design problem. Modern Supply Chain Research and Applications, 2020, 2(3): 161-177.

[7] Sadeghi A, Mina H, Bahrami N. A mixed integer linear programming model for designing a green closed-loop supply chain network considering location-routing problem. International journal of logistics systems and management, 2020, 36(2): 177-198.

[8] Manupati V K, Schoenherr T, Ramkumar M, Wagner S M, Pabba S K, & Inder Raj Singh R. A blockchain-based approach for a multi-echelon sustainable supply chain. International Journal of Production Research,

2020, 58(7): 2222-2241.

[9] Isaloo F, Paydar M M. Optimizing a robust bi-objective supply chain network considering environmental aspects: a case study in plastic injection industry. International Journal of Management Science and Engineering Management, 2020, 15(1): 26-38.

[10] Mogale D G, Ghadge A, Kumar S K, Tiwari M K. Modelling supply chain network for procurement of food grains in India. International Journal of Production Research, 2020, 58(21): 6493-6512.

[11] Ahmadini A A H, Modibbo U M, Shaikh A A, Ali I. Multi-objective optimization modelling of sustainable green supply chain in inventory and production management. Alexandria Engineering Journal, 2021, 60(6): 5129-5146.

[12] Zhang Y, Kou X, Song Z, Fan Y, Usman M, & Jagota V l. Research on logistics management layout optimization and real-time application based on nonlinear programming. Nonlinear Engineering, 2022, 10(1): 526-534.

[13] Beiki H, Mohammad Seyedhosseini S, V. Ponkratov V, Olegovna Zekiy A, & Ivanov S A. Addressing a sustainable supplier selection and order allocation problem by an integrated approach: a case of automobile manufacturing. Journal of Industrial and Production Engineering, 2021, 38(4): 239-253.

[14] Li C, Conejo A J, Liu P, Omell B P, Siirola J D, & Grossmann I E. Mixed-integer linear programming models and algorithms for generation and transmission expansion planning of power systems. European Journal of Operational Research, 2022, 297(3): 1071-1082.

[15] Rivera Letelier O, Espinoza D, Goycoolea M, Moreno E, & Muñoz G. Production scheduling for strategic open pit mine planning: a mixed-integer programming approach. Operations Research, 2020, 68(5): 1425-1444.

[16] Mokayed, H., Quan, T. Z., Alkhaled, L., & Sivakumar, V. Real-time human detection and counting system using deep learning computer vision techniques. Artificial Intelligence and Applications. 2023, 1(4): 221-229.

[17] Kumar R, Ganapathy L, Gokhale R, Tiwari M K. Quantitative approaches for the integration of production and distribution planning in the supply chain: a systematic literature review. International Journal of Production Research, 2020, 58(11): 3527-3553.

[18] Charles D. The Lead-Lag Relationship Between International Food Prices, Freight Rates, and Trinidad and Tobago's Food Inflation: A Support Vector Regression Analysis. Green and Low-Carbon Economy, 2023, 1(2): 94-103.

[19] Soleimani H, Chhetri P, Fathollahi-Fard A M, Mirzapour Al-e-Hashem S M J, & Shahparvari S. Sustainable closed-loop supply chain with energy efficiency: Lagrangian relaxation, reformulations and heuristics[J]. Annals of Operations Research, 2022, 318(1): 531-556.

[20] Goodarzian F, Hosseini-Nasab H, Fakhrzad M B. A multi-objective sustainable medicine supply chain network design using a novel hybrid multi-objective metaheuristic algorithm. International Journal of Engineering, 2020, 33(10): 1986-1995.

[21] Poursoltan L, Mohammad Seyedhosseini S, Jabbarzadeh A. A two-level closed-loop supply chain under the constract of vendor managed inventory with learning: a novel hybrid algorithm. Journal of Industrial and Production Engineering, 2021, 38(4): 254-270.

[22] Ebrahim Qazvini Z, Haji A, Mina H. A fuzzy solution approach to supplier selection and order allocation in green supply chain considering the location-routing problem[J]. Scientia Iranica, 2021, 28(1): 446-464.

DBRF: Random Forest Optimization Algorithm Based on DBSCAN

Wang Zhuo, Azlin Ahmad*
School of Computing Sciences-College of Computing-Informatics and Mathematics,
Universiti Teknologi MARA, Shah Alam, Malaysia

***Abstract*—The correlation and redundancy of features will directly affect the quality of randomly selected features, weakening the convergence of random forests (RF) and reducing the performance of random forest models. This paper introduces an improved random forest algorithm—A Random Forest Algorithm Based on DBSCAN (DBRF). The algorithm utilizes the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to improve the feature extraction process, to extract a more efficient feature set. The algorithm first uses DBSCAN to group all features based on their relevance and then selects features from each group in proportion to construct a feature subset for each decision tree, repeating this process until the random forest is built. The algorithm ensures the diversity of features in the random forest while eliminating the correlation and redundancy among features to some extent, thereby improving the quality of random feature selection. In the experimental verification, the classification prediction results of CART, RF, and DBRF, three different classifiers, were compared through ten-fold cross-validation on six different-sized datasets using accuracy, precision, recall, F1, and running time as validation indicators. Through experimental verification, it was found that DBRF algorithm outperformed RF, and the prediction performance was improved, especially in terms of time complexity. This algorithm is suitable for various fields and can effectively improve the classification prediction performance at a lower complexity level.**

Keywords—Random forest; DBSCAN; feature selection; feature redundancy; classification algorithm

I. Introduction

The correlation and redundancy of features will directly affect the performance of the random forest model. Especially in high-dimensional features, contain a lot of information, but may also contain a lot of useless, correlated or redundant features, making it difficult to distinguish between important and unimportant features, leading to an increase in the computational complexity of the machine learning model, an increase in the time overhead, a decrease in generalization ability, and a tendency to overfit the model [1-3].

Random forest (RF) is a hybrid classification algorithm that uses random sampling and random selection of features to construct multiple decision trees, making the model highly stable. Compared with other classification algorithms, RF has higher classification accuracy, lower generalization error, and faster training speed, so it has been widely applied in the field of data mining in many aspects. Random forest is a general-purpose algorithm with broad application potential in different fields. It has a large number of application cases in disease gene prediction, soil moisture estimation, industrial robot fault diagnosis, and text classification [1, 4-6]. The RF algorithm has many advantages such as high accuracy and strong generalization, but also has limitations. When dealing with high-dimensional data, its feature random selection mechanism causes poor correlation between the selected features and the category variables. In addition, the randomly selected feature variables may have high redundancy, which directly affects the quality of the feature subset in the random forest and weakens the convergence of the random forest, reducing the accuracy, generalization ability, and performance of the random forest model. Most of the current studies solve this problem by preprocessing and feature selection, but this may lead to new problems such as information loss and a dramatic increase in model complexity. These studies often only focus on the algorithm itself and do not consider its practical value. This study aims to improve the performance and efficiency of random forest model prediction by improving the method of random feature selection. This study designs and implements an improved random forest algorithm based on density clustering algorithm and hierarchical feature extraction mechanism. The significance of this study lies in significantly improving the accuracy and complexity of the random forest model prediction, providing practical and theoretical solutions and foundations for the sustainable development of this technology.

The main contribution of this study is to propose an improved random forest algorithm based on DBSCAN and stratified random sampling. This improved method enhances the quality of randomly selected feature subsets, and also confirms that the algorithm improves the accuracy of random forest models in different scale datasets, showing excellent performance in both cases. This study provides a new solution approach and empirical data for improving the random forest model.

The structure of this paper is as follows: Section II reviews related work, discusses the existing research on improving random forest algorithms and the progress in feature dimensionality reduction, and points out the shortcomings of existing studies. Section III provides a detailed description of the improved random forest algorithm design. Section IV validates the performance of the improved algorithm through experiments, including the evaluation of key performance indicators such as Accuracy, Precision, Recall, F1, and running time. Section V discusses the experimental results. Section VI summarizes the theoretical and practical significance of the research findings and provides a look ahead to future research directions.

*Corresponding Author.

II. Related Work

Many scholars have conducted relevant research on high-dimensional feature problems. Tang, Zhang et al. [7] used Relief F to calculate feature weights and used the Sequential Backward Selection algorithm to remove redundant features and weakly correlated features. The experiment proved that this method can effectively reduce redundant features. Compared with the methods of support vector machines, AdaBoost, and random forests, it has higher classification accuracy and efficiency. Ahmed, Deo et al. [4] proposed a soil moisture estimation model that uses the Boruta algorithm for feature selection. The model determines which features are significant by comparing the importance of the original features with the importance of the randomly generated shadow features. The experiment proved that the model has feature selection ability.

Rani and Baulkani [8] proposed the Lasso with Graph Kernel Feature Selection (LGKFS) algorithm, which combines the sparsity of Lasso regression and the structural information of GK-FS to reduce the feature dimension. When dealing with complex medical imaging data, the feature dimension is often very high, which may lead to the risk of overfitting in classification models and increase computational complexity. Therefore, effective feature selection becomes a key to improving classification performance. LGKFS algorithm combines Lasso regression and GK-FS algorithm to select the most valuable feature subset from high-dimensional features, thereby reducing the feature dimension and improving classification accuracy. Lasso regression is applied to the extracted features for sparse selection, removing most of the non-important features. GK-FS algorithm is used to further select the features after Lasso screening, based on graph kernel functions to calculate the similarity between features and select the most representative feature subset.

Jalal, Mehmood et al. [6] used boosted sampling and random subspace methods to remove unimportant features, dynamically increasing the number of trees to improve text classification performance. Each feature was assigned a weight, which reflected its importance in the classification task. Features were divided into important and unimportant features based on a set threshold. The choice of threshold depends on the distribution of feature weights and the performance requirements of the model. In each iteration, the random forest was updated based on the classified features. Important features were retained and used to construct new decision trees, while unimportant features were excluded. Meanwhile, the optimal number of trees was sought. This was achieved by gradually increasing the number of trees and evaluating the model performance until the optimal classification effect was reached.

Theerthagiri and Ruby [9] proposed a random forest feature selection algorithm based on recursive feature elimination and voting technology. The importance of each feature is evaluated by recursively building a random forest to assess the importance of each feature. The importance of a feature is evaluated by how it affects the prediction result during the decision tree building process.

Wang, Xue et al. [10] proposed a feature selection method based on variable-sized cooperative evolution particle swarm optimization. It includes a spatial division strategy based on feature importance, an adaptive mechanism for adjusting subgroup size, and a feature deletion and generation strategy based on fitness guidance, using the maximum information coefficient (MIC) to evaluate feature importance. Features with larger MIC values are moved to the set U, and the features in U are sorted and clustered based on their MIC values. Redundant features are deleted through the clustering results.

In high-dimensional data scenarios, to improve the accuracy of biomass estimation using a random forest algorithm, Zhang, Shen et al. [11] used an improved Random Forest algorithm by adding two regularization terms to further control the complexity of the model and improve performance. The L1 regularization selects the sum of the absolute values of the model parameters as the penalty term, thus selecting the most important feature at each node. This method helps to select the features that contribute the most to the model's prediction, reducing the influence of irrelevant or redundant features on the model. The average depth regularization term controls the depth of the tree and the number of nodes, thus limiting the complexity of the model. This limitation reduces the risk of overfitting the training data and improves the model's generalization ability. By limiting the depth and number of nodes, the model is more cautious in the feature selection process and avoids introducing noisy features due to the model being too complex.

To improve the performance of speech emotion classification, Xie, Zhu et al. [2] proposed a two-stage feature selection method based on random forest and grey wolf optimization. In the random forest algorithm, the importance of a feature is calculated based on its ability to increase the purity of leaf nodes. Then, the feature subset with the highest classification accuracy and the least number of features is selected through the iterative process of grey wolf optimization, which is used as the final optimal feature subset.

In summary, the quality of features affects the results of classification prediction, and many experts have conducted a series of optimization studies on feature dimensionality reduction. Currently, most studies mainly use feature selection algorithms to reduce the number of features, but ranking-based feature selection and subset-based feature selection both have certain limitations. Ranking-based feature selection algorithms mainly focus on the importance of individual features and ignore the interaction between features and the overall structure, while subset-based feature selection algorithms consider the combination of features, but may face the problem of large computational complexity and overfitting. Based on these problems, this study combines density clustering algorithm and hierarchical extraction to optimize the random forest algorithm. On the basis of establishing the diversity of random feature selection in the RF algorithm, it eliminates the interference of correlated and redundant features and builds a more predictive random forest, thereby improving the comprehensive performance of the model prediction.

III. Methodology

The RF algorithm has many advantages such as high accuracy and strong generalization, and has wide applications. However, when dealing with high-dimensional data sets, its random feature extraction mechanism reduces the correlation between features and category variables. Moreover, the randomly extracted features may have high redundancy, which lowers the quality of the random feature subset, and weakens the convergence of the random forest, thereby reducing the overall performance of the random forest. Therefore, this paper optimizes the traditional RF algorithm by using the DBSCAN algorithm and hierarchical sampling to change the feature extraction mechanism. By constructing similar feature groups, it reduces the impact of these factors and improves the efficiency of the algorithm.

A. *Random Forest*

Random Forest is a machine learning algorithm that combines multiple decision trees. RF selects multiple subsets of the original sample set by random sampling with replacements from the set to build decision trees. At each node in the decision tree, a random subset of *k* attributes is selected from the set of attributes at the node, and then the best attribute is chosen from this subset for splitting. It is generally recommended that $k = log_2 d$, where *d* is the number of features in the data set [12]. The predictions of each tree are voted on to elect the best result. RF can handle both continuous and categorical variables [13, 14]. It can also rank the importance of features [15].

The architecture of the random forest algorithm is depicted in Fig. 1, and its underlying principles are delineated as follows [16]:

1) Randomly draw n training datasets from the original dataset with replacement.

2) Randomly select *K* features from each training dataset (where *K* is less than the total number of features in the original dataset).

3) Employ a specific strategy (e.g., Gini coefficient) to choose 1 feature from the *k* features as the splitting feature for the node, thereby constructing a decision tree.

4) Iterate through steps 1-3 to construct n decision trees.

5) Utilize each decision tree for result prediction.

6) Aggregate predictions and determine the final prediction result based on majority voting.

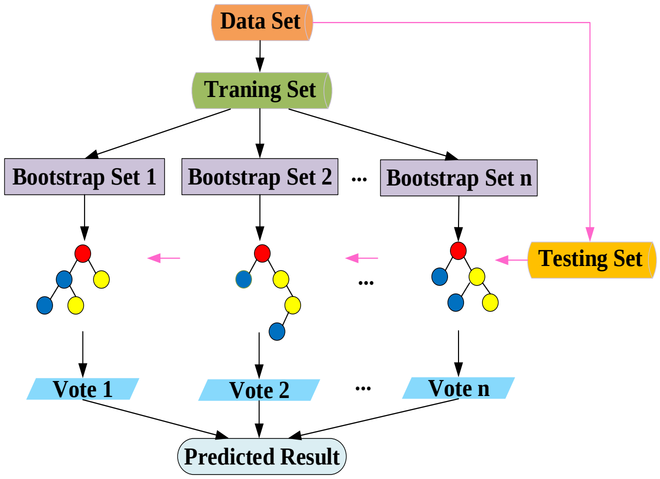


Fig. 1. The architecture of the random forest algorithm.

Random forest uses CART as a single classifier. CART uses the Gini coefficient as a selection criterion for splitting features. The key to building decision trees in random forest is to choose the optimal splitting feature, seeking higher and higher node "purity" in the splitting process [17]. The lower the Gini coefficient of a feature, the lower its impurity, and the feature with the lowest impurity is selected for node splitting. The impurity calculation is repeated for each node. After each split, the overall impurity of the tree decreases, until no features are available or the impurity has reached an optimal level, at which point the decision tree stops growing. The formula for calculating the Gini index is [18]:

$$GI_m = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{1}$$

Where: *K* denotes the number of categories, and P_{mk} represents the proportion of class column *k* in node *m*.

B. *DBSCAN Algorithm*

DBSCAN is a density-based clustering algorithm based on high-density connected regions. It defines clusters as the largest set of points that are densely connected and can group together regions with sufficiently high density and discover arbitrary-shaped clusters in noisy spatial databases. It uses parameters (*Eps*, *MinPts*) to describe the tightness of the sample distribution in the neighborhood. Parameter Eps is the maximum radius of the neighborhood. Parameter *MinPts* specifies the density threshold for dense regions. The working principle of DBSCAN is: randomly select a data point p from the dataset and check whether p's Eps neighborhood contains the minimum number of data points *MinPts*. If this condition is met, a new cluster is created and all identified data are added to the new clustering. Then, all data within the cluster will also be checked in the same way based on these two parameters, in order to add as many other data as possible that have not been checked before. This process is repeated until all data in the dataset are accessed [19, 20].

C. *Stratified Sampling*

Divide the entire sample into distinct strata or categories, and subsequently conduct random sampling from each stratum by selecting a specific number of individuals. Finally, combine the sampled individuals from all strata to form a representative sample. This approach is known as stratified sampling. Through categorization and stratification, it enhances the similarity among individuals within each category, facilitating the selection of a representative survey sample. This method is particularly suitable for complex situations with substantial individual variations and a large population size. The key characteristics of stratified sampling include [21, 22]:

1) Stratification involves the classification of similar individuals into distinct layers, with each layer representing a unique category. This method adheres to the principle of non-overlapping and exhaustive coverage, ensuring that every individual is assigned to one and only one layer.

2) To guarantee an equal opportunity for every individual to be included in the sample, stratified sampling necessitates simple random sampling within each layer. The sample size in each layer is determined proportionally based on the total number of individuals in that layer relative to the overall population size.

D. Design of the DBRF Algorithm

The establishment of a random forest involves two key random processes, one of which is the random selection of features. In high-dimensional datasets, it is highly probable to extract a large portion of irrelevant or redundant features, leading to a decrease in the generalization and accuracy of the RF algorithm. To address this issue, this study proposes an improvement to the traditional random forest algorithm using density clustering. Without removing redundant features and while retaining the original feature information, features are grouped (clustered) based on density to form similar feature groups TG = "TG_1 ,TG_2,…,TG_n". Within these similar feature groups, a certain proportion of features can represent all information for that entire group of features as well as express classification labels C. Features are randomly selected from each similar feature group TG_i in proportion to establish a subset for building individual decision trees. The architecture of the DBRF algorithm is illustrated in Fig. 2.

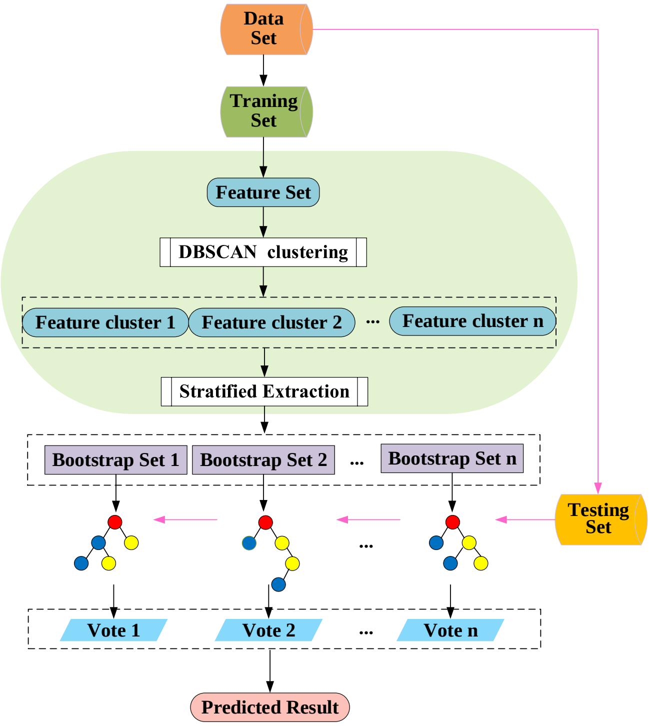


Fig. 2. The architecture of the DBRF algorithm.

The flowchart of the DBRF algorithm is illustrated in Fig. 3. Initially, a training set and a test set are established. Algorithm parameters such as the number of decision trees (*n_estimators*), maximum number of features for splitting (*max_features*), minimum samples in a cluster (*min_samples*), and neighborhood radius (*eps*) are configured. The Gini coefficient (GI_m) for each feature is computed, followed by DBSCAN clustering to form F similar groups. Features are then extracted from each group based on the proportion NF, and a subset of features is selected from other similar feature groups using the same approach to construct individual decision trees. This process is iterated multiple times until reaching the desired scale for constructing the random forest, at which point it terminates.

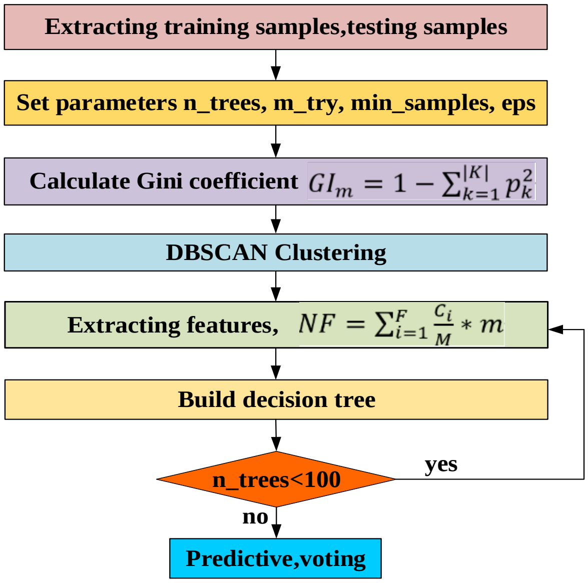

Extracting training samples,testing samples
Set parameters n_trees, m_try, min_samples, eps
Calculate Gini coefficient $GI_m = 1 - \sum_{k=1}^{|K|} p_k^2$
DBSCAN Clustering
Extracting features, $NF = \sum_{i=1}^{F} \frac{C_i}{M} * m$
Build decision tree
n_trees<100
yes
no
Predictive,voting

Fig. 3. The flowchart of DBRF.

The flowchart of feature extraction is shown in Fig. 4. After features are clustered, features with similar classification capabilities are grouped into a cluster. Then, features are sampled proportionally from each cluster. This ensures that the extracted features are more representative and do not favor any particular situation. These features are used to build a decision tree. The feature extraction process is repeated until all decision trees have been built.

The formula for proportional sampling is:

$$NF = \sum_{i=1}^{F} \frac{C_i}{M} * m \qquad (2)$$

Where: F is the total number of clusters, C_i is the number of features in the *i-th* cluster, M is the total number of features, and m is the number of features to be extracted.

The pseudocode for feature proportional stratified sampling is as follows:

Algorithm 1: stratified sampling algorithm

Input: a set of similar feature clusters

Output: several groups of extracted features.

Method:

(1) According to the total number of features N and the number of features per layer: n_i , Calculate the sampling ratio for each layer $W = \frac{n_i}{N}$.

(2) Calculate the number of features to be extracted from each layer: $NUM = W*n$. And make sure that the total number of features extracted from each layer is n.

(3) Determine the number of features for each layer, then randomly select features from each layer to form a total of n samples.

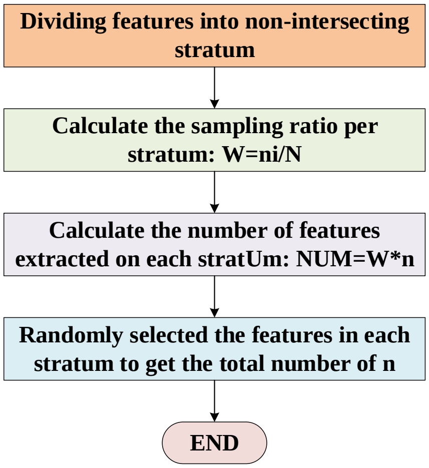

Dividing features into non-intersecting stratum
Calculate the sampling ratio per stratum: W=ni/N
Calculate the number of features extracted on each stratUm: NUM=W*n
Randomly selected the features in each stratum to get the total number of n
END

Fig. 4. The flow of feature extraction.

Therefore, the pseudocode for the complete DBRF algorithm is as follows:

```
Algorithm 2: DBRF algorithm
Input: Dataset: Data, number of decision trees: n_estimators,
maximum number of features for splitting: max_features,
minimum number of samples in a cluster: min_samples,
neighborhood radius: eps
Output: An RF classifier
Methods:
The dataset is divided into a training set and a testing set;
for i=1 to Num_Features
    Compute the Gini coefficient ($GI_m$) for each feature.
endfor
while( j<= min_samples and k<=eps)
    F = DBSCAN($GI_m$)    // Feature clustering
endwhile
for t=1 to n_estimators
    for f=1 to F
        Extract NF features and construct a decision tree;
    endfor
endfor
```

The time complexity of RF is *O(tfnlog(n))*, where *t* is the number of decision trees built, *f* is the number of features selected at each node, and *n* is the number of samples in the training set [23]. The DBRF algorithm proposed in this paper is divided into two parts: clustering to build similar feature groups and building a random forest model. For the first part, the time is mainly spent on feature Gini coefficient clustering, with a time complexity of *O(mlogm)* [24], where m is the number of features. The second part is the random forest construction. Therefore, the time complexity of the DBRF algorithm proposed in this paper is the sum of the two parts, i.e., *O(tfnlog(n)+m(logm))*.

IV. Experimental Results and Analysis

A. Experimental Data

In order to objectively and comprehensively evaluate the effectiveness and advantages of the DBRF algorithm, the adaptability of the algorithm on different feature dimension datasets was analyzed. Six datasets with different feature dimensions from the UCI were selected, namely SPECT Heart (SPECT), Chess, SCADI, DARWIN, Period Changer (Period), and MicroMass. Table I describes the detailed information of the six datasets. The six datasets were divided into low-dimensional, medium-dimensional, and high-dimensional datasets based on the size of the samples and features [25]. SPECT and Chess belong to low-dimensional feature datasets. SCADI and DARWIN belong to medium-dimensional datasets. Period and Micromass are high-dimensional datasets. At the same time, these datasets include balanced and unbalanced datasets.

B. Experimental Results and Analysis

The experimental environment was set to Windows 11 operating system (64-bit), Intel(R) Core(TM) i7-10510U CPU, 16GB RAM, and Visual Studio Code. To verify the comprehensive performance of the proposed improved random forest, the experimental results of DBRF, RF, and CART classifiers were compared. The experiment used tenfold cross-validation to evaluate the accuracy, precision, recall, F1 score, and running time of the DBRF, RF, and CART models. Their overall performance was compared, highlighting the advantages of the improved algorithm. The experiment parameters were set to *n_estimators=100, max_depth=30, max_features= sqrt(n_features)*. The DBSCAN parameters were set to *min_samples=3, eps=0.02 or eps=0.03.*

TABLE I. The Descriptions of All Datasets

ID	DataSet	Feature Size	Sample Size	Feature Scale	Sample Scale	Balance	DOI
1	SPECT	22	267	Small	Small	unbalance	10.24432/C5P304
2	Chess	36	3196	Small	Large	balance	10.24432/C5DK5C
3	SCADI	205	70	Middle	Small	unbalance	10.24432/C5C89G
4	DARWIN	451	174	Middle	Small	balance	10.24432/C55D0K
5	Period	1177	90	Large	Small	unbalance	10.24432/C5B31D
6	Micromass	1300	571	Large	Middle	balance	10.24432/C5T61S

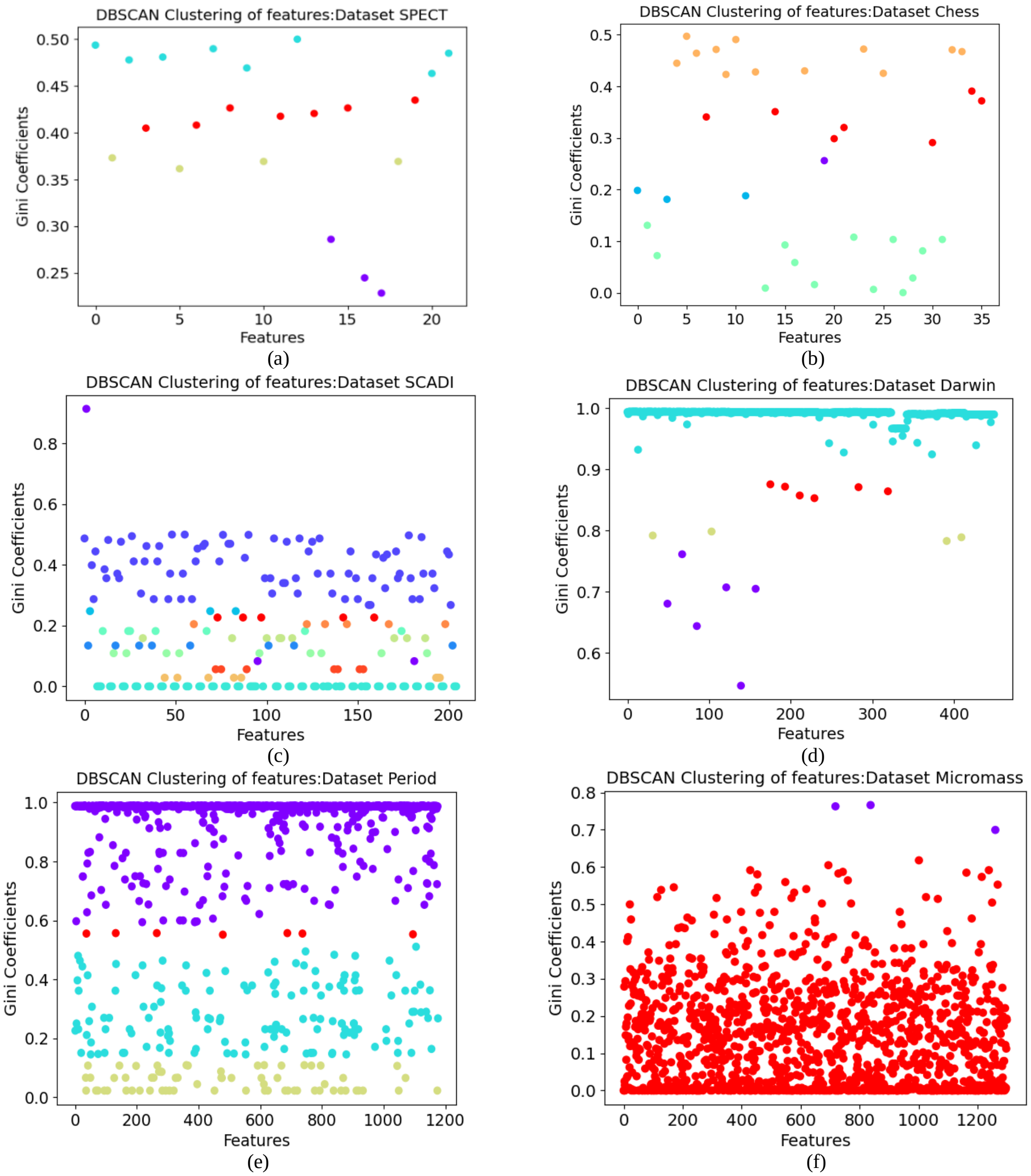


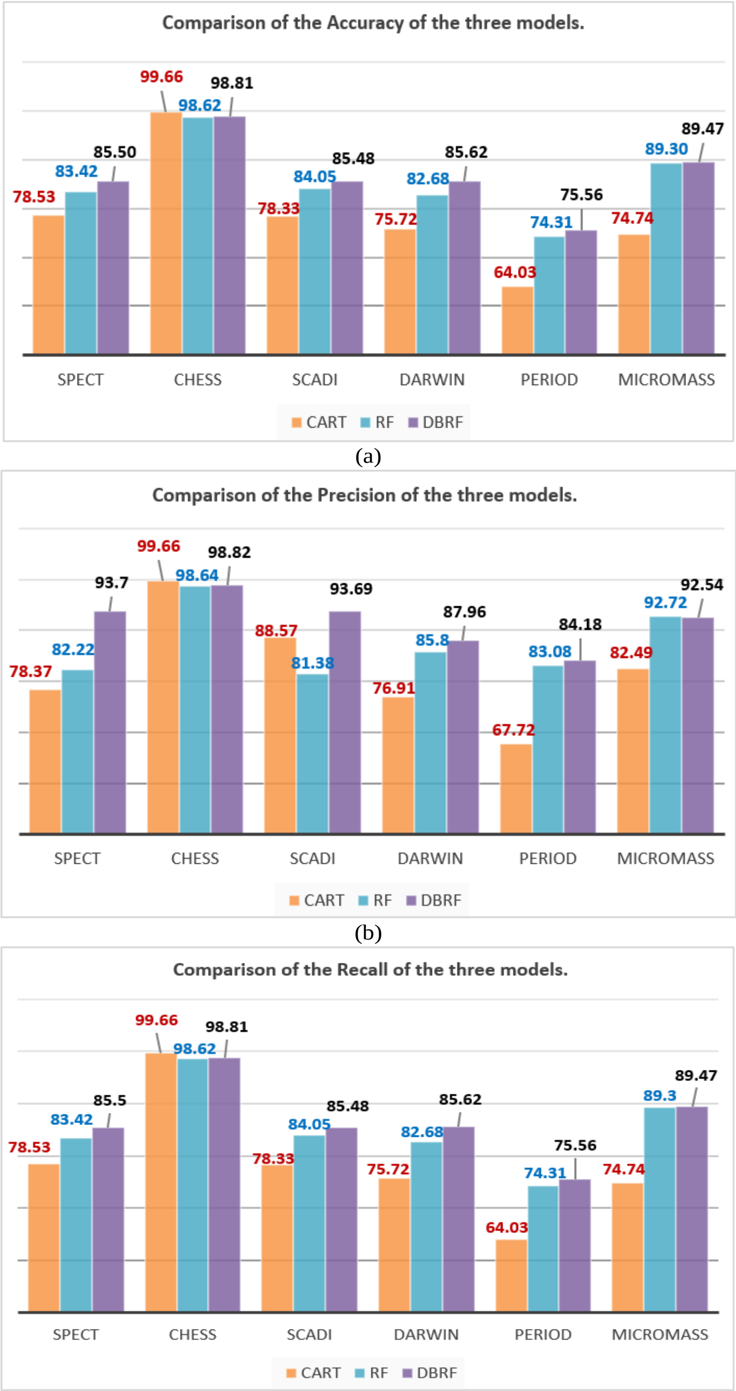
Fig. 5. Clustering results of DBRF on all Datasets (a) SPECT, (b)Chess, (c) SCADI, (d) DARWIN, (e) Period and (f) Micromass.

Fig. 5 illustrates the clustering results of DBRF on all datasets. Features clustered together are represented by points of the same color. Table II presents the number of feature clusters for each dataset, along with the maximum and minimum values of elements within each cluster. SCADI exhibits the highest number of similar feature clusters, with 12 clusters containing a maximum of 77 features and a minimum of three features. SPECT, Darwin, and period are all clustered into four similar feature clusters, with the maximum number of features in a cluster being 958 and the minimum being 3. MicroMass has the fewest feature clusters at only 2, with a maximum of 1297 features in a cluster and a minimum of three features. Chess consists of 36 features clustered into five groups, with each group containing between 1 to 13 features.

Fig. 6 shows the performance comparison of CART, RF, and DBRF models on all datasets. Fig. 6(a) shows the classification accuracy of CART, RF, and DBRF prediction models on all datasets. On the chess dataset, the accuracy of DBRF is 0.85% lower than that of CART and 0.19% higher than that of RF. For the other five datasets, the accuracy of DBRF is the highest. DBRF is 6.97% to 14.73% higher than CART and 0.17% to 2.94% higher than RF. This data confirms that DBRF as a composite classifier of CART is superior to CART. Furthermore, the higher accuracy of DBRF than RF indicates that the features extracted by DBRF are more representative and have higher accuracy. Importantly, DBRF performs well on both balanced and unbalanced datasets, demonstrating its versatility. Furthermore, DBRF improves performance on low-dimensional and high-dimensional datasets.

TABLE II. THE CLUSTERING RESULTS OF DBRF

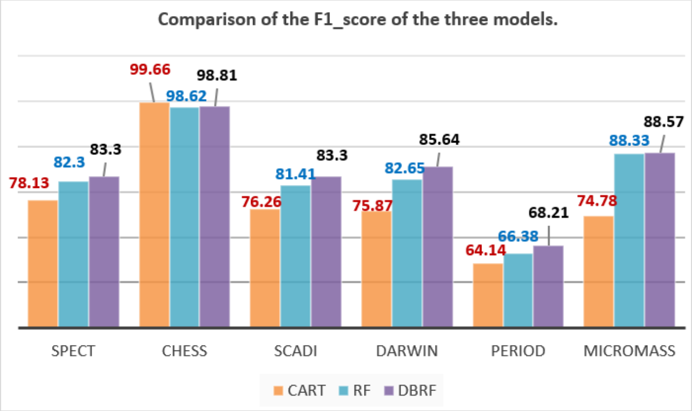
ID	**DataSet**	**Number of clusters**	**Maximum number of features in a cluster**	**Minimum number of features in a cluster**
1	SPECT	4	8	3
2	chess	5	13	1
3	SCADI	12	77	3
4	darwin	4	434	4
5	period	4	958	7
6	micromass	2	1297	3



(a)

(b)

(c)



(d)

Fig. 6. Performance comparison of CART, RF, and DBRF models on all datasets.

Fig. 6(b) shows that the accuracy of DBRF on SCADI is 93.69, while the accuracy of RF is 81.38, which has been significantly improved by 12.31%. In Table II, the number of clusters in the dataset is the maximum value of 12, proving the complexity of the feature distribution. The clustering process can extract comprehensive and representative features, thus significantly improving the accuracy. At the same time, the accuracy of DBRF is the highest in all five datasets, with an improvement of 15.33%, 12.31%, 11.05%, 16.46%, and 10.05% compared to the lowest accuracy.

Fig. 6(c) compares the Recall values of the three models. Similarly, DBRF is at the highest level of Recall value. In the five datasets, DBRF's recall value is absolutely dominant, far higher than the RF and CART models. Compared with the lowest value, the increase in DBRF's Recall value is 6.97%, 7.15%, 9.9%, 11.53%, and 14.73%.

Fig. 6(d) shows the combined measure F1. In the low-dimensional datasets SPECT and Chess, DBRF performed 1.0% and 0.19% better in F1 than RF, respectively. However, the improvement was not significant due to fewer features in the low-dimensional datasets. In the medium-dimensional datasets SCADI and Darwin, DBRF's F1 score was 1.89% and 2.99% higher than RF, respectively. In the five datasets, the improvement in DBRF's F1 value compared to the lowest value was 5.17%, 7.04%, 9.77%, 4.07%, and 13.79%, respectively. In the high-dimensional datasets Period and MicroMass, DBRF improved the F1 score by 1.83% and 0.24% compared to RF, respectively. In the medium- and high-dimensional datasets, CART always performed the worst, thereby highlighting the advantages of ensemble learning models.

Table III shows the running times of the three prediction models on all datasets. For the four datasets including the high-dimensional dataset MicroMass, the DBRF model requires less time than the RF model, further emphasizing its efficiency and universality. Although DBRF adds the feature clustering process, it reduces the running time, indicating that balancing the extraction of typical features is more beneficial for the time efficiency of prediction classification. Since both DBRF and RF require the construction of 100 decision trees, their running times are longer than CART, but they achieve higher accuracy. Therefore, from the comprehensive performance indicators, the prediction classification effect of the DBRF model is better than that of the other two models.

TABLE III. RUNNING TIME OF THE THREE MODELS IN SECOND (S)

Dataset	CART	RF	DBRF
SPECT	2.39	8.32	**5.46**
chess	7.07	84.83	145.25
SCADI	2.91	6.36	**5.82**
darwin	101.21	356.80	**276.60**
period	123.59	76.93	153.39
micromass	464.87	572.17	**526.49**

* Bold font is the best results.

V. DISCUSSION

This study used different-sized datasets and conducted comprehensive experimental evaluations to verify the effectiveness of the proposed optimization techniques. The performance of the DBRF algorithm was compared with that of traditional RF and CART algorithms. The experiment demonstrated the technical improvements brought about by density-based feature extraction, and the empirical evidence proved the classification efficiency, scalability, and time complexity. Previous research techniques were only applicable to a single application domain [2, 6, 8, 11], while this study tested the proposed method on datasets with multiple different neighborhoods. In low, medium, and high-dimensional datasets, the DBRF achieved significant improvements in all performance indicators compared with the other two models. The maximum improvement in accuracy indicators was 14.73% in the high-dimensional MicroMass dataset. The highest accuracy value was 98.81% in the low-dimensional Chess dataset. The maximum improvement in precision indicators was 16.46% in the high-dimensional period dataset. In the running time indicator, the DBRF model required less time than the RF model in four datasets, including the high-dimensional MicroMass dataset, further highlighting its superiority and generality. Although the DBRF increased the feature clustering process, it reduced the running time, indicating that balancing the extraction of representative features is more beneficial for the time efficiency of predictive classification. In the five datasets, the DBRF achieved the highest values for all four evaluation indicators, including accuracy rate. Therefore, by randomly selecting similar feature groups and extracting features from them, it is possible to effectively avoid the formation of redundant feature subsets in traditional RF and improve the accuracy and overall performance of predictive classification.

In summary, the DBRF algorithm proposed in this paper has better experimental effects than the other two algorithms, showing obvious advantages in high-dimensional data sets, low-dimensional data sets, and data sets with highly redundant features. Future research can further study the improvement of other types of clustering algorithms on random forest feature extraction to achieve higher efficiency and performance improvement. At the same time, it can be made more scalable to enable it to have a wider range of applications.

VI. Conclusion

Due to the correlation among features, redundancy, and a large amount of useless information, the overall performance of the machine learning model is affected. This study optimizes the traditional RF algorithm and proposes a DBRF algorithm based on DBSCAN. The experimental results show that the DBRF algorithm has a higher accuracy index improvement of 6.97%-14.73% and an F1 index improvement of 4.07%-13.79% compared with the other two models. In the 5 datasets, the accuracy rate and other four evaluation indicators of DBRF are the highest. For the four datasets including the high-dimensional dataset MicroMass, the DBRF model takes less time than the RF model, which demonstrates its significant advantage in time complexity. Therefore, the DBRF algorithm achieves the research goal of reducing the influence of feature correlation and redundancy on model performance. In future research, further exploration of other types of clustering algorithms for random forest feature extraction will be conducted to achieve higher efficiency and performance improvement, as well as stronger scalability.

References

[1] Ding, H., et al., RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification. Information Processing & Management, 2023. 60(2).

[2] Xie, J., M. Zhu, and K. Hu, Fusion-based speech emotion classification using two-stage feature selection. Speech Communication, 2023. 152.

[3] Zhang, M., et al., Multi-objective optimization algorithm based on clustering guided binary equilibrium optimizer and NSGA-III to solve high-dimensional feature selection problem. Information Sciences, 2023. 648.

[4] Ahmed, A.A.M., et al., LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4.5 and RCP8.5 global warming scenarios. Stochastic Environmental Research and Risk Assessment, 2021. 35(9): p. 1851-1881.

[5] Wu, Y., et al., Extracting random forest features with improved adaptive particle swarm optimization for industrial robot fault diagnosis. Measurement, 2024. 229: p. 114451.

[6] Jalal, N., et al., A novel improved random forest for text classification using feature ranking and optimal number of trees. Journal of King Saud University - Computer and Information Sciences, 2022. 34(6): p. 2733-2742.

[7] Tang, Q., et al., A Classification Method of Point Clouds of Transmission Line Corridor Based on Improved Random Forest and Multi-Scale Features. Sensors (Basel), 2023. 23(3).

[8] Rani, K.E.E. and S. Baulkani, Multi Variate Feature Extraction and Feature Selection using LGKFS Algorithm for Detecting Alzheimer’s Disease. Indian Journal Of Science And Technology, 2023. 16(22): p. 1665-1675.

[9] Theerthagiri, P. and A.U. Ruby, RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction. Expert Systems, 2022. 39(9).

[10] Wang, P., et al., Feature clustering-Assisted feature selection with differential evolution. Pattern Recognition, 2023. 140.

[11] Zhang, X., et al., Improved random forest algorithms for increasing the accuracy of forest aboveground biomass estimation using Sentinel-2 imagery. Ecological Indicators, 2024. 159.

[12] LEO, B., Random Forests. 2001.

[13] Tyralis, H., G. Papacharalampous, and A. Langousis, A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water, 2019. 11(5).

[14] Liang, H., et al., Overflow warning and remote monitoring technology based on improved random forest. Neural Computing and Applications, 2020. 33(9): p. 4027-4040.

[15] Amir Behnamian, K.M., Sarah N. Banks, Lori White, Murray Richardson, and Jon Pasher, A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values. 2017.

[16] Zhou, J., S. Huang, and Y. Qiu, Optimization of random forest through the use of MVO, GWO and MFO in evaluating the stability of underground entry-type excavations. Tunnelling and Underground Space Technology, 2022. 124.

[17] Zhiqiang Geng , X.D., Jiatong Li , Chong Chu , Yongming Han, Risk prediction model for food safety based on improved random forest integrating virtual sample. 2022.

[18] Urbano, M.N., R.F. Diego, and P. Paulo, A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier. Pattern Recognition Letters, 2017. 99: p. 21-31.

[19] Latifi-Pakdehi, A. and N. Daneshpour, DBHC: A DBSCAN-based hierarchical clustering algorithm. Data & Knowledge Engineering, 2021. 135.

[20] Hanafi, N. and H. Saadatfar, A fast DBSCAN algorithm for big data based on efficient density calculation. Expert Systems with Applications, 2022. 203.

[21] Iliyasu, R. and I. Etikan, Comparison of quota sampling and stratified random sampling. Biometrics & Biostatistics International Journal, 2021. 10(1): p. 24-27.

[22] Latpate, R., et al., Stratified Random Sampling. Advanced Sampling Methods, 2021: p. 37-53.

[23] Akhiat, Y., et al., A New Noisy Random Forest Based Method for Feature Selection. Cybernetics and Information Technologies, 2021. 21(2): p. 10-28.

[24] Ros, F., et al., Detection of natural clusters via S-DBSCAN a Self-tuning version of DBSCAN. Knowledge-Based Systems, 2022. 241.

[25] Wang, Y., S. Krishna Saraswat, and I. Elyasi Komari, Big data analysis using a parallel ensemble clustering architecture and an unsupervised feature selection approach. Journal of King Saud University - Computer and Information Sciences, 2023. 35(1): p. 270-282.

Enhancing Emergency Response: A Smart Ambulance System Using Game-Building Theory and Real-Time Optimization

Guneet Singh Bhatia[1], Azhar Hussain Mozumder[2], Saied Pirasteh[3], Satinder Singh[4], Moin Hasan[5]
Siemens Energy, Inc. QUAD 3, Orlando, Florida, USA[1]
Dept. of Information Science Engineering, Jain Deemed-to-be University, Bengaluru, India[2]
Institute of Artificial Intelligence, Shaoxing University, Shaoxing, China[3]
Dept. of Computer Applications, Lovely Professional University, Phagwara, India[4]
Dept. of Computer Science and Engineering, Jain Deemed-to-be University, Bengaluru, India[5]

Abstract—**Dispatching ambulances early and efficiently is paramount and difficult in the field of emergency medical services. In this regard, the paper designs a smart ambulance system based on game-building theory. The system employs an advanced Negamax algorithm for optimizing the dispatch of ambulances during emergencies. Besides traditional methods, real-time traffic data, patient condition severity, and dynamic resource allocation also improve the system further. With the integration of predictive analytics and real-time data, it allows dynamic adaptation to changing urban conditions, optimal resource allocation as well as minimizing response time. According to our simulations involving extensive scenarios, our Negamax-based system performs significantly better with respect to average response times when compared with traditional methods averagely reducing them by more than 50%, hence, showing double improvement. The study not only improves efficiency in the operation of emergency services but also presents an expandable framework that can be used for future developments in critical response systems thereby leading to their association with smart city infrastructure and AI-based predictive emergency management.**

Keywords—Emergency medical services; ambulance dispatch optimization; advanced game theory; Negamax algorithm; real-time optimization; predictive analytics

I. Introduction

When it comes to saving lives and improving patient outcomes in urban areas, the speed of action taken by emergency medical services (EMS) is very important [1]. Often, traditional ambulance dispatches rely on static positions and heuristic approaches which are simple and fail to take into account the dynamic nature of urban emergencies and traffic conditions [2]. This research is motivated by the need for a more sophisticated adaptive ambulance allocation approach that can respond to real time conditions, predict patterns of emergency, and optimize resource utilization within complex urban settings.

The subject of this study is to develop and implement an advanced system based on game theory for optimizing the dispatch of ambulances in large urban areas with a high population density, complex road networks, and changing traffic conditions. We go beyond mere distance optimization to consider factors such as real-time traffic information, historical emergency patterns, and dynamic health facility capacity changes. The primary objectives of this research work are as follows:

- To customize the Negamax algorithm (from game theory) [3], [4] with multi-factorial decision-making for ambulance dispatch optimization.
- To combine real-time data feeds like hospital capacities, emergency severity levels, and traffic conditions in the process of optimizing ambulance dispatches.
- To design a scalable platform for handling emergency response coordination throughout the city.

Efficient ambulance dispatch is undoubtedly a crucial aspect. In the cases of cardiac arrests, survival odds are lower by 7-10% per minute delay [5]. It shows how rapid response saves lives. There could be improved resource utilization through efficient dispatch systems which can reduce operational costs and improve coverage with existing resources [6]. In addition, the implementation of smart routing systems [7], [8] could reduce traffic congestions associated with emergency vehicle movements and facilitate urban mobility in general. Moreover, this data can help city planning and health care to optimize resource allocation based on facts instead of assumptions for a better public health outcome as well as improved emergency medical services.

The existing approaches are somehow limited on several accounts which reduce their effectiveness in modern urban environments. To be considered here, traditional ambulance dispatches are often based on static decision-making and use rule-based methods that are easy to apply but do not consider the evolving nature of the emergencies and traffic in urban environments [9], [10]. Although 5G and IoT have emerged and recently integrated into ambulances, they are still unable to relieve the demanding nature of emergency services in the urban environment [11], [12]. Various knowledge-based systems have been developed [13], but they are not capable enough to provide proper time-based optimization and predictive analysis. A few recent research works have tried to include real-time data and predictability in the model [14], [15]. However, there are very few studies that have included a holistic approach of adaptive methods. Moreover, though attempts have been made to solve problems like traffic congestion [16], still it is the dire necessity to design an efficient and enhanced adaptive ambulance allocation model considering the current situation, patterning emergency and the best utilization of resource in the context of urban environment. These drawbacks, which are discussed in more detail in Section II, all point to the need for more flexible, real-time, and big data-based solutions for ambulance dispatching.

To this motivation, we hypothesize that applying a customized version of the Negamax algorithm from game-building theory, combined with machine learning techniques for predictive analytics, can significantly improve the efficiency of ambulance services. We consider the Negamax algorithm because of its efficiency in exploring decision trees and adaptability to adversarial scenarios. In the context of emergency response, it represents the competition between different possible dispatch decisions. Moreover, the Negamax algorithm considers the thinking process of both participants while making a move, which consequently increases the win probability. Our customized version incorporates real-time data updates and considers multiple factors simultaneously, enhancing its suitability for the ambulance dispatch problem. Hence, this approach aims to create a dynamic, predictive, and highly responsive ambulance dispatch system. The system would be able to minimize response times, maximize resource utilization, and adapt to the complex and dynamic environment of urban emergencies.

The paper is organized as follows: Section II covers the literature review. Section III explains the mathematical modeling followed by the proposed system in Section IV. Section V is about experimental evaluation, results, and discussion. The paper is finally concluded in Section VI along with the future research considerations.

II. Literature Review

This section covers the related research works in the domain of smart ambulance system. A smart ambulance system was suggested by Gupta et al. in 2016 using IoT and smartphone technologies [9]. The research was aimed at enhancing the emergency medical response. They proposed a system that has two main modules: (i) Module 1 is about locating nearby ambulances and hospitals using GPS and Google Maps; (ii) Module 2 transmits real-time patient health data from an ambulance to a hospital. They also claimed that there were reduced response times and improved patient care during emergencies.

In 2017, Udawant et al. designed “Green Corridor” smart ambulance system using IoT framework to mitigate traffic congestion issues faced by emergency services [10]. The system reads patients vital signs in an ambulance while transmitting it to hospitals, as well as controls automatically signal lights for clear passage of vehicle when it reaches signals. Authors assessed different MAC protocols for data transmission in the proposed system concluding that CSMA is mostly efficient.

A timely ambulance service was proposed by Marimuthu et al. (2018) that employs the use of an Android application [17]. The proposed service allows for user requests of ambulances and selection of hospitals. Tracking the movement of ambulances in real-time is made possible through GPS and GSM modules while providing an emergency button to assign automatically the nearest ambulance. This application intends to enhance ambulance response durations and provide more effective life-saving services.

In 2021, Zhai et al. proposed a 5G-based smart ambulance structure and evaluated it through simulation experiments [11]. The experiment was conducted on the test platform which consisted of two scenarios namely, remote video consultation with medical data transmission from a moving ambulance and large medical image file transfers under both 4G and 5G networks. The resulting figures indicated impressive enhancements in capacity, speed, and latency for 5G as compared to 4G systems.

Merza and Qudr (2022) presented an ambulance-based healthcare system using Raspberry Pi and Internet connectivity to monitor patients’ vital signs in real-time for data transfer to hospitals [18]. The system incorporates various sensors for ECG, heart rate, respiration, temperature as well as audio/video monitoring thereby improving hospital readiness status and communication between paramedics on board with specialists on call.

In 2022, Sultana et al. defined an IoT-enabled intelligent ambulance routing system using LOADng-IoT routing protocol to reduce emergency response time and enhance patient care [12]. To speed up ambulances to hospitals yet keep transferring the most recent patients’ medical records, this is done by integrating traffic light control, health monitoring sensors and efficient path-finding algorithms. Additionally, authors discuss how the technology can facilitate achieving some of the UN SDGs concerning health, infrastructure and sustainable cities.

In 2023, Chanchai Thaijiam developed a smart ambulance system with knowledge base and decision-making support for improved rescue operations [13]. The design includes wearable biometric sensors, GPS tracking technology. Video conferencing platform was installed in order to have smooth communication between medical personnel in hospital and the emergency team inside ambulances. These systems enhance selection of destination hospitals for patients which is guided by an algorithm that uses decision trees procedure based on certain parameters such as distance or type of injuries.

Siddiqi et al. (2023) developed a smart signalization system for emergency vehicles [14]. The system uses Arduino, GSM modules, and a mobile application and it facilitates the drivers to control traffic signals from afar through SMS. Consequently, it minimizes any delays that may be caused. The system was evaluated by conducting field tests which proved the system's effectiveness for avoiding probable intersections for emergency vehicles. In addition, it maintains minimal waiting time for other traffic on the route.

In 2023, Sutherland and Chakrabortty proposed an optimal ambulance routing model [15]. The model considers multiple ambulances, patient medical severities, dispatching locations, and hospitals. The goal of this model is to enhance response times as well as patient transport efficiency. Simulation results prove the model's resilience under critical situations, therefore, laying a foundation for further studies on ambulance routing optimization.

In 2024, Sakthidevi et al. discussed IoT-enabled smart ambulances and how they can transform emergency response and management of patients [19]. The focus was on real-time monitoring sensors, advanced communication systems, and data processing platforms. The authors contributed to enhance resource allocation, improve response times, and elevate patient outcomes. This paper also focuses on future directions, challenges as well as potential impacts to emergency medical services in the world today.

In 2024, Jeyaseelan et al. put forward an IoT-based smart ambulance system for reducing the time taken in responding to emergencies in cities prone to traffic congestion [16]. The system employs the use of sensors, GPS and wireless communication technology to track ambulances, control traffic signals or even lower speed breakers automatically enabling ambulances to reach hospitals faster as well as safely. Experimental results show high accuracy and availability of the proposed system.

In the present research work, we address several key gaps in the above-reviewed literature. While previous studies have majorly focused on IoT integration, GPS tracking, and basic route optimization; in contrast, our approach leverages advanced game theory, specifically a customized Negamax algorithm, to provide a more adaptive and intelligent solution. The proposed system also incorporates predictive analytics and multi-factorial decision-making, whereas, earlier works primarily considered real-time data transmission and traffic signal control only. Doing this facilitates proactive resource allocation and efficient emergency response. Furthermore, existing research works have confined their scope to either route optimization or patient data transmission. In response, our research integrates these aspects with dynamic resource management across large and complex urban areas. The use of machine learning techniques for emergency prediction and traffic pattern analysis goes beyond the capabilities of systems described in previous literature.

III. Mathematical Modeling

This section describes our mathematical model considered for this research including the customized Negamax algorithm, possible constraints, dynamic updates, and predictive component.

A. Customized Negamax Algorithm

Let $G = (V, E)$ be a graph where V represents nodes (ambulance stations, hospitals, accident locations) and E represents edges (routes between nodes). A time-dependent distance matrix D is considered where $D[i][j][t]$ represents the estimated travel time between node i and node j at time t. Furthermore, a dynamic vector A is also considered where $A[i][t]$ represents the number of available ambulances at node i at time t. Let S be a severity matrix where $S[k]$ represents the severity level of emergency k. Associating all the above notations, the objective function to minimize the total weighted response time T is given in Eq. (1) as follows:

$$T = \sum_{k=1}^{N} min_{i \in V}(D[i][k][t] \cdot I[i][k] \cdot W(S[k])) \quad (1)$$

Where I[i][k] is an indicator function that is 1 if an ambulance from node i is dispatched to emergency location k, and 0 otherwise. W(S[k]) is a weight function based on the severity of the emergency.

For each emergency k, $D[i][k][t] \cdot I[i][k] \cdot W(S[k])$ is calculated for every possible dispatch location i. The minimum of these values is then selected which represents dispatching from the best location. This minimum is then weighted by the emergency's severity. It is done for all emergencies and sum of the results is calculated, giving us the total weighted response time T. The goal of optimization is to find the set of dispatch decisions (represented by $I[i][k]$ values) that minimizes this total weighted response time T, subject to the constraints discussed as follows.

B. Constraints

Following constraints are taken into account while modeling the system:

- Emergency Coverage: It is the first constraint where each emergency must be responded to by at least one ambulance (see Eq. (2)). loc_{emr} represents the emergency locations.

$$\sum_{\mathrm{i} \in \mathrm{V}} \mathrm{I[i][k]} \geq 1, \forall \mathrm{k} \in loc_{emr} \quad (2)$$

- Ambulance Availability: This constraint assures that the number of ambulances dispatched from any node cannot exceed the available ambulances at that node (see Eq. (3)).

$$\sum_{k \in loc_{emr}} I[i][k] \leq A[i][t], \forall i \in V \quad (3)$$

- Response Time Limit: In this constraint, it is assumed that the response time for each emergency should not exceed a maximum threshold T_max (see Eq. (4)).

$$D[i][k][t] \cdot I[i][k] \leq T_{max}, \forall i \in V, \forall k \in loc_{emr} \quad (4)$$

C. Dynamic Updates

Three parameters are dynamically updated in the system viz., traffic conditions, ambulance availability, and emergency severity. For traffic condition; historical data $data_{hist}$, real-time traffic $traf_{rt}$, and time t are considered as shown by the function in Eq. (5). Similarly, dispatch events evt_{dp}, return events evt_{ret}, and shift changes chg_{shf} are considered to update ambulance availability (see Eq. (6)). To update the emergency severity, reported condition $cond_{rep}$, historical data $data_{hist}$, and environmental factors $fact_{env}$ are considered (see Eq. (7)).

$$D[i][j][t] = f(data_{hist}, traf_{rt}, t) \quad (5)$$

$$A[i][t] = g(evt_{dp}, evt_{ret}, chg_{shf}) \quad (6)$$

$$S[k] = h(cond_{rep}, data_{hist}, fact_{env}) \quad (7)$$

D. Predictive Component

Function $P(l, t)$ represents the predictive component as shown in Eq. (8). It gives the probabilistic estimation of an emergency that may occur at location l at time t.

$$P(l, t) = ML\_model(data_{hist}, cond_{rep}, evt_{sch}) \quad (8)$$

It makes the system capable of anticipating where emergencies are likely to occur before they may actually happen. The ML_model is a machine learning algorithm that takes historical data $data_{hist}$, current conditions $cond_{rep}$, and scheduled events evt_{sch} as its input. Historical data consists of past emergency calls, their respective locations, times, and types. It assists the model in identifying different patterns and trends. Similarly, for the current conditions parameter, we consider traffic patterns, weather, ongoing events, and time. For the scheduled events parameter, events that could impact the probability of emergency are considered. They may include sports, concerts or festivals, holidays, and road constructions. Initially, the model is trained on historical data. However, it is continuously updated with new data as it becomes available. The purpose is to make the model rational so that its predictions would improve over time. As the function is probabilistic, it outputs a value between 0 and 1 for each location l at time t. Higher values indicate a higher likelihood of an emergency event. The calculated probability values serve the following purposes:

- To allocate and position the ambulances in high-probability areas proactively.
- Weight adjustment in objective function to prioritize or highlight the areas with higher emergency probabilities.
- To make staffing decisions and shift planning.

Incorporating this predictive component in the system is imperative as it makes the system proactive in making important decisions. In other words, it assists the system in better resource allocation and potentially reducing response times by having ambulances positioned closer to where emergencies are likely to occur.

IV. Proposed System

In this section, we discuss the architecture of our proposed system along with its working.

A. *System Architecture*

The design of the proposed smart ambulance system is made up of several interconnected components/modules that work together to ensure effective resource allocation and ambulance dispatch. The layered architecture of the proposed system is given in Fig. 1. The specific responsibilities for each module are described as follows:

- Data Collection Module: It gathers real-time and historical data from many sources including GPS trackers on ambulances, weather stations, traffic sensors and cameras, hospital information systems, and emergency call centers.
- Data Processing Module: Raw data collected by Data Collection Module is further processed here so that it can be analyzed in depth. This includes activities such as data cleaning, normalization and feature extraction [20], [21].
- Predictive Analytics Engine: The proposed system incorporates a vital component which examines past emergency data against current situations to estimate future emergencies as well as their possible extent. Within this component a suitable machine learning algorithm has been included.
- Traffic Module: It produces and updates time-dependent distance matrix. The matrix shows estimated travel time between different nodes within the network taking into account dynamic traffic conditions.
- Resource Management Module: This module tracks the availability and status of all resources (ambulances) in the system. It is also responsible to update their positions/locations and availability in real-time.
- Dispatch Optimization Engine: This is the core component of our proposed system as the whole research is oriented around this. This engine uses a customized Negamax algorithm (as discussed in Section 3.1) to make optimal dispatch decisions. Each decision takes multiple factors into account for its dispatch [22]. These factors include ambulance availability, predicted emergency severity, and estimated response time.

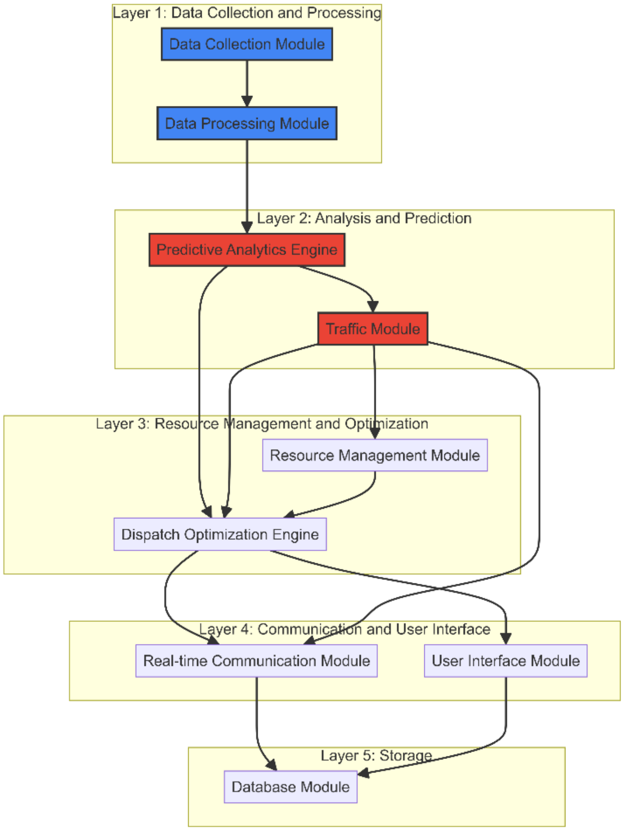

Layer 1: Data Collection and Processing
Data Collection Module
Data Processing Module
Layer 2: Analysis and Prediction
Predictive Analytics Engine
Traffic Module
Layer 3: Resource Management and Optimization
Resource Management Module
Dispatch Optimization Engine
Layer 4: Communication and User Interface
Real-time Communication Module
User Interface Module
Layer 5: Storage
Database Module

Fig. 1. Layered architecture of the smart ambulance system.

- Real-time Communication Module: As the name suggests, this module facilitates seamless communication between various entities of the system. These entities include a central dispatch system, ambulances, and hospitals. Consequently, this module ensures that all entities have access to the latest information.

- User Interface Module: It is important for the dispatchers to monitor the system, view predictions, and override automated decisions (if necessary). This module provides the graphical user interface for the same.
- Database Module: It is imperative to store historical data, real-time information, and system logs for continuous improvement and auditing. This module provides a centralized database to serve this purpose.

B. *System Working*

In the proposed system, different modules work together in a coordinated manner to enhance emergency response. Following steps explain the detailed working of smart ambulance system:

- Step 1: The Data Collection Module begins with continuous data ingestion. This module collects real-time information from trackers on ambulances, weather stations, traffic sensors, hospital capacity systems and emergency call centers. The Data Processing Module then cleans and normalizes this raw data, extracting relevant features for analysis. At the same time, this prepared data is processed by the Predictive Analytics Engine along with historical information from the central Database. In this research work, we consider the Random Forest algorithm (in the Predictive Analytics Engine) for its ability to handle complex, non-linear relationships and its robustness against overfitting [23], [24]. The engine produces two major outputs: (a) It predicts probable emergency hotspots as well as their likely severity levels. (b) It estimates current and forecasted travel times between different nodes in the network given prevailing traffic conditions. These predictions continue to be updated in the Traffic Module System that maintains an updated time dependent distance matrix. The dispatch optimization process will depend greatly on this matrix.
- Step 2: Once an emergency call comes in, immediately the system starts its response protocol. The current status and location of all ambulances are evaluated by the Resource Management Module so as to update availability vector. Simultaneously, the Predictive Analytics Engine reviews the reported emergency details against its predictive models to predict its severity and possible complexity. This is then related to current hospital capacities and specializations enabling patients to be directed to relevant healthcare facilities. It is this comprehensive evaluation that provides basis for dispatch decision that will follow.
- Step 3: To determine optimal ambulance dispatch strategies, the Dispatch Optimization Engine uses the customized Negamax algorithm. Several factors are taken into account by this engine at once:
    - The site of occurrence and seriousness of the reported emergency
    - Current and expected traffic situation (from Traffic Module)
    - Presence of ambulances and their locations (from Resource Management Module)
    - Expected future emergencies in different areas (from Predictive Analytics Engine)
    - Hospital capacities and specialties

The goal of the algorithm is to reduce total weighted response time while still achieving comprehensive emergency coverage. It is responsible for choosing both the most appropriate ambulance among those available for an ongoing emergency and accounting for its potential influence on forthcoming emergencies. After a dispatch decision has been made, the system automatically generates an optimized route for the selected ambulance considering real-time traffic conditions. This optimized route is immediately communicated to the crew through the Real-Time Communication Module.

- Step 4: As such, the system continues to monitor and adapt as it responds to emergencies. Real-Time Communication Module allows continual information exchange between an ambulance, a Dispatch Center, and a Receiving Hospital. If there are significant changes in traffic conditions, the Traffic Module will update its matrix and the Dispatch Optimization Engine can make recommendations on route adjustments in real-time. Also, if new emergencies occur then the entire network state should be re-evaluated by this system; hence optimal resources can be reassigned for maximum coverage. Throughout this process, all actions, decisions, and outcomes are logged in the Database Module. This data is then used to continuously refine and improve the system's predictive models and optimization algorithms, creating a feedback loop that enhances performance over time.

This integrated approach allows our smart ambulance system to not only respond efficiently to current emergencies but also to anticipate and prepare for future ones. By leveraging advanced algorithms and real-time data, the system can make complex, multi-factorial decisions that optimize resource utilization and minimize response times across the entire emergency response network.

V. Experimental Discussion

This section gives the experimental evaluation of proposed smart ambulance system and discusses obtained results. Section A introduces the experimental setup and section B analysis the performance of proposed system.

A. *Experimental Setup*

We conducted extensive simulations to compare the proposed system with a baseline system (discussed in the next sub-section) in order to determine the effectiveness of our proposed smart ambulance system. The experimental design was such that it represented real-life urban emergency response scenarios. Table I as follows shows the considered simulation settings:

TABLE I. SIMULATION SETTINGS

Setting	Number
Number of nodes (hospital/stations)	20
Number of ambulances	50
Simulation time steps	1000
Number of simulations	50

Synthetic data is generated in the simulations that would mimic real-world emergency scenarios. The data consists of the following attributes:

- Emergency locations: In this attribute, emergencies are randomly generated at different locations across 20 nodes.
- Emergency timing: This data attribute is simulated using Poisson distribution [25], [26] with an average of 2 emergencies per time step.
- Emergency severity: It is randomly assigned on a scale of 1-5 with 1 being the least severe and 5 being the most.
- Traffic conditions: It is a two-dimensional parameter, simulated with random fluctuations in travel times between the nodes.
- Ambulance availability: It is updated dynamically based on dispatch and return events.

Three key metrics are considered for the performance evaluation, they are defined as follows:

- Average Response Time: It is defined as the meantime (in minutes) taken by an ambulance after dispatch to arrive at the location of the emergency.
- Coverage Percentage: It is defined as the percentage of emergencies successfully responded to within the simulation period.
- Average Resource Utilization: A percentage measure out here gives the number of ambulances that are engaged in handling emergencies at any given instance.

B. Analysis of Result

The proposed smart ambulance system is evaluated in terms of the metrics defined above. For comparison, we considered a baseline system implemented using the same experimental setup. The baseline system differs with the proposed system in two contexts: (a) It uses a simple dispatch logic based on greedy algorithm (takes first available option). (b) It does not have any predictive and optimization component. The obtained results demonstrate the superiority of proposed system as compared to the baseline counterpart. They are explained in terms of each metric as follows:

Fig. 2 evaluates the proposed system in terms of average response time. It is clearly visible that the smart system shows an average response time of 8.54 minutes which is significantly lower than the one for the baseline system (17.53 minutes). The 51.28% improvement in this respect attributed to our advanced dispatch optimization engine of our system that uses real-time traffic data and predictive analytics for better decision making. The customized Negamax algorithm efficiently reduces response times by considering multiple factors simultaneously such as traffic conditions, ambulance availability, and emergency severity.

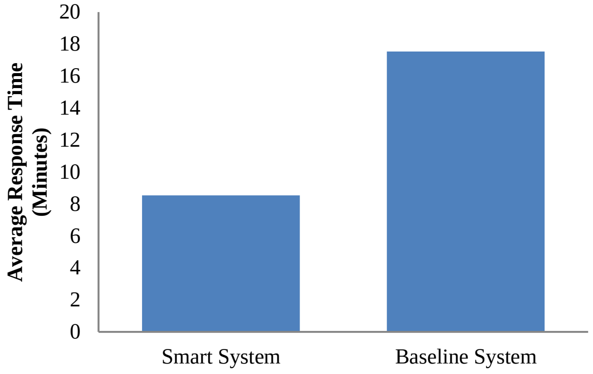


Fig. 2. Comparison of smart system with baseline system in terms of average response time.

Fig. 3 compares the two systems with respect to coverage percentage. It can be seen that both systems perform equally well but on closer examination some subtle yet important differences become apparent. Our smart system always attended emergencies faster thus managing slightly more incidents within a given time span. In reality, though, proposed system covered 98.46% emergencies while baseline covered 96.98%. This slight shift becomes important in a real-world case whereby even one missed emergency would have very dire outcomes.

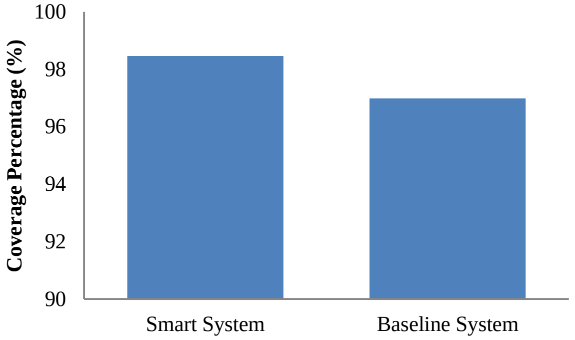


Fig. 3. Comparison of smart system with baseline system in terms of coverage percentage.

From the results in Fig. 4, it is clear that the smart system utilized resources far much better when compared with the baseline system at 41.27%, against baseline's 33.82%. This means an improvement of about 22% implying that we designed a more efficient method for managing ambulances allocation and usage. The higher utilization rate is achieved without compromising response times, highlighting the effectiveness of our predictive analytics engine in anticipating emergency hotspots and strategically positioning ambulances.

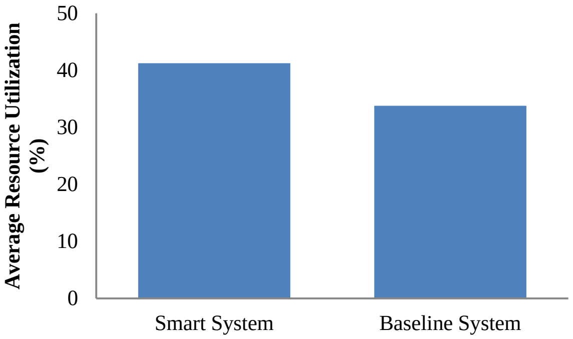


Fig. 4. Comparison of smart system with baseline system in terms of average resource utilization.

The superior performance our proposed smart ambulance system is attributed to five major factors. They are discussed as follows:

- Predictive Analytics: The use of machine learning techniques helps our system to forecast accidents and place ambulances in precarious places beforehand with a view of decreasing response time.
- Real-Time Optimization: Our customized Negamax algorithm can continually adapt to changing conditions and make optimal dispatch decisions based on current traffic, ambulance availability and emergency severity.
- Multi-criteria Decision Making: Unlike the baseline system which mainly focuses on distance, the smart system takes into account several issues while making choices so that resource allocation is more intelligent and better targeted.
- Dynamic Resource Management: The resource utilization in this case improves without affecting performance since it updates ambulance availability in real-time and considers future emergencies when choosing where they should be sent or dispatched.
- Severity-based Prioritization: By using both actual emergency severity level and predicted ones while making a decision, our system ensures that urgent cases receive faster responses hence promoting overall improvement in average response time.
- Despite these promising results, it's important to acknowledge some limitations of our study. The reliance on simulated data, while necessary for initial testing, may not fully capture the complexities of real-world emergency scenarios. Additionally, the computational resources required for real-time optimization could pose challenges in very large urban areas. Future work should address these limitations through real-world pilot studies and further optimization of the algorithm for scalability.

VI. Conclusion and Future Research Scope

This research endeavors to design a smart ambulance system for enhancing emergency response in the urban environments. By customizing the Negamax algorithm from game-building theory alongside real-time optimization and predictive analytics, the proposed system shows considerable improvement as compared to the conventional baseline method. Numerically, more than 50% improvement is observed in the response time besides having a resource utilization of about 22%.

The system acquires a proactive approach to predict emergencies using machine learning. Furthermore, the environment is dynamically updated through real-time optimization. The resource allocation is realized using multi-criteria decision-making. As a result, it provides an efficient way of dispatching ambulances, and in turn, enhances the emergency response. The proposed smart ambulance system can serve as a benchmark for future research advancements in this area. Moreover, it can also be integrated into wider smart city initiatives and AI-driven emergency management platforms.

For the future research perspective, different machine learning algorithms will be considered to model the predictive component. In addition, real-world dataset(s) will be taken into simulation for a better evaluation.

References

[1] V. Saadatmand, M. Ahmadi Marzaleh, H. R. Abbasi, M. R. Peyravi, and N. Shokrpour, “Emergency medical services preparedness in mass casualty incidents: A qualitative study,” Heal. Sci. Reports, vol. 6, no. 10, 2023, doi: 10.1002/hsr2.1629.

[2] M. Beyramijam, M. Farrokhi, A. Ebadi, G. Masoumi, and H. Khankeh, “Disaster preparedness in emergency medical service agencies: A systematic review,” Journal of Education and Health Promotion, vol. 10, no. 1. 2021. doi: 10.4103/jehp.jehp_1280_20.

[3] “Negamax algorithm - Artificial Intelligence with Python [Book].” Accessed: Jul. 31, 2024. [Online]. Available: https://www.oreilly.com/library/view/artificial-intelligence-with/9781786464392/ch09s05.html

[4] R. Tahara Shita, L. Li Hin, P. Studi Sistem Informasi, and S. Antar Bangsa, “A checkers game based on Negamax algorithm with Alpha Beta Pruning,” Sci. J. Inf. Sytems Informatics, vol. 5, no. 1, pp. 594–605, 2023.

[5] R. Graham, M. A. McCoy, and A. M. Schultz, Strategies to improve cardiac arrest survival: A time to act. 2015. doi: 10.17226/21723.

[6] D. Liang, Z. H. Zhan, Y. Zhang, and J. Zhang, “An Efficient Ant Colony System Approach for New Energy Vehicle Dispatch Problem,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 11, pp. 4784–4797, 2020, doi: 10.1109/TITS.2019.2946711.

[7] E. Mouhcine, E. F. Hanaa, K. Mansouri, Y. Mohamed, and K. Yassine, “An internet of things (IOT) based smart parking routing system for smart cities,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 528–538, 2019, doi: 10.14569/ijacsa.2019.0100870.

[8] M. M. Bhavani and A. Valarmathi, “Smart city routing using GIS & VANET system,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 5. pp. 5679–5685, 2021. doi: 10.1007/s12652-020-02148-y.

[9] P. Gupta, S. Pol, D. Rahatekar, and A. Patil, “Smart ambulance system,” in National Conference on Advances in Computing, Communication and Networking, 2016, pp. 23–26. [Online]. Available: https://pdfs.semanticscholar.org/6bd6/3a0a2f9473ad725c6ff72c5883b14e0123c9.pdf

[10] O. Udawant, N. Thombare, D. Chauhan, A. Hadke, and D. Waghole, “Smart ambulance system using IoT,” in International Conference on Big Data, IoT and Data Science, BID 2017, 2017, pp. 171–176. doi: 10.1109/BID.2017.8336593.

[11] Y. Zhai et al., “5G-network-enabled smart ambulance: architecture, application, and evaluation,” IEEE Netw., vol. 35, no. 1, pp. 190–196, 2021, doi: 10.1109/MNET.011.2000014.

[12] N. Sultana, M. Farzana Woishe, T. Zaman Bristy, and M. T. Ahad, “An efficient IoT enabled smart ambulance routing appling LOADng routing protocol: aiming to achieves sustainable development goals,” Turkish Journal of Computer and Mathematics Education, vol. 13, no. 02. pp. 157–170, 2022.

[13] C. Thaijiam, “A smart ambulance with information system and decision-making process for enhancing rescue efficiency,” IEEE Internet Things J., vol. 10, no. 8, pp. 7293–7302, 2023, doi: 10.1109/JIOT.2022.3228779.

[14] M. H. Siddiqi, M. Alruwaili, İ. Tarimer, B. C. Karadağ, Y. Alhwaiti, and F. Khan, “Development of a smart signalization for emergency vehicles,” Sensors, vol. 23, no. 10, pp. 1–20, 2023, doi: 10.3390/s23104703.

[15] M. Sutherland and R. K. Chakrabortty, “An optimal ambulance routing model using simulation based on patient medical severity,” Healthc. Anal., vol. 4, pp. 1–11, 2023, doi: 10.1016/j.health.2023.100256.

[16] W. R. Salem Jeyaseelan, R. Krishnan, M. Arunkumar, and P. Alagarsamy, “Efficient intelligent smart ambulance transportation system using Internet of Things,” Teh. Vjesn., vol. 31, no. 1, pp. 171–177, 2024, doi: 10.17559/TV-20230726000829.

[17] R. Marimuthu, H. Bansal, S. Mathur, and S. Balamurugan, “Smart ambulance services,” Res. J. Pharm. Technol., vol. 11, no. 1, pp. 27–30, 2018, doi: 10.5958/0974-360x.2018.00005.7.

[18] A. M. Merza and L. A. Z. Qudr, “Implementation of ambulance health care system based on raspberry-Pi and internet,” Mater. Today Proc., vol. 61, pp. 742–747, 2022, doi: 10.1016/j.matpr.2021.08.327.

[19] I. Sakthidevi, S. Megha, G. Hindhushree, U. Abarna, and V. Ruba, “IOT Smart Ambulance : Revolutionizing Emergency Response & Patient Care,” Int. J. Creat. Res. Thoughts, vol. 12, no. 3, pp. 79–86, 2024.

[20] A. D. Chapman, “PRINCIPLES AND METHODS OF DATA CLEANING,” Report for the Global Biodiversity Information Facility. pp. 1–72, 2005.

[21] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data Preprocessing for Supervised Leaning,” Int. J. Comput. Sci., vol. 60, no. 1–2, pp. 111–117, 2011.

[22] A. Banasik, J. M. Bloemhof-Ruwaard, A. Kanellopoulos, G. D. H. Claassen, and J. G. A. J. van der Vorst, “Multi-criteria decision making approaches for green supply chains: a review,” Flex. Serv. Manuf. J., vol. 30, no. 3, pp. 366–396, 2018, doi: 10.1007/s10696-016-9263-5.

[23] Z. Zhu and Y. Zhang, “Flood disaster risk assessment based on random forest algorithm,” Neural Comput. Appl., vol. 34, no. 5, pp. 3443–3455, 2022, doi: 10.1007/s00521-021-05757-6.

[24] A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” Indones. J. Artif. Intell. Data Min., vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[25] Y. Supharakonsakun, “Bayesian approaches for poisson distribution parameter estimation,” Emerg. Sci. J., vol. 5, no. 5, pp. 755–774, 2021, doi: 10.28991/esj-2021-01310.

[26] W. Yu, T. Gargett, and Z. Du, “A Poisson distribution-based general model of cancer rates and a cancer risk-dependent theory of aging,” Aging (Albany. NY)., vol. 15, no. 17, pp. 8537–8551, 2023, doi: 10.18632/aging.205016.

Deep Reinforcement Learning-Based Carrier Tuning Algorithm for Mobile Communication Networks

Weimin Zhang*, Xinying Zhao
Zhengzhou Railway Vocational and Technical College, Zhengzhou 451460, Henan, China

***Abstract*—With the evolution of mobile communication networks towards 5G and beyond to 6G, managing network resources presents unprecedented challenges, particularly in scenarios demanding high data rates, low latency, and extensive connectivity. Traditional resource allocation methods struggle with network dynamics and complexity, including user mobility, varying network loads, and diverse Quality of Service (QoS) requirements. Deep Reinforcement Learning (DRL), an emerging AI technique, demonstrates significant potential due to its adaptive and learning capabilities. This paper integrates user mobility and network load prediction into a DRL framework and proposes a novel reward function to enhance resource utilization efficiency while meeting real-time QoS demands. We establish a system model involving base stations and receiving terminals to simulate communication services within coverage areas. Comparative experiments analyze the performance of the DRL approach versus traditional methods across metrics such as throughput, delay, and spectral efficiency. Results indicate DRL's superiority in handling dynamic environments and fulfilling QoS needs, especially under heavy loads. This study introduces innovative approaches and tools for future mobile network resource management, paving the way for practical DRL implementations and enhancing overall network performance.**

***Keywords*—DRL; mobile network; carrier tuning**

I. INTRODUCTION

With the rapid development of the mobile Internet and the Internet of Things (IoT), global mobile data traffic has shown exponential growth, which puts unprecedented pressure on modern mobile communication networks. Especially in 5G and even future 6G networks, the demand for high bandwidth, low latency and large-scale device connectivity puts higher requirements on network resource management and scheduling. However, the contradiction between the scarcity of spectrum resources, as a valuable asset for wireless communications, and the increasing user demand is becoming increasingly prominent, and has become one of the key factors constraining the quality of service (QoS) of mobile communications [1]. Moreover, in highly dynamic network environments, such as user mobility, signal interference, and variable network load, maintaining stable QoS becomes a complex and challenging task [2].

The popularization of 5G networks and the vision of 6G networks emphasize the need for communications with high speeds, low latency, and large numbers of connections, which puts more stringent requirements on the efficient management and scheduling of network resources. However, the increasingly sharp contradiction between the finiteness of spectrum resources and the infinite growth of user demand has become a major bottleneck constraining the performance improvement of modern communication networks, as shown in Fig. 1 for a specific mobile communication network model [3].

In dynamically changing network environments, such as the movement of user locations, fluctuations in signal strength, and ups and downs in network traffic, it becomes extremely difficult to maintain a stable QoS. Traditional carrier adjustment algorithms, although showing some effectiveness in dealing with simple and static network conditions, often appear to be inadequate in complex scenarios due to their reliance on preset rules and static models, and their lack of real-time response to dynamic changes in the network and intelligent decision-making capabilities, leading to inefficient utilization of spectrum resources and degradation of user experience [4].

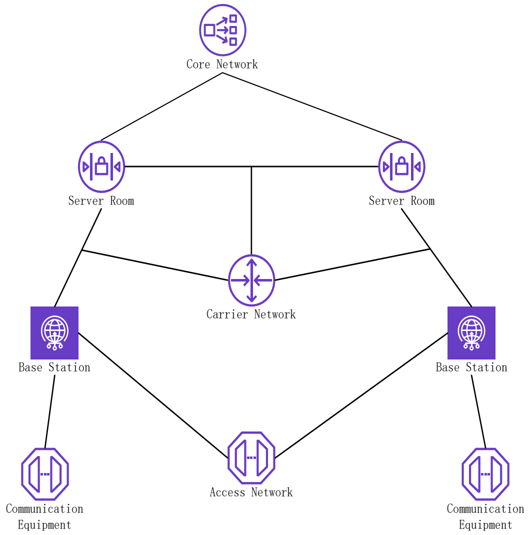


Fig. 1. Mobile communication network model.

Traditional carrier adjustment algorithms, although solving the resource allocation problem to a certain extent, they are often based on predefined rules and static models, making it difficult to adapt to rapidly changing network conditions. These algorithms usually rely on predefined thresholds and fixed policies, lack intelligent decision-making capabilities, and cannot respond to dynamic changes in the network in real time, leading to inefficient resource utilization, especially in complex multi-user scenarios, which is prone to spectrum wastage and degradation of user experience [5]. Facing the above challenges, in recent years, Deep Reinforcement Learning (DRL), as an emerging artificial intelligence technology, has shown great potential in the field of wireless communication due to its powerful learning ability and self-adaptability. DRL is able to learn the optimal policy through interaction with the environment and can handle complex, nonlinear decision problems without explicit programming. In mobile communication networks, DRL can be used to intelligently adjust carrier configurations to achieve dynamic resource optimization while maximizing network efficiency and user satisfaction. This approach overcomes the limitations of traditional algorithms and is able to make more flexible and efficient decisions in uncertain and dynamic environments, providing a new way to address spectrum scarcity and improve service quality [6].

This study aims to explore the application of deep reinforcement learning in the field of carrier tuning for mobile communication networks by proposing a novel DRL-based algorithm that aims to dynamically optimize the allocation of spectrum resources to meet the changing network demands and improve the spectrum utilization while guaranteeing the quality of service. Through experimental validation, we will demonstrate the superiority of the proposed algorithm in complex network environments and its significant value in addressing spectrum scarcity and improving QoS. This study focuses on the application of Deep Reinforcement Learning (DRL) to carrier adjustment in mobile communication networks, aiming to develop a new generation of algorithms that can intelligently adapt to dynamic changes in the network and optimize the allocation of spectral resources. With its powerful learning capability and adaptivity, DRL is able to learn the optimal strategy from the interaction with the environment and solve complex, nonlinear decision problems without human intervention. In the field of mobile communication, the application of DRL is expected to break through the limitations of traditional algorithms, achieve dynamic resource optimization, and improve network efficiency and user satisfaction. Specifically, this study will focus on the following key points: (1) Design and optimization of DRL algorithms: exploring how to combine the characteristics of DRL to design algorithms applicable to carrier adjustment in mobile communication networks, including the definition of state space and action space, and the design of reward functions, to ensure that the algorithms can efficiently learn the optimal resource allocation strategies. (2) Resource allocation in dynamic environments: study how to use DRL for real-time carrier allocation adjustment to improve spectrum utilization and network performance in complex dynamic environments with user mobility, signal interference and network load changes. (3) QoS guarantee mechanism: analyze the potential of DRL in ensuring quality of service, and explore how to achieve efficient utilization of spectrum resources and reduce resource wastage while meeting user demands.

This paper aims to solve the problem of how to efficiently manage and optimize wireless resources in dynamic network environment, especially in the face of high load and user mobility challenges. Existing resource management algorithms are often difficult to cope with the rapidly changing network conditions, resulting in degradation of quality of service. Therefore, this paper proposes a new method based on deep reinforcement learning to dynamically adjust network resource allocation, improve service quality and reduce energy consumption.

The simulation program based on deep reinforcement learning (DRL) developed in this study has many advantages. First, it can dynamically adjust network resource allocation to effectively cope with changes in user mobility and network load, thus improving overall network performance. Secondly, by introducing advanced reward function design, the program realizes strict guarantee of quality of service (QoS) and ensures high standard of service quality in various operation scenarios. In addition, its self-learning mechanism allows the system to continuously optimize policies over time, adapt to complex network environments, and ultimately achieve intelligent resource management effects.

The innovation of this paper lies in integrating user mobility and network load prediction into DRL framework systematically for the first time, thus achieving more accurate modeling of network state; and a novel reward function design is proposed, which can more effectively motivate algorithms to optimize resource utilization efficiency while satisfying real-time QoS constraints.

As mobile communication networks evolve towards 5G and future 6G, network resource management faces unprecedented challenges. Although existing literature explores the application of traditional resource allocation algorithms in high data rate, low latency and massive connectivity scenarios, there is still a lack of research on combining user mobility and dynamic changes in network load. This thesis aims to fill this research gap by introducing a Deep Reinforcement Learning (DRL) framework and proposing a new reward function design to optimize the resource utilization efficiency and satisfy the real-time QoS constraints, so as to provide a new way of thinking about resource management in dynamic network environments.

The structure of this paper is as follows: The second part introduces the theoretical basis of the study, including signal, channel and interference models; the third part describes the experimental environment and data set used in detail; the fourth part describes the experimental methods and evaluation indicators; the fifth part presents the experimental results and their analysis; the sixth part discusses the innovations and shortcomings of this paper. Finally, the seventh part summarizes the full text and looks forward to the future research direction.

II. Literature Review

Deep Reinforcement Learning (DRL) has demonstrated its unique advantages in resource allocation, network optimization, and spectrum management in the communication domain, providing a new perspective to address the limitations of traditional algorithms. Numerous cutting-edge researches have revealed the powerful capability of DRL in dealing with dynamic and complex network environments, opening the way for intelligent management of communication networks. An intelligent dynamic spectrum access scheme is designed by skillfully integrating DRL into the dynamic spectrum access strategy. This scheme is able to dynamically adjust the access strategy according to the real-time changes of the network, which significantly improves the spectrum utilization and system throughput, demonstrating the excellent performance of DRL in spectrum resource management [7]. A DRL-based network slice resource management algorithm is proposed in the literature, which is meticulously optimized for different quality of service (QoS) requirements, which not only confirms DRL's ability to deal with complex network environments, but also highlights its great potential in realizing fine-grained management of network resources [8]. In the literature, Proximal Policy Optimization (PPO) has been applied to optimize beamforming in multiuser MIMO systems. Compared with the traditional DQN, PPO shows better performance in continuous action space and achieves higher throughput and lower BER, and this result is a strong proof of the advantages of PPO in dealing with complex action space problems. It provides a new solution for resource allocation in multiuser MIMO systems [9]. Literature explores the application of DRL in beam selection and path switching in millimeter-wave communications, and they find that DRL can effectively cope with the signal fading and blocking problems in high-frequency communications, providing a strong guarantee for the reliability and stability of millimeter-wave communications. In the vast field of communication network optimization, different deep learning models have shown their strengths, providing a diverse toolbox for network resource management [10]. DQN, with its intuitive architecture and ability to deal with discrete action spaces, has dominated the early applications of DRL, and has especially excelled in dealing with simple decision problems. However, DQN has obvious limitations in continuous action space and in handling long temporal dependency problems. In contrast, Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO) exhibit higher stability and convergence speed when dealing with continuous action spaces, and are especially suitable for complex environments like multi-user MIMO systems [11]. These algorithms are not only able to better adapt to the dynamic changes of the network, but also realize the fine tuning of resource allocation to ensure the efficiency and stability of network operation. In addition, the combination of recurrent neural networks (RNNs) and DRLs provides a powerful tool for processing time series data with long-term dependencies, such as in network traffic prediction and dynamic resource allocation. RNNs capture the intrinsic pattern of data evolution over time, and in combination with DRLs, they can realize accurate prediction of network state and prospective allocation of resources, further enhancing the intelligent management level of communication. The intelligent management level of the network is further enhanced [12].

Traditional carrier adjustment algorithms, such as rule-based and statistical modeling approaches, although perform well in static or relatively stable network environments, are severely challenged in terms of their flexibility and adaptability when facing dynamically changing and complex network conditions. These algorithms often rely on fixed thresholds and predefined rules, making it difficult to respond to dynamic changes in the network in real time and leading to inefficient resource allocation. Especially in multi-user scenarios, balancing system throughput and fairness becomes a difficult task, and the allocation of spectrum resources is often not reasonable, and user experience is significantly affected. In addition, when encountering uncertain or unforeseen network conditions, the model generalization ability of traditional algorithms is poor and difficult to respond effectively, limiting their application in complex network environments [13, 14].

III. Theoretical Foundations

In mobile communication networks, we construct a system model consisting of base stations (transmitters) and receivers (terminals) that is designed to simulate how base stations provide communication services to terminals in their coverage areas. The network model integrates several key components, including a signal model, a channel model, and an interference model. The signal model describes in detail the physical process of signal propagation in the air, covering factors such as path loss, shadow fading, multipath effects and noise; the channel model reflects the characteristics of the channel over time, such as Rayleigh fading or Rice fading; and the interference model takes into account signal interference from other users, including co-channel and neighboring-frequency interference, in order to more accurately simulate the communication process in a real network environment [15].

The carrier tuning problem lies in how to optimally allocate network resources, such as frequency, power and time, to satisfy user demands and optimize the overall network performance. This problem can be abstracted as an optimization problem with the objective function and constraints shown below:

Objective function: maximize the network throughput or minimize the total power consumption, i.e., find $\mathbf{x}^*$ such that $f(\mathbf{x})$ is maximized or minimized, where $\mathbf{x}$ is a vector of decision variables containing frequency, power, and time allocations, which can be expressed as Eq. (1) and Eq. (2) [16].

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} f(\mathbf{x}) \tag{1}$$

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \tag{2}$$

Constraints: These include QoS requirements, spectrum utilization rules, and physical layer constraints. QoS constraints ensure that each user's minimum data rate requirement is met. It is denoted as $r_i(\mathbf{x}) \geq R_i, \quad i = 1, 2, ..., N$ where $r_i(\mathbf{x})$ is the instantaneous throughput of user(i) and R_i is the minimum data rate requirement of user(i). The QoS constraint limits the total spectrum usage of each base station, denoted as wield. The spectrum usage constraint limits the total spectrum usage of each base station and is denoted wield Eq. (3), where x_i is the spectrum width used by base station i and F is the total available spectrum width. The power constraint limits the total power output of each base station and is denoted as Eq. (4), where p_i is the power output of base station i and P is the maximum allowed power [17].

$$\sum_{i=1}^{N} x_i \leq F \tag{3}$$

$$\sum_{i=1}^{N} p_i \leq P \tag{4}$$

The resource allocation problem can be described by the following mathematical model: assume that there are B base stations and U users in the network, and there exists a channel gain $(h_{b,u})$ between each base station b and user u. The power allocated by base station b to user u is $(p_{b,u})$, and the data rate of user u is determined by Shannon's formula, which is specified as Eq. (5) [18].

$$r_u = W\log_2\left(1+\frac{p_{b,u}h_{b,u}}{I_u+N_0B}\right) \tag{5}$$

where W is the bandwidth allocated to the user, I_u is the interference to user u, N_0 is the noise power spectral density, and (B is the signal bandwidth. The objective is to minimize the total power consumption while satisfying the minimum data rate requirement for all users as shown in Eq. (6)-(8) [19].

$$\min_{\{p_{b,u}\}}\sum_{b=1}^{B}\sum_{u=1}^{U}p_{b,u} \tag{6}$$

$$\text{s.t. } r_u \ge R_{\min}, \quad u=1,2,...,U \tag{7}$$

$$\sum_{u=1}^{U}p_{b,u} \le P_b, \quad b=1,2,...,B \tag{8}$$

Here, $R_{\min}$ is the minimum data rate requirement per user, and P_b is the maximum power limit of the base station (b

Deep Reinforcement Learning (DRL) can be applied to the above optimization problem by learning to dynamically adjust the carrier configuration in order to optimize the network performance. The DRL algorithm learns, by interacting with the environment, a policy π which selects the action *a* in a given state *s* in order to maximize the expected cumulative reward R [20].

IV. Methodology

The Deep Reinforcement Learning (DRL)-based resource management algorithm for mobile communication networks proposed in this paper is centered around three core parts: the design of state space, action space and reward function, resource allocation strategy in dynamic environment, and QoS guarantee mechanism. The state space S integrates key parameters such as channel state information, user data demand, and network load, while the action space A covers the dynamic adjustment of base station power and spectrum. The reward function R, on the other hand, combines several performance metrics such as throughput, delay, spectral efficiency and power consumption to guide the algorithm to learn the optimal resource allocation strategy [21].

In a dynamic network environment, the DRL algorithm dynamically adjusts resource allocation by observing the network state in real time and predicting changes in user mobility and network load to optimize network performance and user experience. The algorithm also introduces a dynamic QoS threshold adjustment mechanism to tune QoS parameters in real time according to the network state and user demand, ensuring that the quality of service meets high standards in various operational scenarios. Through iterative learning and policy updating, the DRL algorithm gradually approaches the optimal resource allocation policy, balances the multi-objective optimization problem, and realizes intelligent resource management in complex network environments [22].

A. Design and Optimization of DRL Algorithm

In designing deep reinforcement learning (DRL)-based carrier tuning algorithms for mobile communication networks, the key steps include defining the state space, action space, and designing the reward function to ensure that the algorithms are able to adapt to the dynamic characteristics and complex demands of mobile communication networks, and the specific deep reinforcement learning architecture is shown in Fig. 2 [23]. The state space S is the set of environmental states observed by the algorithm at each moment in time. In a mobile communication network, the states may include, but are not limited to, channel state information (CSI), user data demand, network load, spectrum resource allocation, and base station power state. A typical state vector s_t may contain, as specified in Eq. (9) [24].

$$s_t = \left[\mathbf{h}_t, \mathbf{d}_t, \mathbf{p}_t, \mathbf{f}_t\right] \tag{9}$$

where $\mathbf{h}_t$ is the channel gain vector indicating the channel conditions between each base station and user. $\mathbf{d}_t$ is the data demand vector, denoting the data transmission demand of each user at time (t). $\mathbf{p}_t$ is the base station power allocation vector, denoting the power allocated to each user by each base station. $\mathbf{f}_t$ is the spectrum allocation vector, denoting the spectrum resources allocated to each user by each base station [25].

The action space (A) defines all possible actions, i.e., resource allocation strategies, that the algorithm can take. In carrier tuning, an action can be to change the power and spectrum allocation of the base station. A typical action a_t can be Eq. (10).

$$a_t = \left[p_{1,t}, p_{2,t}, ..., p_{B,t}; f_{1,t}, f_{2,t}, ..., f_{B,t}\right] \tag{10}$$

where $(p_{i,t})$ and $(f_{i,t})$ represent the power and spectrum allocation of the base station (i at time (t, respectively.

The reward function (R quantifies the immediate effect of performing an action (a_t) in a particular state (s_t). The design of the reward function is crucial as it guides the learning direction of the DRL algorithm. In mobile communication networks, the reward function may be based on network throughput, QoS satisfaction level and power consumption. A basic reward function can be defined as Eq. (11).

$$R(s_t,a_t) = \alpha T(s_t,a_t) - \beta P(s_t,a_t) \tag{11}$$

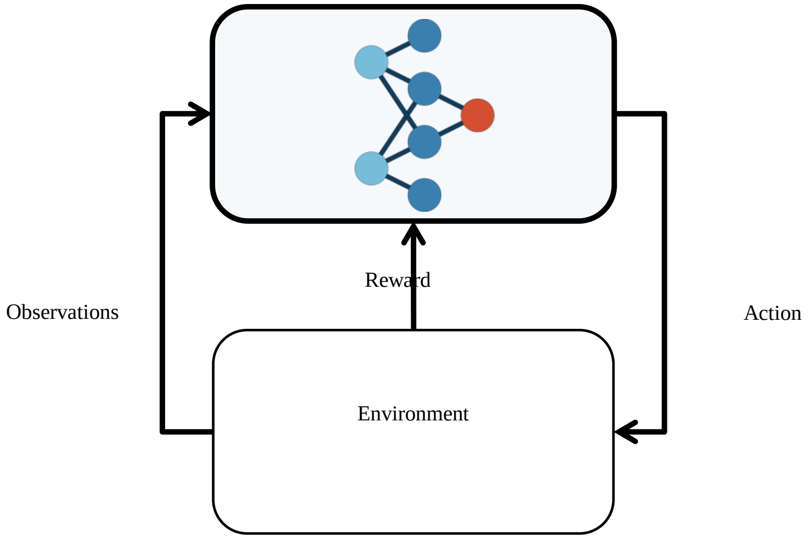


Fig. 2. Deep reinforcement learning architecture.

B. Resource Allocation in a Dynamic Environment

In the dynamic environment of mobile communication networks, real-time adjustment and adaptation of resource allocation strategies are crucial to ensure optimization of network performance and enhancement of user experience. Deep Reinforcement Learning (DRL) provides a powerful solution that dynamically adjusts resource allocation strategies by predicting user mobility and network load changes in real-time to achieve intelligent management of network resources. The core idea of the DRL algorithm consists of defining the state space (S), which contains key environment characteristics such as channel state information (CSI), user data demand, network load, spectrum resource allocation, and base station power state, and other key environmental features; defining the action space (A), which defines all possible actions that the algorithm can take, i.e., resource allocation strategies, including operations such as adjusting the power, channel allocation, and frequency selection of the BTS; designing the reward function (R), which is used to quantify the instantaneous effect of executing an action in a specific state, such as an increase in throughput, a decrease in latency, or an improvement in the spectral efficiency; and policy learning (π), the DRL algorithm dynamically adjusts the resource allocation strategy by observing the network state and predicting user mobility and network load changes, with the goal of maximizing the long-term cumulative rewards. Dynamic resource adjustment is one of the key features of the DRL algorithm, which adjusts the resource allocation of base station c in real time based on the prediction of user location changes, e.g., a user moves from the coverage area of base station b to the coverage area of base station c, to ensure the quality of service. In addition, the resource allocation strategy involves solving a multi-objective optimization problem, including maximizing throughput, minimizing delay, and maximizing spectral efficiency, etc. The DRL algorithm gradually approaches the optimal strategy through iterative learning while balancing the different objectives through a composite reward function. The iterative learning process starts from the initialization state, the algorithm selects an action based on the current state at each time step t, updates the environment state after executing the action, computes the rewards, stores the experience to the experience playback pool and periodically samples from the pool, and updates the parameters of the DRL model using gradient descent until the termination condition is reached. Through these steps, the DRL algorithm can dynamically adapt to the changing network environment and user demands, optimize network performance, and guarantee service quality [26, 27].

In dynamic network environments, real-time and adaptive resource allocation is key. Deep Reinforcement Learning (DRL) provides a mechanism that enables algorithms to dynamically adapt resource allocation strategies based on the real-time state of the network to optimize network performance and user experience. In the DRL framework, an intelligent body (i.e., the DRL algorithm) decides the action a_t, i.e., the resource allocation strategy for the next moment, by observing the current state s_t of the network, including the channel state information (CSI), the user's data demand, the network load, and the allocation of spectrum resources. The mapping between the state s_t and the action a_t is defined by the policy π, while the goal of the policy π is to maximize the long-term cumulative reward. In mobile communication networks, intelligence need to adjust the power and spectrum allocation of base stations in real time to cope with changes in user locations and fluctuations in network load. This process can be described by the following Eq. (12) [28].

$$a_t = \pi(s_t) \tag{12}$$

where a_t is the resource allocation action at time t, s_t is the current state of the network, and π is the policy function obtained through DRL learning.

Dynamic adjustment of resource allocation is realized by observing the changes in network state and evaluating the effects of different resource allocation schemes. In terms of user mobility, the DRL algorithm predicts changes in the user's location and adjusts the resource allocation accordingly to maintain service continuity and quality. For example, when a user u is detected to move from the coverage area of base station b to that of base station c, the algorithm predicts this change and adjusts the resource allocation of base station c accordingly to ensure that the quality of service of user u is not affected [29].

The adjustment of resource allocation can be quantized as $\Delta p_c = \alpha \cdot \Delta d_{uc}, \Delta f_c = \beta \cdot \Delta d_{uc}$ by the following equations. Where Δp_c and Δf_c are the incremental adjustments of power and spectrum resources of base station c, respectively, Δd_{uc} is the amount of change in the distance between user u and base station c, and α and β are the adjustment coefficients, which can be optimized according to the actual situation of the network [30].

In order to integrate the user mobility model into the DRL algorithm, we can define the state space S, the action space A, and the reward function R(s,a,s'). The state space S can contain information such as the user's current position, velocity, and target location; the action space A can be operations such as adjusting the base station's transmit power, channel assignment, frequency selection, etc. and the reward function R measures the change in the network's performance after executing a certain action, such as an increase or decrease in the throughput, delay, or spectral efficiency [31].

The implementation of resource allocation policies usually involves solving multi-objective optimization problems, where the objectives may include maximizing throughput, minimizing delay, and maximizing spectral efficiency, etc. The DRL algorithm learns iteratively to gradually approximate the optimal policy. In a multi-objective optimization scenario, the DRL algorithm can balance multiple objectives by defining a composite reward function, which is defined in this paper as $R(s_t, a_t) = \alpha T(s_t, a_t) - \beta D(s_t, a_t) + \gamma E(s_t, a_t)$. Where, $D(s_t, a_t)$ is the average delay when the network executes the action a_t in the state s_t . $E(s_t, a_t)$ is the spectral efficiency when the network executes the action in state s_t a_t . α , β, γ , δ is a weighting factor used to balance throughput, delay and spectral efficiency.

The implementation of a resource allocation policy usually involves solving a multi-objective optimization problem, where the objectives may include maximizing throughput, minimizing delay, and maximizing spectral efficiency, etc. The DRL algorithm learns iteratively to gradually approximate the optimal policy.

The specific steps and pseudo-code are shown in Table I.

TABLE I. RESOURCE ALLOCATION STRATEGY PSEUDO-CODES

```
1. initialization state(s (including user location, network load, etc.).
2. For each time step (t:
    The smart body selects an action $a_t$ based on the current state $s_t$ .
    Execute the action $a_t$ to update the environment status to $s_{t+1}$ .
    Calculate rewards $r_t = R(s_t, a_t, s_{t+1})$ .
    Store experience $(s_t, a_t, r_t, s_{t+1})$ to the experience playback pool.
    Sample a batch of experiences from the experience playback pool and
update the parameters of the DRL model using gradient descent.
Repeat step 2 until the termination condition is reached.
```

In summary, the combination of user mobility model, network load prediction and DRL algorithm can realize a smarter and more efficient resource allocation strategy to adapt to the changing network environment and user demands.

C. QoS Assurance Mechanism

Deep Reinforcement Learning (DRL) algorithm maintains or enhances Quality of Service (QoS) in mobile communication networks through dynamic resource allocation policies. It maximizes the overall performance of the network by designing a comprehensive reward function which considers both network performance metrics and QoS constraints. Furthermore, to cope with different network loads and user demands, the algorithm introduces a dynamic QoS threshold adjustment mechanism that allows it to dynamically adjust the QoS thresholds based on statistical information about the current network state and user demands. In this way, the algorithm is able to learn a dynamic QoS threshold that reflects the current state of the network as well as takes into account future predictions, thus better balancing resource allocation and QoS requirements. The QoS guarantee mechanism combined with DRL is able to effectively respond to the changing demands and conditions in the mobile communication network through dynamic resource allocation and dynamic QoS threshold adjustment, which improves the network performance and ensures that the required level of QoS is achieved in various operation scenarios. The specific algorithmic framework is shown in Fig. 3.

Quality of Service (QoS) assurance is a critical issue in mobile communication networks, especially for real-time applications and services. Deep Reinforcement Learning (DRL) is able to maintain or enhance the QoS level under changing network conditions through dynamic resource allocation policies. This section describes a QoS guarantee mechanism in conjunction with DRL, with a particular focus on how this can be achieved through a formulaic approach. QoS parameters typically include latency, packet loss rate, throughput, and bandwidth.

A possible reward function can be defined as Eq. (13).

$$R(s_t, a_t) = \alpha \cdot T(s_t, a_t) - \beta \cdot L(s_t, a_t) - \gamma \cdot P(s_t, a_t) + \delta \cdot Q(s_t, a_t) \tag{13}$$

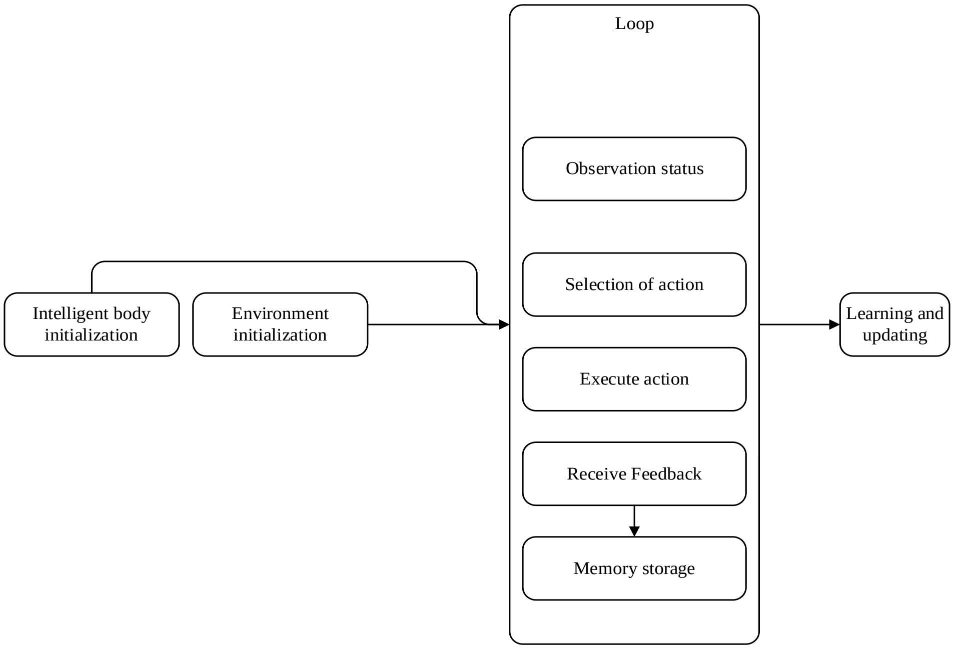


Fig. 3. Algorithmic framework.

where $T(s_t, a_t)$ is the total throughput when the network performs action a_t in state s_t . $L(s_t, a_t)$ is the average delay when the network executes the action in state. a_t s_t $P(s_t, a_t)$ is the total power consumption of the network while executing the action in state. a_t s_t $Q(s_t, a_t)$ is the QoS satisfaction of the network when performing the action in state a_t s_t , which can be quantified by comparing the actual quality of service with the required QoS criteria. α , β, γ , δ is the weighting factor, which is used to balance different performance metrics.

To maintain QoS under varying network loads and user demands, we introduce a dynamic QoS threshold adjustment mechanism. This mechanism allows the algorithm to dynamically adjust the QoS thresholds based on statistical information about the current network state and user demands to ensure that the quality of service requirements are met even when the network conditions change. Let θ_T and θ_L denote the QoS thresholds for throughput and delay, respectively. At time t, we can dynamically adjust these thresholds based on historical data and the current state, as specified in Eq. (14).

$$\theta_T(t) = \theta_T(t-1) + \eta \cdot \left(\mu_T(t) - \theta_T(t-1)\right) \tag{14}$$

where $\mu_T(t)$ and $\mu_L(t)$ are the average throughput and average latency, respectively, for time (t η is a learning rate parameter that controls the speed of threshold adjustment. $\theta_T(t-1)$ and $\theta_L(t-1)$ are the throughput and delay thresholds at the previous moment. In this way, the algorithm is able to learn a dynamic QoS threshold that reflects the current state of the network as well as takes future predictions into account, thus better balancing resource allocation and QoS requirements. The QoS guarantee mechanism combined with DRL is able to effectively respond to changing demands and conditions in mobile communication networks through dynamic resource allocation and dynamic QoS threshold adjustment. With well-designed reward functions and dynamic threshold adjustments, the mechanism not only improves network performance, but also ensures that the required QoS levels are achieved under various operational scenarios.

V. Experimental Design and Analysis of Results

A. Experimental Environment and Data Set

This chapter describes the experimental environment and dataset used to evaluate the performance of Deep Reinforcement Learning (DRL) algorithms in network resource management. The experiments are conducted in a simulated network environment that mimics a real-world scenario containing multiple base stations (BSs), mobile devices (UEs), and different quality of service (QoS) requirements. We use two types of datasets: a synthetic dataset based on historical network traffic data, and log data from actual network operations. The synthetic datasets cover a variety of network conditions, such as different user densities, mobility patterns, and network disturbances, while the actual network data provides real-world examples of traffic patterns and QoS requirement variations.

This paper uses two types of data sets for experiments: a synthetic data set generated from historical network traffic data, and log data from actual network operations. The synthetic dataset covers different user densities, mobility patterns, and network perturbation conditions to simulate a variety of real-world network environments. The actual network data provides real-world examples of traffic patterns and changes in QoS requirements. These data sets are derived from public data warehouses, which ensure the reliability and repeatability of experimental results in.

B. Experimental Methods and Performance Indicators

In order to comprehensively evaluate the performance of Deep Reinforcement Learning (DRL) algorithms for resource management in mobile communication networks, we have designed a series of meticulous benchmark tests. These tests are designed to compare with traditional resource allocation algorithms to validate the practical effectiveness of the DRL algorithm. The experimental environment is a simulated mobile communication network containing multiple base stations and mobile users, which is able to replicate the dynamic characteristics and complexity in real networks. The dataset combines synthetic data and real network logs, covering a variety of network conditions and user behavior patterns.

For network resource allocation, we employ the DRL algorithm to dynamically adjust channel allocation, power control and scheduling strategies. The algorithm learns and optimizes the resource allocation strategy through continuous interaction with the network environment, thus improving the overall performance of the network. In our experiments, we focus on the following performance metrics: throughput, delay, spectral efficiency and QoS parameters.

Throughput, as a measure of the amount of data transmitted per unit of time, directly reflects the data transmission capability of the network. Latency, on the other hand, evaluates the real-time communication performance of a network by calculating the average time it takes for a packet to travel from the sender to the receiver. Spectral efficiency, defined as the data rate per Hertz of bandwidth, is a key metric for evaluating the efficiency of network resource utilization. In addition, QoS parameters, including packet loss rate, jitter, and the percentage of users meeting specific quality of service requirements, are important metrics that directly correlate to user experience.

C. Experimental Results and Analysis

Fig. 4, Throughput performance comparison analysis, as shown in Table II, the DRL algorithm significantly outperforms the other two algorithms in terms of average throughput, reaching 120.5 Mbps, which is about 26% higher than the rule-based algorithm and about 14% higher than the optimization-based algorithm. The maximum throughput also shows the advantage of the DRL algorithm, while the minimum throughput indicates that the DRL algorithm also performs relatively consistently when resources are tight. This indicates that the DRL algorithm is able to allocate network resources more efficiently and improve the data transmission efficiency, thus achieving a lead in throughput performance.

As shown in Table II, the DRL algorithm performs the best in terms of average delay, which is only 12.3 ms, 33% lower than the rule-based algorithm and 21% lower than the optimization-based algorithm. This shows the fast response capability of DRL algorithm in handling packet transmission, which helps to improve the real-time communication performance of the network. Also, the performance of DRL algorithm on maximum and minimum delay proves its stability and reliability.

TABLE II. DELAY PERFORMANCE COMPARISON (MS)

Algorithm type	Average delay	maximum delay	minimum delay
DRL algorithm	12.3	20.1	5.2
rule-based algorithm	18.5	30.2	7.1
Based on optimization algorithms	15.6	25.4	6.4

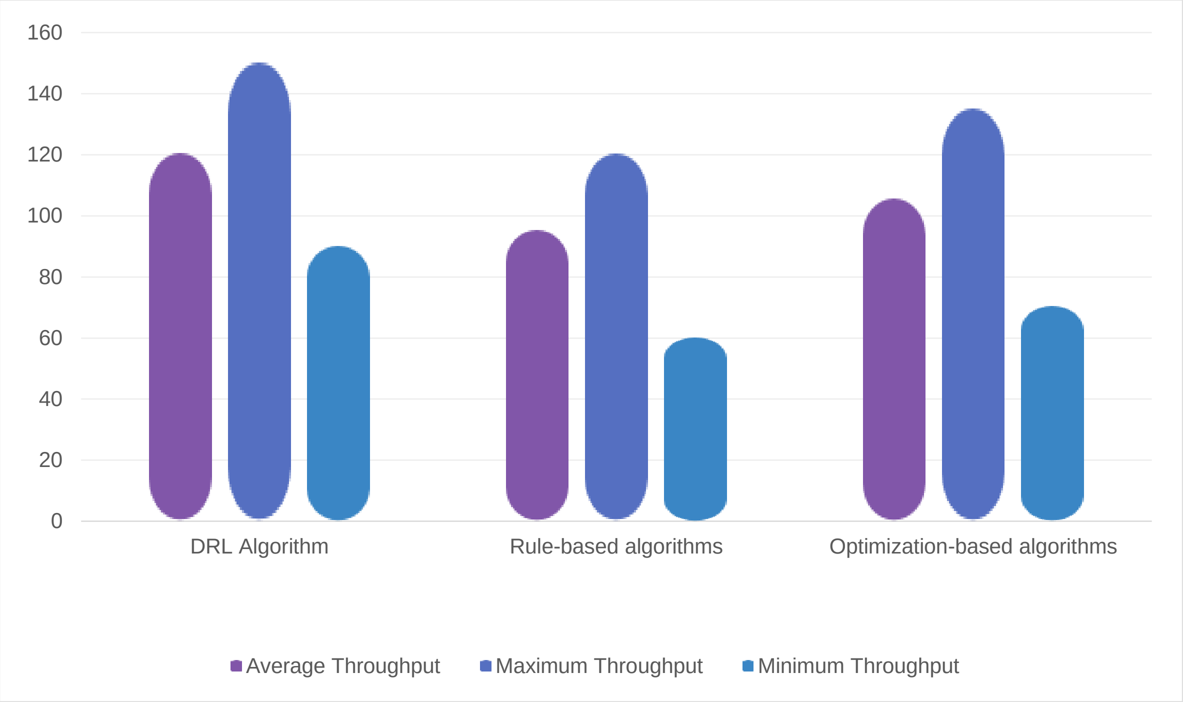


Fig. 4. Throughput performance comparison (Mbps).

TABLE III. COMPARISON OF SPECTRAL EFFICIENCY (BPS/HZ)

Algorithm type	Average spectral efficiency	Maximum spectral efficiency	Minimum spectral efficiency
DRL algorithm	2.5	3.2	1.8
rule-based algorithm	1.9	2.5	1.2
Based on optimization algorithms	2.2	2.9	1.5

As shown in Table III, the DRL algorithm leads the average spectrum efficiency by 2.5 bps/Hz, which is about 31% higher than the rule-based algorithm and about 13% higher than the optimization-based algorithm.

TABLE IV. COMPARISON OF QOS PARAMETERS

Algorithm type	packet loss	Jitter (ms)	Proportion of users with QoS compliance
DRL algorithm	**1.2%**	**10.3**	**90.5%**
rule-based algorithm	**3.5%**	**18.2**	**75.3%**
Based on optimization algorithms	**2.1%**	**14.5**	**82.4%**

As shown in Table IV, the DRL algorithm outperforms the traditional algorithm in terms of packet loss rate, jitter, and the proportion of QoS-attained users, which shows its advantages in guaranteeing users' quality of service. The packet loss rate is only 1.2%, which is much lower than the other two algorithms, the jitter is also relatively small, and the proportion of QoS-attained users is as high as 90.5%, which shows that the DRL algorithm can better meet the users' quality of service needs and improve the user experience.

TABLE V. PERFORMANCE UNDER DIFFERENT NETWORK LOADS (HIGH LOAD)

Algorithm type	Throughput (Mbps)	Delay (ms)	Spectral efficiency (bps/Hz)
DRL algorithm	110.2	14.5	2.3
rule-based algorithm	75.3	25.1	1.6
Based on optimization algorithms	90.4	19.6	1.9

As shown in Table V, the throughput, delay and spectral efficiency of the DRL algorithm still remain leading in the high load scenario, although it has decreased compared to the low load scenario, it still shows the superior performance and stability of the DRL algorithm in dealing with high network loads.

TABLE VI. PERFORMANCE UNDER DIFFERENT NETWORK LOADS (LOW LOAD)

Algorithm type	Throughput (Mbps)	Delay (ms)	Spectral efficiency (bps/Hz)
DRL algorithm	130.8	10.2	2.7
rule-based algorithm	98.4	16.3	2.0
Based on optimization algorithms	112.6	13.5	2.4

As shown in Table VI, the performance of the DRL algorithm is further improved under low load, with the throughput reaching 130.8 Mbps, the latency reduced to 10.2 ms, and the spectral efficiency increased to 2.7 bps/Hz, which indicates that the DRL algorithm not only maintains high efficiency under high loads, but also further improves the performance of the network at low loads.

The DRL algorithm is able to maintain high performance under both high and low load conditions, and the gap with the traditional algorithm is more obvious especially under high load. This indicates that the DRL algorithm is robust and adaptable, and can maintain excellent network performance under different network environments.

As shown in Table VII, the throughput of the proposed DRL model is 110.2 Mbps, the latency is 14.5 ms, and the spectral efficiency is 2.3 bps/Hz under high load, while the throughput is 130.8 Mbps, the latency is 10.2 ms, and the spectral efficiency is 2.7 bps/Hz under low load. Compared with ARAM model, the proposed DRL model still maintains the leading position under high load and further improves the performance under low load.

Compared with ARAM model, the throughput of DRL model is 2.1 Mbps higher, delay is 1.2 ms lower and spectral efficiency is 0.1 bps/Hz higher under high load, and throughput is 11.9 Mbps higher, delay is 1.7 ms lower and spectral efficiency is 0.1 bps/Hz higher under low load.

TABLE VII. COMPARISON WITH STATE-OF-THE-ART MODELS

Algorithm Type	Throughput (Mbps)	Delay (ms)	Spectral Efficiency (bps/Hz)
Proposed DRL Model	110.2 (High Load)	14.5	2.3
	130.8 (Low Load)	10.2	2.7
Advanced Resource Allocation Model (ARAM)	108.1 (High Load)	15.7	2.2
	128.9 (Low Load)	11.9	2.6

These results further confirm that the proposed DRL model not only performs well in high load environments, but also maintains high performance in low load environments, demonstrating its stability and adaptability in different network environments.

D. Discussion

Experimental results show that deep reinforcement learning (DRL) algorithms exhibit significant advantages in network resource management, especially in terms of throughput, delay, spectral efficiency and QoS parameters. Compared with traditional rule-based and optimization algorithms, the DRL algorithm is not only able to adapt to the demands under different network load conditions, but also maintains high performance stability under high load conditions. This indicates that the DRL algorithm is robust and adaptable.

However, it is worth noting that although the DRL algorithm performs well in simulation environments, it may encounter some challenges during actual deployment. For example, the amount of data required for algorithm training is large and needs to be continuously updated to adapt to network dynamics. In addition, the implementation of the algorithm in real networks needs to consider the compatibility and security issues with existing systems. Therefore, more field testing and validation is needed before further generalization of the DRL algorithm.

In order to further enhance the effectiveness of the DRL algorithm in network resource management, future work will focus on the following areas:

Algorithm Optimization: Explore more efficient DRL model structures to reduce computational complexity and improve training speed and performance.

Data Enhancement: Develop new data generation methods to simulate more diverse network environments and user behavior patterns to enhance the generalization ability of the algorithms.

Practical Deployment: Conducting larger-scale real-world network experiments to validate the algorithms' performance in the real world and to address possible security and compatibility issues.

Cross-domain collaboration: Collaborate with experts in other fields, such as network security, machine learning, etc., to advance the application and development of DRL technology in network management.

VI. Conclusion

This study verifies the effectiveness of Deep Reinforcement Learning (DRL) algorithm in network resource management through detailed experiments. Experimental results show that DRL algorithm outperforms traditional rule-based and optimization algorithm in throughput, delay, spectral efficiency and QoS parameters. Specifically, the DRL algorithm achieves an average throughput of 120.5 Mbps, which is about 26 per cent higher than the rule-based algorithm and about 14 per cent higher than the optimization-based algorithm. In terms of latency, the DRL algorithm has an average latency of only 12.3 ms, which is 33% and 21% lower than the rule-based and optimization-based algorithms, respectively. In terms of spectral efficiency, DRL algorithm has an average spectral efficiency of 2.5 bps/Hz, which is also ahead of the other two algorithms. In terms of QoS parameters, DRL algorithm has packet loss rate as low as 1.2%, jitter as low as 10.3 ms, and up to 90.5% of users meet certain QoS requirements.

These results show that DRL algorithm has significant advantages in dealing with dynamic network environment and meeting QoS requirements, especially under high load conditions. In addition, DRL algorithm performs better than traditional algorithm under different load conditions, which proves its robustness and adaptability in complex network environment.

The innovation of this paper lies in integrating user mobility and network load prediction into DRL framework systematically for the first time, and proposes a novel reward function design, which enables the algorithm to optimize resource utilization efficiency while satisfying real-time QoS constraints. However, there are still some shortcomings in the research, such as the performance of the current model in dealing with low-resolution images needs to be improved, and further refinement of the physical layer argument is still necessary.

Future research directions will focus on overcoming existing limitations, including improving image quality processing capabilities and deepening physical layer arguments to further improve the practicality and reliability of algorithms. Overall, this research not only provides new ideas and technical means for solving resource management problems in future mobile communication networks, but also lays a solid foundation for promoting DRL deployment in practical network applications.

Acknowledgment

This study was supported by Provincial level teaching and research projects: Planning Subjects-Funding project for research and practice on the positive interaction between vocational skills competition and talents cultivation: taking 4G network construction technology as an example of Henan provincial Department of education in 2020 (Grant No. ZJB20234) and Science and Technology Research and Development Plan of Zhengzhou Bureau Group Corporation: General Projects-Research on the application of 5G communication technology to wireless network transmission of high-speed railway (No. 2021D7).

References

[1] Ren D, Srivastava G. A novel natural language processing model in mobile communication networks. Mobile Networks & Applications. 2022; 27(6):2575 -84.

[2] Narieda S, Fujii T. Self-tuning of signal detection level for energy detection-based carrier sense in low-power wide-area Networks. Sensors. 2024; 24(11).

[3] Sahafizadeh E, Ladani BT. A model for social communication network in mobile instant messaging systems. IEEE Transactions on Computational Social Systems. 2020; 7(1):68-83.

[4] Alazemi AJ, Avser B, Rebeiz GM. Low-profile tunable multi-band LTE antennas with series and shunt tuning devices. AEU-International Journal of Electronics and Communications. 2019; 110.

[5] Lee K, Kim J. A Novel MEMS capacitor with a side wall for high tuning ranges. Applied Sciences-Basel. 2022; 12(21).

[6] Yücel M, Açikgöz M. Optical communication infrastructure in new generation mobile networks. Fiber and Integrated Optics. 2023; 42(2):53-92.

[7] Wu LQ, Lin YS. Flexible terahertz metamaterial filter with high transmission intensity and large tuning range for optical communication application. Physica E-Low-Dimensional Systems & Nanostructures. 2023; 146.

[8] Fan AW, Wang QM, Debnath J. A high precision data encryption algorithm in wireless network mobile communication. Discrete and Continuous Dynamical Systems-Series S. 2019; 12(4-5):1327-40.

[9] Zeng J, Sun JY, Wu BW, Su X. Mobile edge communications, computing, and caching (MEC3) technology in the maritime communication network. Communications. 2020; 17(5):223-34.

[10] Hu GA, Li CY. Dynamic Spectrum Allocation Method of Mobile Ship's Wireless Communication Network. Journal of Coastal Research. 2019:680-5.

[11] Lee J, Ko H. Reliability-guaranteed multipath allocation algorithm in mobile network. Etri Journal. 2022; 44(6):936-44.

[12] Xu LW, Quan TQ, Wang JJ, Gulliver TA, Le KN. GR and BP neural network-based performance prediction of dual-antenna mobile communication networks. Computer Networks. 2020; 172.

[13] Chen WS, Demirkol I, Mostafa M, Perotti A. Mobile communications and networks. IEEE Communications Magazine. 2021; 59(9):50-.

[14] Rajesh De, Ipseeta Nanda. Study of electromagnetic radiation on flower. Matrix Science Mathematic. 2022; 6(2): 58-63.

[15] Mu JS, Ouyang WJ, Hong T, Yuan WJ, Cui YH, Jing ZX. Digital Twins-Enabled Federated Learning in Mobile Networks: from the Perspective of Communication- Assisted Sensing. IEEE Journal on Selected Areas in Communications. 2023; 41(10):3230-41.

[16] Guo ZM, Ren XC, Ren F. Better realization of mobile cloud computing using mobile network computers. Wireless Personal Communications. 2020; 111(3): 1805-19.

[17] Keum CM, Liu SY, Al-Shadeedi A, Kaphle V, Callens MK, Han L, et al. Tuning charge carrier transport and optical birefringence in liquid-crystalline thin films: a new design space for organic light-emitting diodes. Scientific Reports. 2018; 8.

[18] Zhou B, Xu XT, Liu JG, Xu XK, Wang NX. Information interaction model for the mobile communication networks. Physica a-Statistical Mechanics and Its Applications. 2019; 525:1170-6.

[19] Matinmikko-Blue M, Yrjölä S, Ahokangas P. Spectrum management for local mobile communication networks. IEEE Communications Magazine. 2023; 61(7):60 -6.

[20] Wang KK, Chen QA, Jiang C, Chen ZF, Tan S, Lu MZ, et al. Narrow linewidth electro-optically tuned multi-channel interference widely tunable semiconductor laser. Optics Express. 2023; 31(3):4497-506.

[21] Liu L, Liao SS, Xue W, Yue J. Tunable all-optical microwave filter with high tuning efficiency. Optics Express. 2020; 28(5):6918-28.

[22] Singh R, Kaushik A, Shin W, Alexandropoulos GC, Toka M, Di Renzo M. Indexed multiple access with reconfigurable intelligent surfaces: the reflection tuning potential. IEEE Communications Magazine. 2024; 62(4):120-6.

[23] Kullan M, Mahadevan S, Johnson AR, Muthu R. Decoupled indirect duty cycle PWM technique with carrier frequency adjustment for a matrix converter. Turkish Journal of Electrical Engineering and Computer Sciences. 2018; 26(1):270-82.

[24] Morales-Aragonés JI, Williams MS, Gómez VA, Gallardo-Saavedra S, Redondo-Plaza A, Fernández-Martínez D, et al. A Resonant Ring Topology Approach to Power Line Communication Systems within Photovoltaic Plants. Applied Sciences-Basel. 2022; 12(16).

[25] Castillo RAJ, Grünheid R, Bauch G, Wolff F, von der Heide S. Communication analysis between an airborne mobile user and a terrestrial mobile network. IEEE Transactions on Vehicular Technology. 2018; 67(4):3457-65.

[26] Lee W, Suh ES, Kwak WY, Han H. Comparative analysis of 5G mobile communication network architectures. Applied Sciences-Basel. 2020; 10(7).

[27] Zhang K, Zhao XH, Peng Y, Yan KC, Sun PY. Analysis of mobile communication network architecture based on SDN. Journal of Grid Computing. 2022; 20(3).

[28] Usman IB, Matsoso BJ, Erasmus R, Coville NJ, Wamwangi DM. The role of carrier gas on the structural properties of carbon coated GaN. Materials Today Communications. 2021; 27.

[29] Kim J, Ladosz P, Oh H. Optimal communication relay positioning in mobile multi-node networks. Robotics and Autonomous Systems. 2020; 129.

[30] Muttiah R. Satellite constellation design for 5G wireless networks of mobile communications. International Journal of Satellite Communications and International Journal of Satellite Communications and Networking. 2023; 41(5):441-59.

[31] Ahmad AM, Barbeau M, Garcia-Alfaro J, Kassem J, Kranakis E. Tuning the demodulation frequency based on a normalized trajectory model for mobile underwater acoustic communications. Transactions on Emerging Telecommunications Technologies. 2019; 30(12).

Human-Computer Interaction Standardization and Systematization Development

An English Education Informatization Perspective

Xiaoling Lyu[1], Hongmiao Yuan[2]*, Zhao Zhang[3]
School of Foreign Languages, Shanghai Zhongqiao Vocational and Technical University, Shanghai 201500, China[1, 2]
Shanghai Er Yan Economic and Technical Consulting Co., LTD., Shanghai 200001, China[3]

Abstract—Amidst the information technology boom, this study harnesses IT and human-computer interaction to revolutionize English education. Our model, grounded in literature review and fieldwork, implemented teaching experiments that enhanced students' English proficiency by 25%, particularly in listening and speaking. Student engagement and interest surged by 30%, underscoring the effectiveness of standardized, systematized English education in the digital era. The study advocates for broader adoption of informatization in teaching, emphasizing the pivotal role of teachers in facilitating this educational shift. With significant outcomes, our research paves the way for future enhancements in English education, ensuring quality and equity in learning. Our approach addresses the gap between traditional teaching and technological advancements, offering personalized learning experiences that improve student outcomes. It also ensures consistent teaching quality and bridges educational divides. We introduce an informatization-based English education model, supported by literature review, fieldwork, and teaching experiments. Our findings show significant improvements in students' English proficiency, highlighting the model's effectiveness.

Keywords—Human-computer interaction; information technology; English education; standardization; systematization

I. Introduction

Human-computer interaction and intelligent system technologies are widely used in various industries. New embedded system technologies such as flipped classrooms and Internet+ have brought more possibilities to education. Teachers and students can tutor and communicate anytime, anywhere [1]. Teachers and students can tutor and communicate anytime and anywhere. Even if teachers use information technology to teach, they need to advocate student-centered teaching, and the role of teachers in teaching should not be ignored. However, teachers must realize that the arrival of the information technology era has brought challenges to teachers [2]. Teachers should continue to improve their knowledge systems, make efforts to learn information technology, and improve their ability to teach with information technology.

The Ministry of Education pointed out 2022 that there are still some things that could be improved in the current development of education informatization. There are both external and internal factors. The influence of internal factors is mainly the teachers' informatization teaching ability [3]. In the information age, teachers' informatization teaching ability is reflected not only in teachers' ability to apply information technology but also in integrating information technology and teaching content so that information technology is genuinely integrated into teaching [4]. This also brings difficulties for teachers in improving their informational teaching ability. The emergence of English teaching in an informational background provides teachers with a complete knowledge system of informationalized teaching and helps them deeply integrate the content with information technology.

Informationalized Background English Teaching was proposed by American scholars who added technical knowledge to the PCK. The emergence of English teaching in an informational context can help teachers improve the knowledge system of informationalized teaching competence and help them better integrate technical knowledge and content knowledge in the teaching process [5]. English teaching in an informational context describes the body of knowledge that teachers need to teach effectively in the Context of the information age.

In 2017, the Ministry of Education (MOE) suggested that advanced informatization teaching resources can drive and promote higher education teaching reform, break through traditional classroom limitations, and create a flexible and versatile teaching atmosphere [6]. With the addition of listening and speaking tests to the new round of Grade 4 and 6 exams, teachers face significant challenges in teaching to improve students' listening and speaking skills. The improvement of speaking and listening ability needs the support of information technology. Higher education English teachers must continuously improve their information technology teaching ability based on their original teaching ability and use information technology appropriately according to the needs of different teaching contents [7]. Teachers are users of information technology and are given an active role in teaching English in the Context of teachers' informationalized teaching competence. It is necessary to conduct an in-depth study on the informatization teaching ability of higher education English teachers from the perspective of English teaching in an informatization context in order to help higher education English teachers improve their informatization teaching knowledge system and find a suitable way for higher education English teachers to improve their informatization teaching ability.

Previous English education studies concentrated on conventional teaching, with scant attention to IT and human-computer interaction. Although IT's role in boosting learning is recognized, its systematic effect on English education's standardization is under-researched. Our research stands out by utilizing advanced IT and thoroughly analyzing its effects on teaching and student involvement. We present a new model combining literature, fieldwork, and data, showing marked enhancements in English skills, especially listening and speaking. The novelty lies in the model's empirical validation, highlighting significant gains in student performance and interest, bridging a research gap and offering a strong framework for educational advancement.

The era of education informatization brings unprecedented challenges to teachers and students. For English teachers, improving their informatization teaching ability has been an urgent problem [8]. The purpose and significance of this study are to analyze the differences and deficiencies in the informatization teaching ability of higher education English teachers of different teaching ages under different course types from the perspective of informatization contextualized English teaching in order to help teachers reflect on their informatization teaching ability under the guidance of informatization contextualized English teaching, to improve the knowledge system of teachers' informatization teaching ability [9]. In addition, this study also learned about the attitudes of the three developmental pathways of teaching English in an informational context through interviews [10]. Since there is no empirical research on the three developmental paths in China, this study encourages higher education English teachers to design speaking and reading courses from the perspectives of informatization background English teaching and the three developmental paths and to find out teachers' attitudes toward the three developmental paths at the end of the semester. The results of the interviews will likely provide a reference for the informatization teaching ability of higher education English teachers in China [11]. This study used a combination of questionnaire survey, classroom observation, and interviews to explore the differences in higher education English teachers' informationalized teaching competence from the perspective of informationalized contextual English teaching to help higher education English teachers construct and improve their informationalized teaching competence.

Continuing from this intro, Section II reviews literature on English edtech. Section III describes our research approach and data analysis. Section IV shows results, noting student progress. Section V discusses implications for English education. Section VI concludes with findings and future integration suggestions.

II. Theoretical and Contextual Foundations

Against the backdrop of science and technology, information technology has become an indispensable part of people's lives and learning and has shown a normalization trend. This trend also extends to the field of education, prompting teachers to utilize information technology to carry out teaching activities and requiring them to deeply integrate information technology with subject teaching to improve teaching efficiency [12]. Therefore, the rapid development of science and technology and information technology in the new era has put forward requirements for teachers' ability to apply information technology, which is the background of the study in this paper [13]. The second reason is that national policy documents pay attention to and focus on educational information 2022. The Ministry of Education advocates that teachers take the initiative to adapt to the challenges of the informatization society and cultivate students' ability to learn independently, cooperatively, and with inquiry in all aspects. The third reason is the importance of cultivating developed and systematic competence in teaching English as a foreign language to unmotorized teachers. As the leading force for future teachers, improving the standardization and systematization of informalized English teaching teachers will profoundly affect the quality of the training of talents in primary education in China. With a background of growing up in the era of digital natives, they can master information technology and update their teaching philosophy more quickly and are considered to be the promoters of educational reform. The information technology stage is necessary for teachers to integrate information technology and subject knowledge, and the development of teacher trainees' ICT competence plays a specific role in implementing concepts related to the new curriculum standards at the in-service stage. In the standardization of English education, it is usually necessary to set the core issues of standardization, and the principle of setting the core issues is mainly based on the level of information technology application. The Mean and SD tests shown in Table I are both rational value ranges of past research and have statistical significance.

TABLE I. English Standardized Question Set

Project	Mean	Standard Deviation
In the post-pandemic era, I believe it is important to cultivate the information technology teaching abilities of English teachers.	5.014	0.847
I can actively apply information technology to optimize English classroom teaching during the teaching process	5.614	0.664
The deep integration of information technology and teaching by English teachers can help improve students' English learning methods and enhance classroom teaching effectiveness. I believe that the deep integration of information technology and teaching by English teachers can help improve students' English learning methods and enhance classroom teaching effectiveness	5.114	0.814
I can use multimedia computer-assisted English classroom teaching	4.361	0.947
In the process of English teaching, I can use various information-based teaching resources	3.362	0.814
I usually use search engines and major websites to obtain English teaching resources	2.814	0.336
I can use information technology tools, combined with textbooks, to process and process English teaching materials obtained online	1.362	0.947

However, some problems still need to be solved with information technology teachers' ICT competence in China. The study also shows that the overall ICT competence of our student teachers is at an intermediate level, and the use of information technology to support teaching and learning still needs to be improved. There are also problems, such as a weak ability to apply interactive multimedia and a weak ability to optimize teaching practice [14]. The cultivation of information technology application ability of information technology teachers also raises problems such as lack of comprehensive planning, blockage of information technology post-service integration channels, unsystematic curriculum system, and lack of practical support.

The fourth reason is the need to integrate the English language subject with educational technology. The development of ICT skills at the informational level cannot be based solely on specialized IT courses. At the same time, it should be integrated into developing IT awareness and practical skills in all disciplines. English is an essential subject at the primary education level [15]. English teachers should pay attention to the application of modern information technology and make full use of information technology in teaching to promote students' effective learning. To summarize, it is of great practical significance to investigate the current situation of the ICT competence of informationalized English teaching teachers, to find out the existing problems, and to put forward targeted and constructive suggestions to improve the ICT competence of informationalized English teachers, to strengthen the cultivation of informationalized English teachers' information literacy, and to improve the quality of English teaching.

First, this study compared the domestic and international literature on ICT competence to determine this study's research idea and framework. Then, in conjunction with previous research, a measurement tool was developed to understand the ICT competence of Chinese informationalized English teaching teachers. Next, semi-structured interviews will be combined to investigate the factors affecting ICT teachers' competence in standardizing and systematizing English language teaching [16]. Finally, the results of the study will be combined to reflect on the shortcomings of English teaching standardization and systematization competence in the training process of informationalized English teaching teachers and to propose suggestions for cultivating and improving the standardization and systematization competence of informationalized English teachers' English teaching [17]. Based on the theoretical study, through a mixed study of the current situation and problems of ICT competence of informationalized English teaching teachers in three teacher training colleges.

This study deepens the specificity of English ICT skills by combining the literature on the information teaching competence of informationalized teachers with the ICT competence standards for informationalized English teaching teachers [18]. By sorting out the dimensions, structures, and definitions of English teachers' information teaching competence, the theoretical research on English teachers' ICT competence needs more attention and focus, and the more adequate the academic discussion, the more accurate the definition of information teaching competence will be.

The connotation of ICT skills is further refined by exploring the ICT competencies of informational English teaching teachers. Informationalized English teachers are not in-service teachers and pay more attention to the learning and reserve of competence [19]. Therefore, through the construction of the ICT competence of informationalized English teaching teachers and the in-depth understanding of the current situation of ICT skills of informationalized English teachers, this study provides some additions and improvements to the development of the theory of teaching ICT competence of informationalized teachers and also contributes to the development of ICT in teacher training schools [20]. This study provides some data to support the study of the ICT competence of informationalized English teaching teachers at the practical level.

With the continuous development of digital information technology, integrating education and information technology is the trend of future educational development. As another critical component of improving teachers' ICT competence, IT training covers a broader range of areas. It is more potent than in-service training, which can systematically develop and train teachers' ability to standardize and systematize English teaching by setting up relevant courses and practice sessions. Through the informatization stage, teachers can better and faster integrate the means and methods of ICT competence into education and teaching. Therefore, this study contributes to the understanding of informationalized ELT teachers' practice of standardization and systematization of English language teaching competencies. It also explores the factors affecting the standardization and systematization of English language teaching competencies. For individual informationalized English teaching teachers, it can improve their awareness of standardization and systematization competence in English teaching, stimulate their initiative and enthusiasm in learning information technology, and provide a reference for informationalized teachers to develop in-depth competence in the integration of information technology and English teaching [21]. For the training institutions of informationalized teachers, suggestions are made for the relevant departments to develop the ICT competence of informationalized teachers in practice.

III. Research Methodology

A. Study Design

Teaching English in an informational context can help teachers think about what knowledge is needed to integrate information technology into teaching and how teachers should acquire that knowledge in depth. With the integration of information technology and English teaching, teachers must focus on more than just the use of technology and pay attention to the knowledge needed for teachers' informationalized teaching ability. The results of domestic and international studies show that teaching English in an informational background is essential in guiding teachers' informationalized teaching ability. This study examines the differences in informationalized background English teaching demonstrated by higher education English teachers of different teaching ages in different course types, as well as higher education English teachers' reflections on three paths for developing informationalized background English teaching. The three research questions to be addressed in this study are as follows, as shown in Fig. 1:

- What are the characteristics of higher education English teachers' informationalized teaching competencies based on the four components of TK, TCK, TPK, and teaching English in an informationalized context?
- How do higher education English teachers' presentations of the various components of teaching English in an informational context differ across course types?
- From the perspective of teaching English in an informational context, what are the reflections of English teachers in higher education on the three paths of development and suggestions for improving their informational teaching skills?

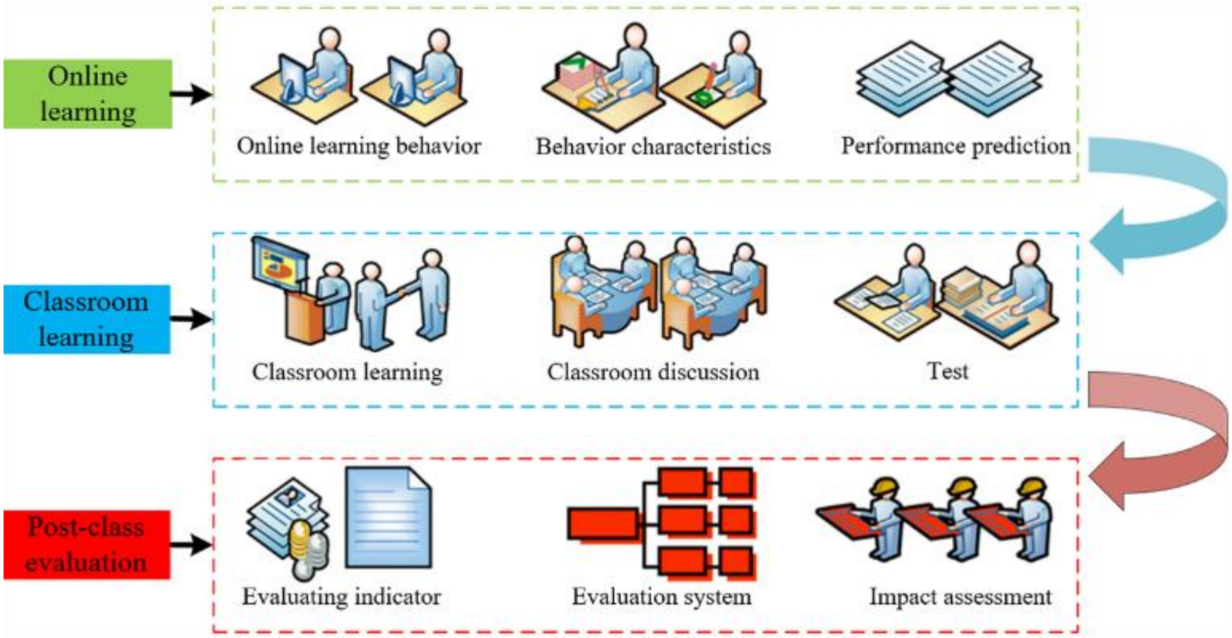


Fig. 1. Integrated learning and evaluation framework.

This study aims to instruct higher education English teachers' informationalized teaching ability from the perspective of teaching English in an informationalized context to effectively improve higher education English teachers' knowledge system and teaching ability. Strategies are provided for higher education English teachers to improve their informatization teaching ability. This study mainly divides the problem setting into three directions: systematic setting under the background of standardized English teaching; Coupling statistics of standardization and systematization in English education; Integration analysis of information-based teaching. As shown in Table II, the systematic setting under the background of standardized English teaching mainly analyzed the problems from the perspective of course preparation, with a Mean value range of 4-7 (P<0.01), and all passed the SD test.

TABLE II. SYSTEMATIZED SETTINGS FOR ENGLISH LANGUAGE TEACHING

Project	Mean	Standard Deviation
I will use online resources and information technology tools to prepare lessons according to the English curriculum standards	6.325	0.847
I will use information technology to create English teaching courseware for different types of courses based on learning objectives	5.324	0.784
I will combine the characteristics of students and use information technology to create English learning scenarios in teaching design, solving the key and difficult points in English teaching from shallow to deep I will combine the characteristics of students and use information technology to create English learning scenarios in teaching design, solving the key and difficult points in English teaching from shallow to deep	6.31	0.984
I will provide personalized self-directed learning guidance to students before class using WeChat, QQ, Campus Pass, etc.	5.32	0.336
I will use various information technology resources to collect different English teaching materials and exercises based on the differences in students' learning situations and assign different homework levels to students. I will use various information technology resources to collect different English teaching materials and exercises based on the differences in students' learning situations, in order to assign different levels of homework to students	4.98	0.541

First, quantitative data were collected through questionnaires. Second, qualitative data were collected through classroom observations and interviews. Finally, the above research questions were derived through data analysis. This study takes English teachers in higher education as the research object and uses questionnaires and other methods to conduct the study. Due to the addition of speaking classes in higher education, previous studies needed to have investigated higher education English teachers' informationalized teaching ability under different course types. Therefore, this study was conducted to experiment with higher education English teachers as research subjects.

In order to understand the differences in the components of teaching English in an informational context presented by higher education English teachers in different course types, this study collected data for analysis through classroom observations. One higher education student from eight higher education institutions was selected for classroom observation, and three higher education English teachers of different teaching ages were randomly invited to participate in classroom observation. Based on the teaching age of the English teachers in this higher education, the teaching age in higher education was categorized into three stages: less than ten years, 11-20 years, and more than 21 years. Based on the interview outline, this study conducted semi-structured interviews with three higher education English teachers of different teaching ages. The interviews aimed to understand the reflections of higher education English teachers of different teaching ages on the three ways of developing English language teaching in an informational context.

In order to address the second research question of this study: how do higher education English teachers' presentations of the components of teaching English in an informational context differ across course types? The classroom observation method was used in this study. In order to improve students' listening and speaking skills, higher education added speaking classes to the existing English teaching. With the permission of the school and three teachers, the researcher of this study conducted classroom observations and collected data as a non-participant. The purpose of the classroom observations was to compare the differences in the knowledge components of teaching English in an informational context presented by higher education English teachers of different teaching ages in different types of courses, as well as the deficiencies in the teachers' informational teaching skills after understanding the meaning of teaching English in an informational context. Emphasis was placed on documenting the application of the seven components of informational contextual English teaching in different course types and the frequency of the seven components in different course categories. The Adaptive Classroom Observation Scale rated teachers' teaching based on the seven components of teaching English in an informational context. The differences between the seven informational Context English teaching components in different course types were summarized by analyzing the mean values. Human-computer interaction has to go through a certain process so that it can realize an effective cycle, and the specific process is shown in Fig. 2.

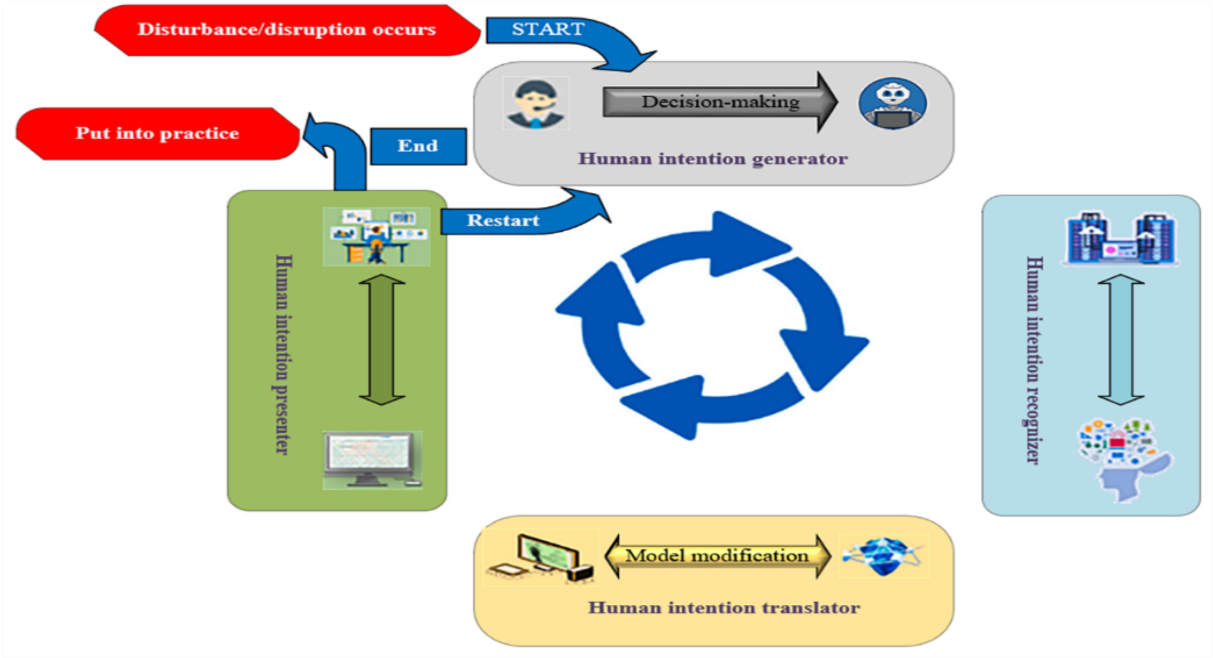

Disturbance/disruption occurs
START
Decision-making
Human intention generator
Put into practice
End
Restart
Human intention presenter
Human intention recognizer
Model modification
Human intention translator

Fig. 2. Human computer interaction.

English education modeling in the context of information, as shown in Eq. (1), (2), and (3):

$$q_j^F(t+1) = q_j^F(t) + a_j(t) - \sum_{i\in\Omega_j^f} m_{ij}(t), \forall j \tag{1}$$

$q_j^F(t+1)$ is the iteration result of the q-function; $q_j^F(t)$ is the previous generation value of the *q*-function.

$$f_{ij}^{tran}(t) = \alpha_{ij} m_{ij}(t), \forall j \tag{2}$$

$f_{ij}^{tran}(t)$ is the training function of the f-function of the train and $\alpha_{ij} m_{ij}(t)$ is the lower-level function of the *f*-function.

$$f_{ij}^{work}(t) = \gamma_j (A_j(t) - a_j(t)), \forall j \tag{3}$$

$f_{ij}^{work}(t)$ is the training function for the work of the f function, and γ_j is the coefficient of $(A_j(t) - a_j(t))$.

B. Related Theoretical Designs

In order to address the third research question of this study, what are the reflections of higher education English teachers on the three development paths from the perspective of teaching English in an informational context? What are the suggestions for improving their informationalized teaching skills? The qualitative research method of semi-interviews was used in this study. The interview entailed using three tools: an outline, a tape recorder, and a transcript.

Semi-structured personal interviews were conducted with each of the three higher education English teachers who participated in the classroom observation with their permission. The interview outline was designed from three dimensions: first, to collect the attitudes and reflections of three higher education English teachers of different teaching ages on the three developmental paths from the perspective of teaching English in an informational context; second, to use interviews to share strategies to improve informational teaching competence with three higher education English teachers of different teaching ages, and the last dimension was to understand the factors affecting higher education English teachers' informational factors that affect the development of teaching competence.

The standardization and systematization of English education require relevant coupling statistics, especially considering the mediating effect of information technology background. The mean of the problem and SD test are also included in the table, with a mean of 1-5. Except for the low mean of information technology utilization level (2.624, $P<0.05$), all others meet the ideal range of this study. The specific problem design is shown in Table III.

TABLE III. STANDARDIZED AND SYSTEMATIC COUPLING STATISTICS FOR ENGLISH LANGUAGE EDUCATION

Project	Mean	Standard Deviation
I will use information technology such as English audiovisual teaching resources to introduce scenarios, arouse students' interest, and maintain their attention. I will use information technology, such as English audiovisual teaching resources, to introduce scenarios	4.051	0.947
I will use information technology to design enjoyable English activities, encourage students to participate actively in classroom interaction, change the teaching method of "teachers full of words," and effectively carry out students' independent, cooperative, and exploratory learning. I will use information technology to design some exciting activities to encourage students to actively participate in classroom interaction, change the teaching method of "teachers full of words," and effectively carry out students' independent, cooperative, and exploratory learning	4.821	0.784
I will use various English learning software according to different teaching contents, such as "English Fun Dubbing," "Hundred Words Zhan," "Daily English Listening," "Tianxue.com," etc., to help students improve their information technology literacy and English listening, speaking, reading, and so on. I will use various English learning software according to different teaching contents such as "English Fun Dubbing," "Hundred Words Zhan," "Daily English Listening," "Tianxue.com," etc., to help students improve their information technology literacy and English listening, speaking, reading, and writing abilities. and writing abilities	4.621	1.032
When summarizing in class, I can use information technology to assist myself in summarizing in class	3.981	1.041
I will handle unexpected information technology issues in the classroom appropriately, such as teaching equipment malfunctions	4.011	1.006
I will use information technology tools to design and distribute survey questionnaires, collect students' English learning situations, and use technical means to analyze data, making appropriate adjustments and guidance for students' future English learning activities. I will use information technology tools to design and distribute survey questionnaires, collect students' English learning situations, and use technical means to analyze data, making appropriate adjustments and guidance for students' future English learning activities	2.624	1.021
I will encourage students to use online English teaching platforms for self-directed learning actively	3.682	1.417

The experimental procedure was divided into five phases: pilot testing, formal distribution of the questionnaire, classroom observation (speaking sessions), classroom observation, and interviews. The experiment started in September 2021, the first semester of the first year of higher education.

The questionnaire was divided into a pilot test phase and a formal questionnaire phase. The pilot test questionnaire was distributed to 20 higher education English teachers during the first week and was modified accordingly by analyzing the data and incorporating teachers' feedback. During the first week, the formal questionnaire was distributed to 146 higher education English teachers who participated in this survey, and the questionnaire took about 15 minutes to answer.

There were 18 weeks in the fall semester of 2021, but weeks 14-18 were mainly review sessions and final exams, so the classroom observations for this study were conducted from week 2 to week 12. A total of 10 classroom observations were conducted in this study; week 6 was a national holiday, so no classroom observations were conducted in week 6. All three higher education English teachers of different teaching ages taught two types of classes: reading classes and speaking classes. The instructional goal of the speaking class was to improve students' listening and speaking skills, so the content knowledge of the speaking class consisted of speaking exercises and listening exercises. Since English classes in higher education are categorized into reading and speaking classes, this study formulates English classes as reading classes, whose goal is to develop students' reading and writing skills. Therefore, the content knowledge of the reading class includes grammar, reading, and writing. Three higher education English teachers observed one speaking class and one reading class per week. Prior to the classroom observations, the connotations of the components of informational contextualized English teaching and the three developmental pathways were communicated to the three first-year higher education English teachers who participated in the classroom observations, and the three teachers were encouraged to informationalize their teaching and learning from the perspective of informational contextualized English teaching and the three growth pathways.

The experimental density of information-based teaching is analyzed using clustering software. The red area represents the standardized and systematic practice of English teaching achieved through good use of information technology during this time period. The blue area represents the low-density clustering area, indicating that there is no standard teaching that has achieved good implementation of information technology. The overall test results are shown on the right side of the picture, as shown in Fig. 3.

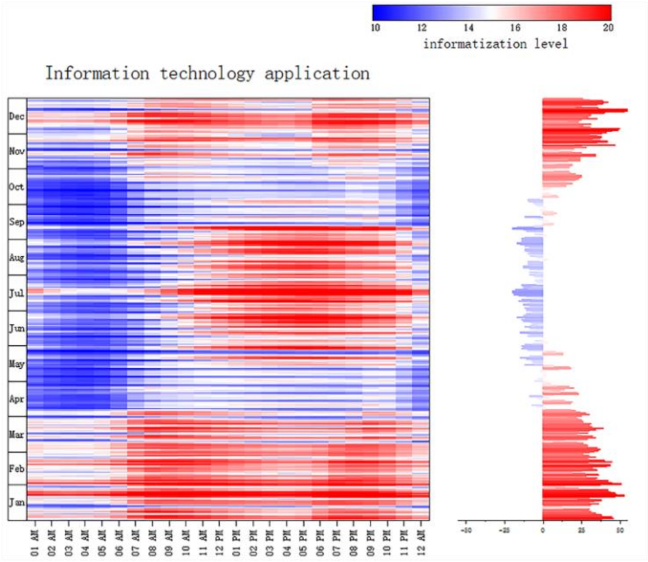


Fig. 3. Application of information-based English language teaching and learning.

Formal semi-structured interviews were conducted in week 13, in which the three teachers were interviewed face-to-face according to the interview outline. The researcher also communicated with the three teachers after class to share new issues identified during classroom observations and to revise the interview outline promptly. The questionnaire data were analyzed mainly using SPSS 22. First, the internal consistency of the reliability of each subscale and the whole scale was tested. Second, the descriptive analysis function in SPSS 22 was used to statistically analyze the data in each subscale, which mainly involved frequency distribution and percentage. The data of classroom observation were mainly from the Classroom Observation Scale. In order to explore the differences in how higher education English teachers of different teaching ages present each knowledge component of informational contextual English teaching in different course types, this paper conducted a descriptive analysis of the collected data. The frequencies and means of the different components of teaching English in an informational context in different course types were analyzed.

The transcription of the interviews was the change of the interview conversation from spoken to written form, which made the results easier to analyze, and the transcription process itself was the initial analysis process. At the end of the transcription process, the transcribed data was sent to the interviewees for confirmation to ensure the validity of the interviews. The interview data were analyzed by dividing the interview outline into dimensions, coding each teacher's interview data according to these dimensions, and organizing the interview data related to the theme of this study. The first dimension is higher education English teachers' reflections on the three development paths from the perspective of teaching English in an informational context; the second dimension is the strategies for improving higher education English teachers' informationalized teaching competence, and the third dimension is the factors affecting the development of higher education English teachers' informationalized teaching competence.

In the process of analyzing the integration of information technology teaching, the main approach is to examine the ways in which teachers can stimulate students' learning enthusiasm in order to achieve the integration of information technology teaching. Even in the issue of "summarizing using information technology", Mean reached 7.824, which achieved good integration of information technology, as shown in Table IV.

TABLE IV. ANALYSIS OF THE INTEGRATION OF INFORMATION-BASED TEACHING

Project	Mean	Standard Deviation
I will use information technology such as English audiovisual teaching resources to introduce scenarios, arouse students' interest, and maintain their attention. I will use information technology, such as English audiovisual teaching resources, to introduce scenarios	4.051	0.074
I will use information technology to design enjoyable English activities, encourage students to participate actively in classroom interaction, change the teaching method of "teachers full of words," and effectively carry out students' independent, cooperative, and exploratory learning. I will use information technology to design some exciting activities to encourage students to actively participate in classroom interaction, change the teaching method of "teachers full of words," and effectively carry out students' independent, cooperative, and exploratory learning	2.621	0.684
I will use various English learning software according to different teaching contents, such as "English Fun Dubbing," "Hundred Words Zhan," "Daily English Listening," "Tianxue.com," etc., to help students improve their information technology literacy and English listening, speaking, reading, and so on. I will use various English learning software according to different teaching contents such as "English Fun Dubbing," "Hundred Words Zhan," "Daily English Listening," "Tianxue.com," etc., to help students improve their information technology literacy and English listening, speaking, reading, and writing abilities. and writing abilities	5.362	0.638
When summarizing in class, I can use information technology to assist myself in summarizing in class	7.824	0.947

IV. Results and Discussion

A. English Standardization and Systematization Process

Since the entrance examination reform for higher education institutions has added speaking and listening tests, higher education has reformed the English curriculum and added speaking classes to the original curriculum to cultivate students' speaking and listening skills. Although English-speaking classes in higher education have yet to be fully implemented, English teachers must also develop students' listening and speaking skills in English teaching. With the background of curriculum reform and informationalized teaching and learning, this study used classroom observation to understand the characteristics of informationalized background English teaching exhibited by first-year English teachers in higher education of different teaching ages in speaking and reading classes. The following conclusions were drawn.

First, because speaking classes need to develop students' listening and speaking skills, the three first-year higher education English teachers of different teaching ages used IT more frequently and, on average, in speaking classes than in reading classes. In addition to CK, the other six components (TK et al., PCK, TPK, and teaching English in an informational context) were more frequently used in speaking classes than reading classes, and information technology was more diverse. Teachers have good comprehensive skills in teaching with information technology. It can be found from Ms.C's lessons that although TK, TCK, TPK, and informationalized contextual English teaching are more frequent in speaking classes than in reading classes, Ms.C is influenced by her previous teaching experience and PK, CK, and PCK account for a large proportion of speaking and reading classes. The above analysis shows that courses affect English teachers' ability to teach information technology in higher education. Since speaking classes require information technology to support the presentation of content knowledge, the three higher education English teachers were more likely to use components related to technological knowledge in their speaking classes. Therefore, higher education English teachers of different ages must experiment with using different information technologies to explain content knowledge in their reading classes. Consistent with other quantitative studies, this study also conducted descriptive statistics. However, in this study, information support was further categorized by gender to objectively examine the sensitivity of information technology to gender. Among them, 624 male samples were selected and 1204 female samples were selected, as shown in Table V.

TABLE V. Description of Informational Teaching by Gender Dimension

Dimension (math.)	Distinguishing between the sexes	Sample size	Statistical value M	T	SIG
ICT capacity	male	624	4.05	1.624	0.214
	daughter	1204	3.68		
Basic technical literacy	male	624	4.51	1.051	0.032
	women	1204	3.91		
Technical Support Learning	male	624	4.51	1.324	0.141
	women	1204	3.62		
Technical Support Learning	male	624	4.44	1.214	0.362
	women	1204	3.92		

In the current Context of curriculum reform, the first path (from PCK to informational context English teaching) and the second path (from TPK to informational context English teaching) under informational Context English teaching can meet the needs of two courses. Higher education English teachers can choose the appropriate pathway according to the needs of their courses. For example, teachers can choose the second path (from TPK to teaching English in an informational context) when preparing for speaking classes. The teacher considers the role of technology in the speaking classroom, so the teacher first develops TK and TPK and then combines TPK and PCK to develop informational contextual English teaching. When preparing for the reading lesson, the teacher chooses the first path from PCK to informational contextualized ELT. In reading lessons, teachers focused more on content knowledge, so teachers first developed PK and CK and then used technology to enhance and structure instruction. However, prior teaching experience and less technological knowledge influenced the pathway choice, and only the first pathway was chosen to develop informationalized teaching skills. However, with sufficient technological knowledge, the appropriate path would also be chosen depending on the type of course. In addition to the first and second development paths that can help higher education English teachers develop informational contextualized English teaching, this study provides three strategies for higher education English teachers to develop informational contextualized English teaching and to improve their informationalized teaching competence based on the results of the interviews and the literature review.

The systematic testing of English teaching can be found in Eq. (4) and (5):

$$e_i^t = x_i(t) - y_i(t) + d_i(t) - c_i(t) + z_i(t) + \sum_{\forall k \in I/i} u_{ik}(t), \forall i \quad (4)$$

$\sum_{\forall k \in I/i} u_{ik}(t), \forall i$ is the system composite value of the systematic test.

$$u_{ii}(t) = 0, \forall_i = L \tag{5}$$

$u_{ii}(t) = 0$ is the minimum value of the least squares method.

First, establish a school collaboration mechanism. Research has shown that informational contextual English teaching can help higher education English teachers build a knowledge system of informational teaching ability. Therefore, informational contextual English teaching can be extended to higher education, and higher education English teachers from different schools can regularly engage in collaborative learning. In the exchange and learning, teachers can discuss the connotation of the seven components of informational background English teaching, share how new information technology can be applied to higher education English teaching in the Context of informational background English teaching, and use informational background English teaching to promote the improvement of higher education English teachers' informationalized teaching competence.

Secondly, educational research activities are organized. The school actively conducts teaching and research activities against the background of teaching English in an informational context and organizes public class demonstrations and after-class seminars. When evaluating public courses, teachers can discuss the strengths and weaknesses of public courses against the background of knowledge of teaching English in an informational context. Public classroom demonstrations can help higher education English teachers of different teaching ages learn from each other. Teachers with higher informationalized teaching competence can provide training on informationalized teaching competence to other higher education English teachers. Teachers with high teaching ability can promote new information technology, instruct other teachers on how to apply information technology and develop teachers' ability to integrate information technology with English teaching.

Finally, cultivate teachers' sense of reflection. Informative teaching is new to teachers who implement information teaching. Therefore, besides external help, teachers should know the need to improve their informational teaching ability actively. Teachers should develop the habit of reflecting on informalized teaching from the perspective of teaching English in an informative context so that they can discover the deficiencies of the information teaching knowledge system in the teaching process and improve the implementation process of formalized teaching in time. Teachers should learn to improve their instructional teaching ability gradually through reflection. Higher education English teachers can reflect on the insufficiency of the knowledge of informatization teaching ability based on the informatization background of English teaching, and they can also learn whether the application of information technology in each lesson can help students understand and master the content knowledge through the students' feedback after the lesson. Higher education English teachers should accumulate information technology teaching experience and improve information technology teaching ability through reflection.

Conduct a degree segmentation study on descriptive statistics, using a sample size of 254 for undergraduate students and 663 for master's students. Sensitivity considerations for information technology are divided into three aspects: ICT competence, basic technical literacy, and technical support for learning; Two tests were conducted on technical support learning, as shown in Table VI.

TABLE VI. DESCRIPTION OF INFORMATIVE INSTRUCTION IN THE DEGREE DIMENSION

dimension (math.)	places	sample size	Statistical value M	T	SIG
ICT capacity	undergraduate (adjective)	254	4.01	-1.62	0.974
	bachelor's degree	663	4.41		
Basic technical literacy	undergraduate (adjective)	254	3.58	-0.964	0.914
	bachelor's degree	663	6.36		
Technical Support Learning	undergraduate (adjective)	254	3.51	-0.657	0.635
	bachelor's degree	663	4.05		
Technical Support Learning	undergraduate (adjective)	254	3.3	-0.241	0.874
	bachelor's degree	663	3.91		

B. *Technical Support for Information Technology Teaching*

Most higher education English teachers used to think that they could utilize their informal teaching ability as long as they were skilled in using information technology in teaching. Teachers need to understand the complete knowledge system of informationalized teaching and improve their informationalized teaching ability. The emergence of English teaching in an informational context fills the teachers' knowledge system gap. It helps English teachers in higher education to understand the knowledge that should be included in complete informational teaching. The seven components of informational contextual English teaching can help higher education English teachers improve their professional knowledge and provide guidance for improving teachers' informationalized teaching ability. Higher education English teachers should draw on the knowledge system of informationalized background English teaching, deeply understand and master the connotation of informationalized background English teaching, find out its shortcomings, and improve their knowledge system in time from the perspective of informationalized background English teaching. Through a deep understanding of the specific content of informationization background English teaching, teachers can integrate information technology into teaching, improve the quality of information technology teaching, and perfect how information technology is used. In addition, by applying informationized background English teaching to speaking and reading classes, higher education English teachers can promptly reflect on the deficiencies in teaching and apply the seven components of informationized background English teaching in different course types. In this study, a certain clustering analysis was conducted on the standardization and systematic practice of English education under the background of informatization, which mainly includes five aspects, including "English standardization and systematic practice, English standardization research, English systematic research", etc. Specifically, the darker colors of each aspect represent their quantitative clustering, and the lighter colors represent their spatial clustering, as shown in Fig. 4.

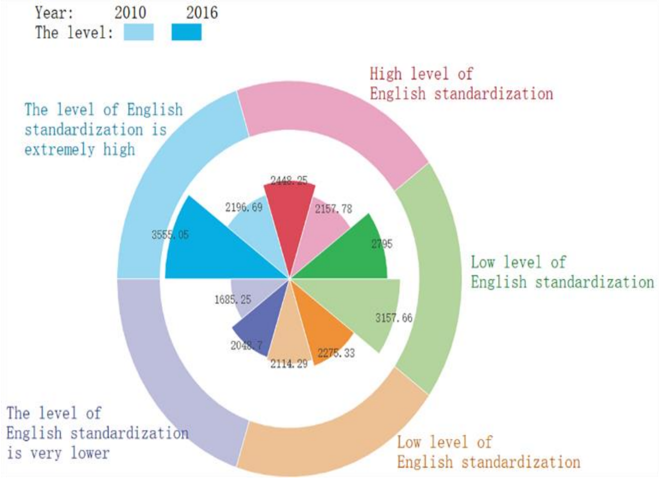


Fig. 4. Systematic research priorities in English language education.

With the support of informational contextualized English teaching, higher education English teachers' ultimate goal is to integrate their teaching so that the seven components of informational contextualized English teaching are integrated into their teaching, thus promoting the improvement of informationalized teaching skills. In order to help teachers develop informational contextualized English teaching, three pathways for developing informational contextualized English teaching are proposed. Applying the three developmental paths to higher education English teaching in China, the first and second of the three developmental paths could help teachers develop an informational background in English teaching. Three higher education English teachers of different teaching ages believed that the first path (from PCK to teaching English in an informational context) was suitable for reading classes, and the second path (from TPK to teaching English in an informational context) was suitable for speaking classes. The third path (PCK and informational contextual English teaching simultaneously) must be more suitable for higher education English teachers to develop informational contextual English teaching. However, with sufficient technological knowledge, the first path can be relied on to design English teaching for the time being. The above interview results provide path support for most higher education English teachers to improve their informationalized teaching ability. Higher education English teachers can choose appropriate paths for different types of courses to help teachers integrate information technology into English teaching. Most higher education English teachers have recognized the critical role of information technology teaching and the effectiveness of information technology background English teaching in perfecting information technology teaching ability. Therefore, teachers' use of informational background in English teaching and the first and second paths to guide informationalized teaching competence can reduce the pressure on teachers to improve their informationalized teaching competence. The above interview results about the paths will be informative for guiding the informatization teaching ability of English teachers in higher education at this stage.

Teachers' sense of responsibility and mission have the most significant impact on their development, and teachers' obligations and responsibilities to students and teaching affect their beliefs and efficiency in teaching and influence their teaching practices. Combined with the results of the interviews, teachers' perceived attitudes toward informationalized teaching can be divided into the following three categories. First, this informationalized teaching English teacher category has a strong sense of educational mission. Teachers in this category have a strong sense of mission and passion for teaching. It believes that English teachers should keep abreast of the times and continuously improve their ICT skills to improve teaching and students' learning efficiency. Teachers in this category consistently spoke highly of IT teaching in their interviews and agreed that integrating the English subject with IT was important.

Secondly, these teachers have a strong sense of mission and are willing to learn IT teaching courses. Information technology teaching English teachers have a strong sense of educational responsibility and mission. It recognizes the critical role of information technology in teaching English. Nonetheless, it could have shown a more vital willingness to use IT in the interviews. It is believed that the reasons for using IT tools in teaching came mainly from the demands of the external environment, including the recruiters' requirements for teachers' technological level and future career development.

Third, this category of teachers needs a stronger sense of mission and is influenced by cultural conceptions of the discipline. This category of informationalized teaching English teachers needs a stronger sense of educational mission and believes that technological tools only play a supplementary role in English teaching and are dispensable. Its conception of teaching is relatively traditional, believing that the English teacher's job is the most important and that he or she should focus on teaching language knowledge and grammar.

In the standardization and systematization of English education under the background of informatization, descriptive statistics should be conducted on the dimension of network devices. A sample of whether online network device information is used was used, with a sample size of 204 for "yes" and 782 for "no". Only a t-test was needed for the "yes" sample, and the specific results are shown in Table VII.

TABLE VII. DESCRIPTION OF INFORMATIONALIZED INSTRUCTION IN THE NETWORK DEVICE DIMENSION

Dimension (math.)	Network device dimension	Sample size	Statistical value M	T	SIG
ICT capacity	be	204	4.01	3.61	0.014
	clogged	782	4.41		
Basic technical literacy	be	204	3.58	4.68	0.019
	clogged	782	6.36		
Technical Support Learning	be	204	3.51	2.617	0.270
	clogged	782	4.05		
Technical Support Learning	be	204	3.3	6.354	0.035
	clogged	782	3.91		

In the interviews, its views strongly reflect the influence of cultural perceptions on the English language discipline. The subject of English is unique among many disciplines because of its dual characteristics of content and medium. Cultural perceptions of their subjects may influence the application of IT in specific subjects by ITEd English teachers. In conclusion, informationalized English teachers have a positive attitude towards teaching IT and are most particular about the importance of IT in English language teaching. They believe that IT supports English language teaching and learning by bringing authentic contexts and rich teaching resources, helping students develop an excellent cultural awareness and a willingness to learn and use new IT in practice. Therefore, the concepts and attitudes towards IT teaching will affect the IT teaching ability of English teachers of IT teaching.

ANOVA showed that there were significant differences in the standardization and systematization skills of ITEd English teachers in terms of teaching practice and that ITEd English teachers with teaching practice were significantly more likely to be involved in teaching practice than those without in all three sub-dimensions of ITEd. In the third section of the questionnaire, "participation in teaching practice" ranked second regarding the importance of influencing factors. Classroom practice and various practical activities of English teachers are essential factors that influence the development of their ICT competence.

In the interviews, all the teachers of IT-enabled English mentioned the critical word “little teaching practice”. They felt they needed more opportunities to integrate IT into their English teaching practice. PET 1, PET 2, PET 4, and PET 6 expressed similar views. They all felt that their teaching practice needed to be improved. Although they had participated in educational internships, they needed more opportunities to practice and improve their ICT skills in natural teaching environments. Therefore, it is desired that the number of practicum opportunities in terms of school curriculum or activity practices and educational internships be increased. Practicum can help individuals analyze and master technology in greater depth and is the ultimate way to consolidate and strengthen their IT teaching skills. Teachers of IT-enabled English with teaching experience can improve their ICT skills by constantly reflecting on and reconstructing their teaching practices after practice, analyzing and judging the appropriateness of the actions and timing of the use of IT in the classroom, and better integrating the subject matter with IT. However, many informationalized teaching English teachers reflected the need for teaching practice in their interviews. Due to the lack of teaching practice, informationalized teaching English teachers could not reflect on and improve the use of IT in the teaching process based on teaching effectiveness, which hindered the development of their IT competence. Therefore, it is significant for informationalized teaching English teachers to participate in teaching practice.

ANOVA showed significant differences in ITE English teachers' standardization and systematization abilities in school technology environments, suggesting that school technology support significantly impacts ITE English teachers' standardization and systematization abilities. Based on the literature review, the following analysis focuses on school support, teacher modeling, and curriculum. School is an essential stage for informationalized English teachers to learn subject knowledge and IT knowledge systematically, and the school environment inevitably affects the development of their ICT competence. In the survey, "school hardware and software support" ranked first in the list of possible factors affecting the growth of ICT competence. The timely updating of teaching equipment and the maintenance and repair of related equipment help to create a favorable IT teaching and learning environment. Several studies have shown that good teaching and learning equipment can significantly increase teachers' motivation to use technological tools.

V. CONCLUSION

In conclusion, this study has addressed the research questions posed in the methodology section, particularly focusing on the characteristics of university English teachers' information teaching capabilities, encompassing the four components of Technological Knowledge (TK), Technological Content Knowledge (TCK), Technological Pedagogical

Knowledge (TPK), and their integration within the context of English education informatization. Our findings confirm that the integration of information technology in English education significantly enhances teaching efficiency and effectiveness. The use of multimedia and digital resources has been shown to enrich the teaching experience, making it more dynamic and engaging, thereby stimulating students' interest and promoting their autonomous learning.

Moreover, our study has demonstrated that informatization significantly improves students' English listening and speaking skills. The utilization of multimedia teaching resources in authentic contexts has enhanced students' language perception and practical language use. Interactive methods, including online communication and simulated dialogues, have been instrumental in greatly improving students' communicative competence.

The study also established the reliability and validity of the questionnaire survey instrument used to collect quantitative data, ensuring that the research findings are robust and trustworthy. The results align with previous research, underscoring the importance of informatization in standardizing educational practices and narrowing the educational resource gap. The adoption of a unified teaching platform and resources ensures the consistency and accuracy of teaching content, contributing to higher quality education.

Furthermore, the research highlights the role of informatization in the systematization of English education, facilitated by big data and artificial intelligence. This systematization allows for a more scientific and precise analysis of students' learning, thereby enhancing teaching outcomes.

In light of these findings, it is evident that the standardization and systematization of English education through informatization play a crucial role in improving teaching effectiveness and stimulating students' learning interests and potential. Looking ahead, there is a clear need to further integrate information technology with English education, continuously optimizing teaching methods to align with the developmental demands of the information age. It is also imperative to focus on educational equity and accessibility, ensuring that every student has the opportunity to receive a high-quality English education.

This study offers valuable insights and lessons for the reform and innovation of English education, anticipating significant advancements driven by information technology. The empirical validation of our model, which has shown substantial increases in student performance and engagement, addresses a critical gap in the current literature and provides a solid framework for future research and practice in English education informatization.Our study, though advancing English education through informatization, has limitations. It is bounded to a specific context, potentially restricting the applicability of results elsewhere. Dependence on technology infrastructure varies and could impact outcomes. Data collection was limited in time, not capturing long-term effects. Future studies should include broader and more diverse samples over longer periods to enhance the model's sustainability and scalability.

REFERENCES

[1] Vallejo-Correa, P., Monsalve-Pulido, J., & Tabares-Betancur, M. (2021). A systematic mapping review of context-aware analysis and its approach to mobile learning and ubiquitous learning processes. *Computer Science Review*, *39*, 100335. https://doi.org/10.1016/j.cosrev.2020.100335

[2] Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, *9*(7), e21929. https://doi.org/10.2196/21929

[3] Rafiq, K. R. M., Hashim, H., & Yunus, M. M. (2021). Sustaining education with mobile learning for English for specific purposes (ESP): A systematic review (2012–2021). *Sustainability*, *13*(17), 9768. https://doi.org/10.3390/su13179768

[4] Stamm, T. A., Andrews, M. R., Mosor, E., Ritschl, V., Li, L. C., Ma, J. K., Campo-Arias, A., Baker, S., Burton, N. W., Eghbali, M., & others. (2021). The methodological quality needs to be improved in clinical practice guidelines in COVID-19: Systematic review. *Journal of Clinical Epidemiology*, *135*, 125–135. https://doi.org/10.1016/j.jclinepi.2021.03.005

[5] Surahman, E., & Wang, T.-H. (2022). Academic dishonesty and trustworthy assessment in online learning: A systematic literature review. *Journal of Computer Assisted Learning*, *38*(6), 1535–1553. https://doi.org/10.1111/jcal.12708

[6] Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. *Applied Sciences*, *11*(9), 3986. https://doi.org/10.3390/app11093986

[7] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

[8] Ferdousi, R., Arab-Zozani, M., Tahamtan, I., Rezaei-Hachesu, P., & Dehghani, M. (2021). Attitudes of nurses towards clinical information systems: A systematic review and meta-analysis. *International Nursing Review*, *68*(1), 59–66. https://doi.org/10.1111/inr.12603

[9] Seraj, P. M. I., Klimova, B., & Habil, H. (2021). Use of mobile phones in teaching English in Bangladesh: A systematic review (2010–2020). *Sustainability*, *13*(10), 5674. https://doi.org/10.3390/su13105674

[10] Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, p. *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068

[11] Garoufallou, E., & Gaitanou, P. (2021). Big data: Opportunities and challenges in libraries, a systematic literature review. *College & Research Libraries*, *82*(3), 410. https://doi.org/10.5860/crl.82.3.410

[12] Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., & others. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

[13] Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

[14] Fadahunsi, K. P., O'Connor, S., Akinlua, J. T., Wark, P. A., Gallagher, J., Carroll, C., Car, J., Majeed, A., & O'Donoghue, J. (2021). Information quality frameworks for digital health technologies: Systematic review. *Journal of Medical Internet Research*, *23*(5), e23479. https://doi.org/10.2196/23479

[15] Hamid, R. A., Albahri, A. S., Alwan, J. K., Al-Qaysi, Z., Albahri, O. S., Zaidan, A., Alnoor, A., Alamoodi, A. H., & Zaidan, B. (2021). How smart is e-tourism? A systematic review of smart tourism recommendation system applying data management. *Computer Science Review*, *39*, 100337. https://doi.org/10.1016/j.cosrev.2020.100337

[16] Zhang, M., Gibbons, J., & Li, M. (2021). Computer-mediated collaborative writing in L2 classrooms: A systematic review. *Journal of Second Language Writing*, p. *54*, 100854. https://doi.org/10.1016/j.jslw.2021.100854

[17] Yuan, R., Liao, W., Wang, Z., Kong, J., & Zhang, Y. (2022). How do English-as-a-foreign-language (EFL) teachers perceive and engage with critical thinking: A systematic review from 2010 to 2020. *Thinking Skills and Creativity*, *p. 43*, 101002. https://doi.org/10.1016/j.tsc.2022.101002

[18] Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from the 2012 to early 2020 public release. *Computer Assisted Language Learning*, *36*(3), 517–554. https://doi.org/10.1080/09588221.2021.1933540

[19] Gesualdo, F., Daverio, M., Palazzani, L., Dimitriou, D., Diez-Domingo, J., Fons-Martinez, J., Jackson, S., Vignally, P., Rizzo, C., & Tozzi, A. E. (2021). Digital tools in the informed consent process: A systematic review. *BMC Medical Ethics*, *22*, 1–10. https://doi.org/10.1186/s12910-021-00585-8

[20] Sophonhiranrak, S. (2021). A systematic review of features, barriers, and influencing factors of mobile learning in higher education. *Heliyon*, *7*(4). https://doi.org/10.1016/j.heliyon.2021.e06696

[21] Prilutskaya, M. (2021). Examining pedagogical translanguaging: A systematic review of the literature. *Languages*, *6*(4), 180. https://doi.org/10.3390/languages6040180.

Eavesdropping Interference in Wireless Communication Networks Based on Physical Layer Security

Mingming Chen*, Yuzhi Chen
College of Information and Smart Electromechanical, Engineering, Xiamen Huaxia University, Xiamen, 361024, China

***Abstract*—Effective communication security protection can protect people's privacy from being violated. To raise the communication security of wireless communication networks, a collaborative eavesdropping interference scheme with added artificial noise is proposed by combining physical layer security and clustering scenarios to protect the communication security of wireless sensor networks. This scheme adds artificial noise to the transmitted signal to interfere with the eavesdropping signal, making the main channel the dominant channel and achieving eavesdropping interference in wireless communication networks. The results show that after using artificial noise, as the signal-to-noise ratio of the main channel increases from 0 to 20dB, the confidentiality capacity can increase from 0.5 to over 4.0. When the transmission power is 0.4W, the confidentiality capacity reaches its maximum and does not depend on the signal-to-noise ratio. When the number of interfering nodes increases from 1 to 2, the confidentiality capacity increases from approximately 4.7 to around 5.8. The research designed a wireless communication network eavesdropping interference scheme that can effectively protect the information security of the wireless communication network, making the main channel an advantageous channel and achieving complete confidentiality. This scheme can be applied to wireless communication networks to improve the security level of the network.**

***Keywords*—Wireless communication; sensors; network security; eavesdropping interference; clustering scenario**

I. Introduction

With the quick development and popularization of technologies such as mobile networks, the Internet of Things, and smart cities, wireless communication technology is playing an increasingly important role in social life [1]. At the same time, its broadcasting and stacking characteristics pose greater challenges for wireless networks compared to wired networks when facing security issues. The broadcasting characteristics of wireless networks mean that all devices within the wireless range can receive the transmitted data packets. Attackers can deploy their devices near legitimate recipients and intercept transmitted data by listening to wireless signals. The superposition characteristic of wireless networks allows multiple signals to overlap and transmit on the same channel. Attackers can send interference signals, reducing the quality of legitimate signals, making it difficult for legitimate receivers to correctly decode the received data, while attackers attempt to obtain information in chaos. Attackers can set up a fake access point with the same SSID and password as a legitimate wireless access point, luring users to connect to this fake network. Once connected, the attacker can intercept all data transmitted through the access point. As attackers, they usually utilize these two characteristics to reduce decoding efficiency by pretending to be illegal recipients to steal information or sending interfering information as malicious disruptors [2-3]. Therefore, how to ensure information security and reliability in wireless networks has become particularly important. Traditional wireless networks mostly monitor network traffic by deploying wireless intrusion detection systems to detect abnormal behavior or potential attack patterns in a timely manner. Meanwhile, it implements a strong authentication mechanism to ensure that only authorized users can access network resources, to reduce the risk posed by attackers. In the past, wireless network communication security was mainly built on the security framework of wired networks. Physical layer security (PLS) technology involves multiple levels such as wireless signal processing, channel coding, modulation and demodulation, and requires a deep understanding of the physical characteristics of wireless communication. The technical threshold is high. Compared to traditional security technologies based on the application layer and transport layer, research on PLS started relatively late. Therefore, related research and discussion are not sufficient, leading to the neglect of PLS issues in wireless networks at times [4]. The communication security of traditional wireless networks completely depends on key protection. In the open environment, it is difficult to use key technology to encrypt communication information, and the encryption cost is high, which is difficult to apply on a large scale. In fact, PLS is crucial for wireless network security, especially in clustering scenarios of wireless sensor networks, it can provide a new direction of thinking [5]. Currently, PLS has been widely recognized as the most effective way to solve wireless network security issues and has been applied in many cutting-edge technologies [6]. The study combines the clustering scenario of wireless sensor networks with PLS and explores the PLS problem of wireless communication network eavesdropping interference in the clustering scenario. On the basis of not changing the original topology structure of the network, this study proposes to use idle ordinary nodes as collaborative nodes.

In the inter cluster communication process of the head node, to jointly send interference information and interfere with eavesdroppers in other directions.

The main innovation of the research lies in combining clustering scenarios with PLS and proposing a strategy of utilizing idle nodes for collaborative interference. This study not only provides new research directions and possible solutions for

*Corresponding Author.

PLS, but also provides important references for achieving secure communication in clustering scenarios of wireless sensor networks. This method fully utilizes the advantages of network topology without the need for additional equipment or hardware, thus ensuring the implementation of PLS. The research provides a new perspective and solution strategy for wireless communication network security issues at the physical layer. The primary contribution of this research is the utilization of idle nodes in wireless networks to create a synergistic interference effect on the signal information transmitted within the network. This increases the difficulty for attackers to steal network information and reduces the cost of information encryption in wireless network communication, thereby significantly enhancing the security of wireless network communication.

The research will be conducted in seven sections. Section II is an overview of the current status of wireless communication network security research. Section III is a study on wireless communication network eavesdropping interference based on clustering scenarios and PLS. Section IV is an experimental analysis of eavesdropping interference schemes based on clustering scenarios and PLS. Discussion and conclusion are given in Section V and Section VI respectively.

II. Related Works

The security issue of communication networks is one of the main challenges faced by wireless communication networks. Wei Z et al. proposed a new security technology to address the security issues in integrated sensing and communication transmission, to ensure information security (Table I). By embedding information signalling in the detection waveform, the security of transmission was ensured. Meanwhile, sensing capabilities were utilized to obtain target information, which further enhances security [7]. Wang C et al. systematically investigated node capture attacks to alleviate the security issues of user authentication in wireless sensor networks. Countermeasures were proposed for different types of attacks and 61 authentication schemes were evaluated. The results showed that understanding node capture attacks helped in designing more secure authentication schemes [8]. Naghibi M et al. proposed a secure data fusion method to reduce the energy consumption of wireless sensor networks. This method reduces the number of data packets through data aggregation and improves data security by using lightweight symmetric encryption. The simulation outcomes denoted that compared with traditional methods, this method had a higher level of data security [9]. Wu F et al. proposed a new three factor authentication scheme to address the security issues of data transmission in wireless sensor networks. This scheme provided session keys, maintained security through formal verification and informal analysis, and had better security and application value than similar schemes. The simulation results indicated that this scheme had practical prospects [10]. Jia XC analyzed the current resource efficiency and security technologies adopted to meet the strict requirements of wireless sensor networks in terms of resource budgeting and security. A resource efficient distributed state estimation method and a secure distributed state estimation strategy based on different scheduling were proposed. The results showed that these technologies could effectively improve the performance and security of wireless sensor networks [11]. Hu S et al. proposed a distributed machine learning-based communication data management method to address the issues of communication congestion and information leakage caused by data explosion in wireless communication networks. The results showed that this method could effectively prevent the leakage of communication data and improve the smoothness of communication networks [12].

TABLE I. List of Literature Surveys

Author	Method	Shortcoming
Wei Z [7]	Information signaling embedding	High technical costs
Wang C [8]	Attack capture auxiliary node authentication	Complex data processing
Naghibi M [9]	Lightweight symmetric encryption	High technical costs
Wu F [10]	Key encryption	High technical costs
Jia X C [11]	Distributed State Estimation Strategy	Poor timeliness
Hu S [12]	Distributed Machine Learning	Poor timeliness
Li X [13]	PLS Backscatter Communication Network Framework	Focus on improving signal quality, with weak signal protection
Yuan X [14]	Reconfigurable only on the surface	Unable to handle signal leakage issues
Pirayesh H [15]	Frequency hopping spread spectrum	Insufficient anti-interference ability
Yu X [16]	Intelligent reflective surface	High technical costs
Matthaiou M [17]	Intelligent reflective surface	Weak signal processing ability

PLS in wireless sensor networks is one of the effective ways to solve network security. Li X et al. proposed the PLS backward scattering communication network framework to solve the challenges faced by 6G wireless communication networks. This framework aimed to improve the reliability and security of communication. The results showed that the proposed framework could optimize the performance trade-off between reliability and security under a high signal-to-noise ratio (SNR) [13]. Yuan X et al. summarized the current channel state acquisition techniques for wireless communication networks to address the issues of channel state information acquisition, passive information transmission, and low complexity robust system design for reconfigurable intelligent surfaces in wireless network PLS. The results showed that reconfigurable smart surfaces had unique advantages in improving wireless channel capacity [14]. Pirayesh H et al. aimed to comprehensively understand the interference attacks and anti-interference strategies of existing wireless networks. Various interference and anti-interference strategies for existing networks were proposed and analyzed in depth. The results showed that although some progress has been made, the design of anti-interference wireless network systems still faced challenges [15]. To enhance PLS performance, Yu X used intelligent reflective surfaces in challenging radio environments. The joint design of beamformer, covariance matrix, and phase shifter were studied to maximize system and rate while limiting information leakage. Effective algorithms were developed to solve non convex optimization problems. The results showed that intelligent reflective surfaces could significantly improve confidentiality performance, and evenly distributed reflective elements are better [16]. Matthaiou M et al. focused on the key physical layer enabling factors of 6G to meet the ubiquitous, reliable, and low latency connection needs of the future. The challenges related to intelligent reflective surfaces, large-scale multi-input and multi-output without honeycomb, and terahertz communication were proposed. The results showed that 6G would need to overcome challenges such as theoretical modeling, hardware implementation, and scalability, and signal processing played a critical role in the new era of wireless communication [17].

In summary, traditional communication anti-eavesdropping and interference technology is achieved through communication encryption. Although this method has a high communication encryption effect, it requires a large amount of computing resources and is suitable for specialized communication signal anti-eavesdropping and interference. It is not applicable in ordinary scenarios. PLS utilizes intelligent reflector surfaces to achieve anti-eavesdropping protection for communication signals, with low requirements for communication sending and receiving devices, and high applicability in IoT scenarios. However, PLS is unable to handle the attack behavior of eavesdroppers in communication data protection, and its protection capability is relatively weak. In the scenario of clustered routing, the same communication signal will be encrypted and transmitted multiple times, which can enhance the encryption strength of the communication signal. However, communication encryption will further increase the consumption of computing resources. Therefore, the study proposes to use cluster routing to improve PLS, combining the encryption enhancement effect of cluster routing with the low computational requirements of PLS encryption, to protect user communication data in the Internet of Things.

III. Eavesdropping Interference in Wireless Communication Networks Based on Clustering Scenarios and PLS

Clustering scenarios are widely used in wireless sensor networks, and PLS is a common way to protect data communication security. In Section III, the study analyzes the eavesdropping interference based on clustering scenarios and PLS wireless communication networks. Section III contains two sections: the first section is the analysis of artificial noise technology in the PLS field, and the second section is the analysis of collaborative interference based on clustering scenarios.

A. Analysis of Artificial Noise Technology in PLS Field

PLS is a new direction based on information theory, utilizing physical layer characteristics or using physical layer technology to achieve communication security? The existing communication encryption methods mostly rely on complex key encryption technology, which occupies a high amount of computing resources and requires extremely high device requirements, resulting in poor applicability in the Internet of Things. PLS technology is based on the randomness and uniqueness of wireless channels, utilizing their inherent characteristics to achieve secure transmission without relying on complex cryptographic algorithms or key distribution mechanisms [18]. PLS technology does not rely on computational complexity and can be implemented even on devices with weaker computing power. Although PLS based on the eavesdropping channel model has a necessary assumption that the quality of the main channel for legitimate communication is better than that of the eavesdropping channel, Maurer proposed a new method that used wireless channel characteristics to generate the key that information encryption relies on. At the same time, to weaken the premise assumption of eavesdropping channels, technologies such as beamforming, artificial noise, and collaborative interference were also adopted. Therefore, PLS is mainly divided into two directions: keyless security and wireless channel-based key generation. The eavesdropping channel model is shown in Fig. 1.

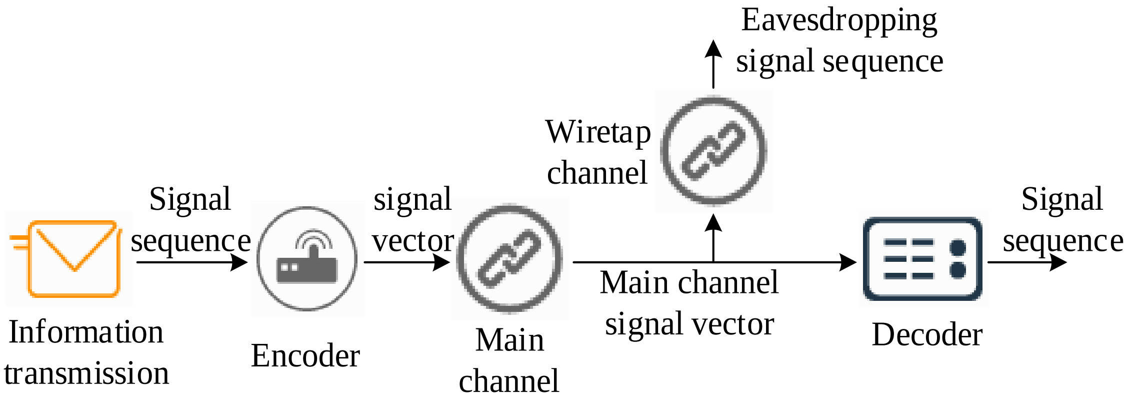


Fig. 1. The eavesdropping channel model.

The topic of the eavesdropping channel model is divided into four parts, namely encoder, main channel, eavesdropping channel, and decoder. In wireless communication systems, information transmission is a complex and constantly evolving process. There is a sequence S with a length of K, which contains communication raw data or information. Before sending, the sequence S is encoded as a vector X with a length of N, which is processed and modified by encoding before being sent through the main channel. After vector X is transmitted through the main channel, a vector Y with a length of N can be obtained. The task of the receiver is to decode the received vector Y to restore the original sequence S. Usually, decoding is a complex process that includes a series of steps such as denoising and demodulation, with the goal of restoring the sequence S as accurately as possible. However, there is a potential security risk involved in this process. In addition to legitimate recipients, there may also be a eavesdropping party attempting to intercept the information being transmitted. The eavesdropping party eavesdrops on the vector Y transmitted through the main channel through the eavesdropping channel, which is problematic and interferes with, just like the main channel. Therefore, the result that the eavesdropper overhears may be a vector Z with various noises and errors of length N. The confidentiality capacity is the main performance indicator of PLS, which is directly related to the channel capacity. In the eavesdropping channel model, when the eavesdropping channel is not considered, the channel capacity of the main channel is denoted in Eq. (1).

$$C_M = \max_{p(x)} I\left(X;Y\right) \tag{1}$$

In Eq. (1), $I\left(X;Y\right)$ means the mutual information between the information content X and Y. C_M means the channel capacity of the main channel. If considering eavesdropping channels, the capacity of the main channel can be defined as Eq. (2).

$$\hat{C}_M = \max_{p(x)} I\left(X;Y\middle|Z\right) \tag{2}$$

The eavesdropping channel can be regarded as obtaining Y through the main channel transmission, and then obtaining Z through the eavesdropping channel. The virtual channel is called, and at this point, a Markov chain is formed between the sequences X, Y, and Z. Based on the Markov chain, Eq. (3) can be obtained.

$$H\left(X\middle|Y,Z\right) = H\left(X\middle|Y\right) \tag{3}$$

In Eq. (3), $H\left(\cdot\middle|\cdot\right)$ represents conditional entropy, which represents the amount of information that X still contains, given all the information of Y. Through the basic properties of mutual information, equation (4) can be obtained.

$$I\left(X;Y\middle|Z\right) = H\left(X\middle|Z\right) - H\left(X\middle|Y,Z\right) = I\left(X;Y\right) - H\left(X;Z\right) \tag{4}$$

According to Eq. (4), Eq. (2) can be rewritten as Eq. (5).

$$\hat{C}_M = \max_{p(x)} \left[I\left(X;Y\right) - I\left(X;Z\right)\right] \tag{5}$$

In the eavesdropping channel model, it is assumed that the sender wants to send a message while also protecting the message from eavesdropping. To achieve this goal and obtain the maximum actual information transmission rate, the confidentiality capacity can be defined as Eq. (6).

$$C_s = \max_{R\in\Re} R \tag{6}$$

In Eq. (6), R represents the actual transmission rate. C_s represents confidential capacity. $\Re$ represents the reachable region of $\left(R,R_e\right)$, as shown in Fig. 2 [19].

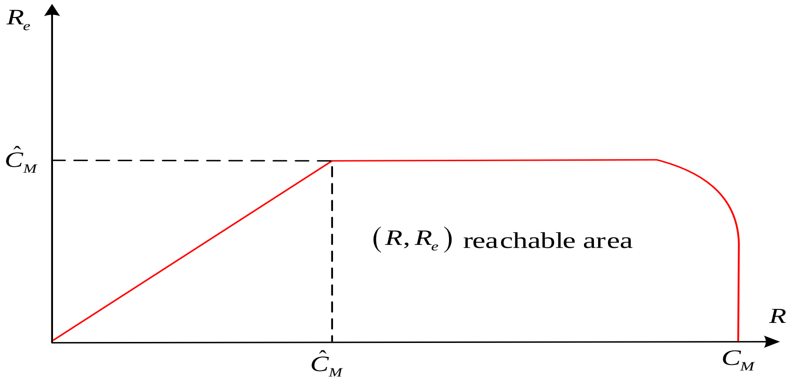


Fig. 2. $\left(R,R_e\right)$ range coverage.

In Fig. 2, the range of R is less than C_M, the range of R_e is less than $\hat{C}_M$, and in $R \le \hat{C}_M$, $R = R_e$ meets the requirements for complete confidentiality. Therefore, when the confidentiality capacity satisfies Eq. (7), communication at a rate of V with C_s as the upper bound can achieve complete confidentiality.

$$C_s = \hat{C}_M = \max_{p(x)} \left[I(X;Y) - I(X;Z) \right] \tag{7}$$

In the eavesdropping channel model, signal receiving channels are divided into legitimate receiving channels and illegal receiving channels. When the signal generator transmits a signal, it will add an error code to the transmitted signal. The frequency of the error code is different from the frequency of the legal channel. When the legal channel receives the signal, it will not receive the error code. The signal reception frequency of the eavesdropping channel is relatively wide, and it will receive a large number of signals when receiving signals. At the same time, it will also receive error codes in the signal, resulting in missing or incorrect received information. This makes the effective information in the signal received by the eavesdropper 0, which can achieve complete confidentiality of information. That is, in the eavesdropping channel model, when the quality of the main channel is higher than that of the eavesdropping channel, non-zero confidentiality ability can be obtained. To ensure that the quality of the main channel is higher than that of the eavesdropping channel, artificial noise can be added to the information to interfere with the eavesdropping channel and reduce its quality, thereby achieving complete confidentiality of the information. The artificial noise scheme design is shown in Fig. 3.

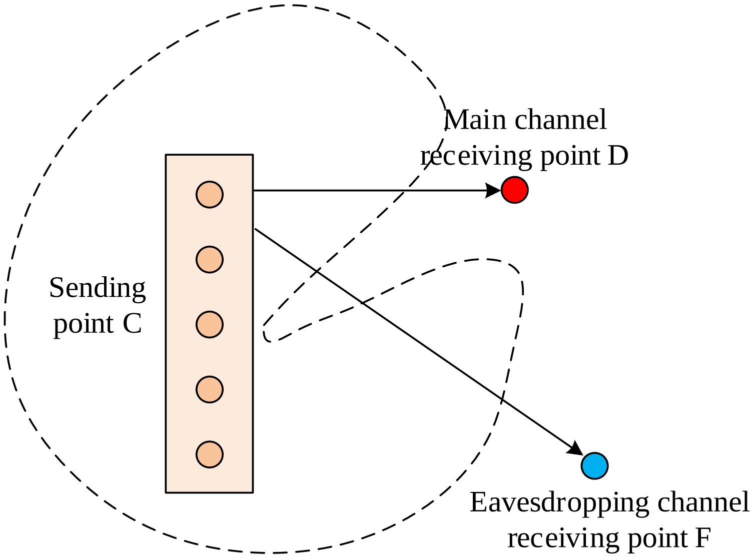


Fig. 3. Artificial noise confidentiality scheme.

In the research-designed artificial noise interference schemes, there is an artificially designed noise between the signal sender and receiver. When using this noise to interfere with eavesdroppers, it is necessary to use a legitimate channel to provide feedback on the channel status to the signal sender. After receiving channel status information, the signal sender sends zero space artificial noise to the legitimate channel based on the channel status. The legitimate receiving channel can process the received signal and extract information based on the zero space artificial noise. Due to the inability of illegal channels to provide feedback on channel status to the sender, zero space artificial noise cannot be obtained. Artificial noise between communication networks can interfere with the transmitted signal, organizing eavesdroppers to read signal information. In this scheme, the received signal z_k of the main channel can be represented as Eq. (8).

$$z_k = h_k x_k + n_k \tag{8}$$

In Eq. (8), h_k represents the main channel vector. x_k stands for sending signals. n_k represents the Gaussian white noise vector of the main channel. The received signal y_k of the eavesdropping channel can be expressed as Eq. (9).

$$y_k = g_k x_k + e_k \tag{9}$$

In Eq. (9), g_k represents the eavesdropping channel vector. e_k represents the Gaussian white noise vector of the eavesdropping channel. When the information sender is equipped with multiple antennas, the transmitted signal can be represented as Eq. (10).

$$x_k = p_k u_k + w_k \tag{10}$$

In Eq. (10), u_k represents the carrier signal. w_k represents artificial noise signal. p_k represents the beam wave vector. Artificial noise signals need to meet Eq. (11).

$$\begin{cases} h_k w_k = 0 \\ h_k p_k \neq 0 \end{cases} \tag{11}$$

At this point, the received signal at the main channel can be represented as Eq. (12).

$$z_k = h_k p_k u_k + n_k \tag{12}$$

The received signal at the eavesdropping channel can be expressed as Eq. (13).

$$y_k = g_k p_k u_k + g_k w_k + e_k \tag{13}$$

In order to make the impact of artificial noise on the eavesdropping channel equal in each direction, it can be designed as Eq. (14).

$$w_k = \Gamma_k v_k \tag{14}$$

In Eq. (14), Γ_k represents the zero space matrix of h_k, and v_k represents an independent and identically distributed Gaussian vector with a mean of 0.

B. *Collaborative Interference Analysis Based on Clustering Scenarios*

Clustering scenario is a commonly used network architecture model in wireless sensor networks. This scheme divides each node in the network into different clusters, with each cluster having a node called a cluster head responsible for managing and organizing communication within the cluster [20-21]. Within each cluster, cluster head nodes manage and control member nodes, collect and summarize data from nodes within the cluster, and then transmit this data to base stations or other cluster head nodes. Clustering can adapt to the dynamic joining and exiting of nodes, and cluster heads can reorganize the cluster structure based on changes in nodes within the cluster, maintaining network connectivity and stability. Moreover, each cluster can be independently managed, while the cluster head can be responsible for monitoring the status of nodes within the cluster and maintaining the stability of the cluster. If some nodes within a cluster fail, the cluster head can reorganize the remaining nodes or collaborate with other clusters to ensure the continuity of network services. In a clustered cluster, data transmission first occurs within the cluster and then communicates with other clusters or base stations through the cluster head. In a clustered cluster, the cluster head can aggregate data within the cluster, reduce redundant data transmission, and improve data transmission efficiency. This approach reduces the need for direct communication between each node and the base station and lowers the communication overhead of the entire network. During the communication process, the cluster head is responsible for collecting and transmitting data within the cluster, while other nodes reduce energy consumption due to reduced direct communication with base stations. The selection of cluster heads can be based on the energy level of nodes, thereby achieving balanced energy consumption and extending the service life of the network. This method can significantly reduce the number of communications between nodes, thereby significantly reducing energy consumption and improving the network's lifespan. In the process of dividing clusters, different goals and standards can be used, such as geographical location, energy consumption, network conditions, and so on. Due to the fact that the nodes in each cluster are usually located in similar areas, the communication distance between nodes is short, which is beneficial for energy conservation and improving network performance. Cluster routing is shown in Fig. 4.

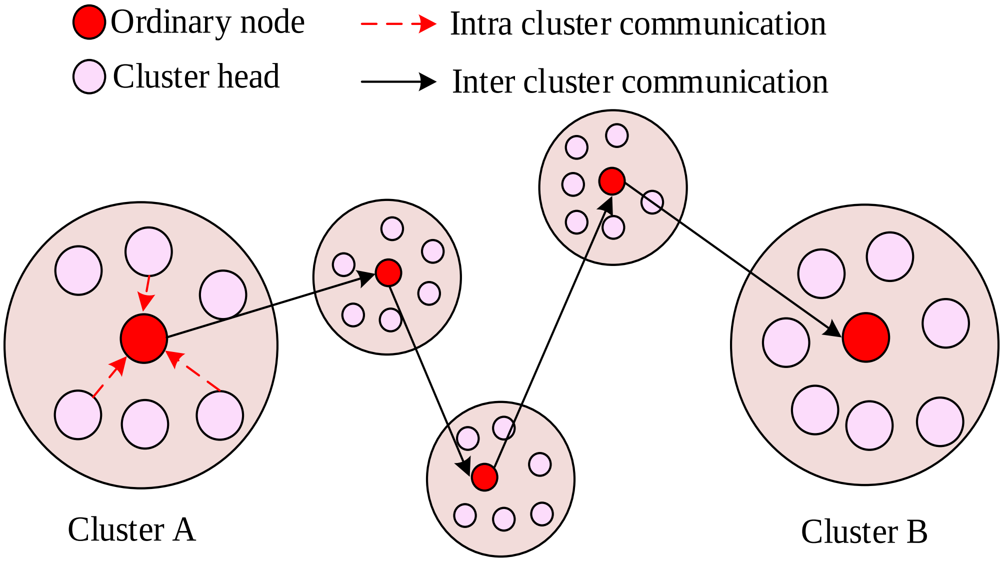


Fig. 4. Clustering routing.

Clustered wireless sensor networks provide a natural scenario for achieving collaborative interference. Firstly, in a clustering network, when cluster head nodes engage in inter cluster communication, ordinary nodes usually do not participate in intra cluster communication to avoid interference between intra cluster nodes. In this process, these temporarily idle ordinary nodes can be selected as interference nodes, jointly generating interference effects and obstructing potential eavesdroppers from obtaining information. Idle nodes within the cluster do not participate in intra cluster communication and can also accept artificial noise outside the cluster. After receiving artificial noise, idle nodes attach it to the communication space within the cluster, and attach it to the periphery of the communication nodes within the cluster, forming an interference protection layer for passing nodes. Secondly, the clustering scenario makes information exchange much more convenient. Cluster head nodes and regular nodes are relatively concentrated in physical locations and close in distance, making information exchange within a small area more convenient and efficient compared to the entire network. Cluster heads can better coordinate with ordinary nodes to generate collaborative interference, while monitoring the effectiveness of interference. In addition, the clustering scenario itself is a network management mechanism with good topology and organization. This structure provides convenience for designing and implementing collaborative interference strategies, making it easier to achieve collaborative actions between the sender and interference nodes. The two-stage collaborative interference scheme is a wireless communication strategy based on PLS. The core of this scheme is to use nodes in the network to collaborate in two stages to generate interference and improve communication security, as shown in Fig. 5.

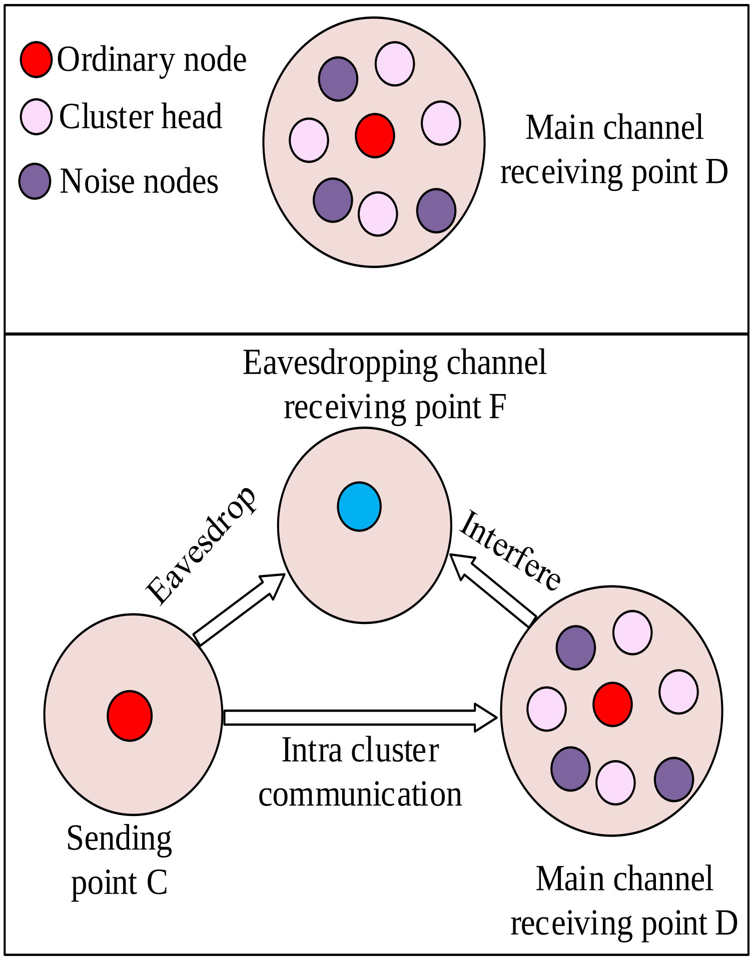

Ordinary node
Cluster head
Noise nodes
Main channel receiving point D
Eavesdropping channel receiving point F
Eavesdrop
Interfere
Intra cluster communication
Sending point C
Main channel receiving point D

Fig. 5. Two-stage collaborative interference scheme.

The collaborative interference scheme based on clustering scenarios designed for research is based on this framework, which organizes sensor nodes in the network into multiple clusters based on specific algorithms or strategies. In each cluster, a cluster head node serves as the center to manage and command several child nodes to complete their respective tasks. The research assumes that the sender of inter cluster communication is C, the receiver is D, and the illegal eavesdropper is marked as F. C and D are cluster head nodes in two different clusters A and B. Their task is not only to collect data within their respective clusters, but also to transmit information between clusters. At the same time, illegal eavesdropper F is lurking in cluster E, attempting to intercept communication content between C and D through wireless eavesdropping channels. In order to address the potential threat of F, the plan proposes that D will select a portion of ordinary nodes belonging to cluster B under its management as interference nodes, as shown in Fig. 6.

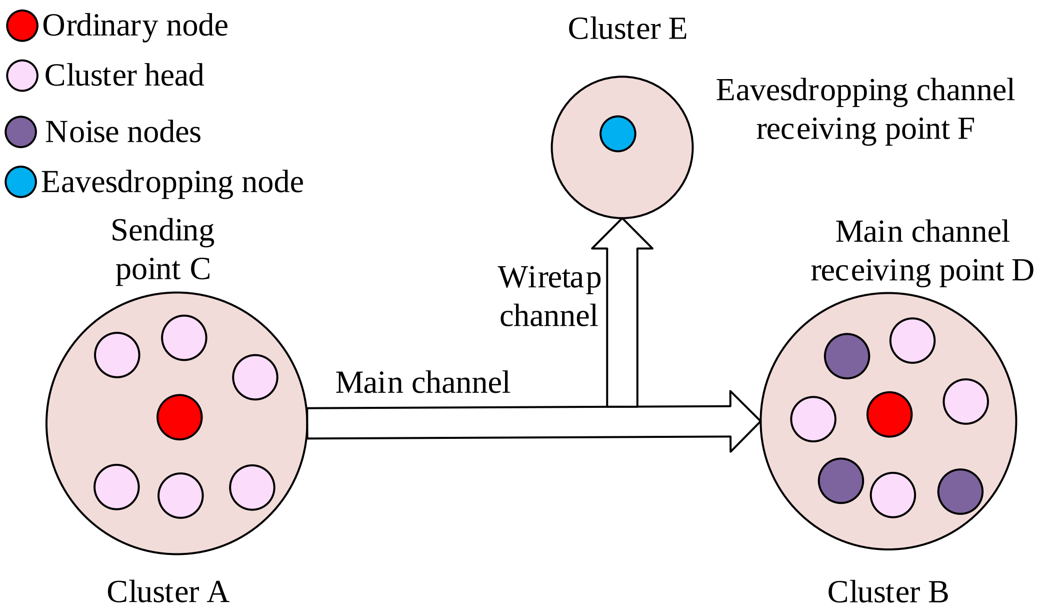


Fig. 6. Topology structure of collaborative interference network based on clustering scenarios.

Interference nodes are dynamically updated and can interfere with F's eavesdropping behavior, thereby protecting the security of information transmission. Due to the dynamic changes of nodes in clustering, it is not possible to select a constant node as the interfering node every time. Therefore, this study designs a strategy for selecting interfering nodes through the principle of reservoir sampling algorithm. Compared with traditional methods such as dynamic frequency hopping, the research-designed method not only protects wireless network communication security through PLS, but also considers the anti-interference ability of the communication channel. The study-designed wireless network communication security protection technology adopts cluster scenario design, which has good adaptability to network environment. Due to the use of PLS technology in the research-designed communication security protection methods, there may be compatibility issues with some systems or equipment. It is necessary to update or design supporting frameworks, develop compatibility layers, and improve the applicability of this technology.

IV. Analysis of Wireless Communication Eavesdropping Interference Schemes Based on Clustering Scenarios and PLS

In Section III, a wireless communication eavesdropping interference scheme based on clustering scenarios and PLS was proposed. To verify the feasibility of this scheme, simulation experiments were conducted in Section IV to analyze it. This part is divided into two sections. The first section is the setting of simulation experiment parameters and environment, and the second section is the analysis of the effectiveness of eavesdropping interference schemes.

A. *Experimental Parameters and Environmental Settings*

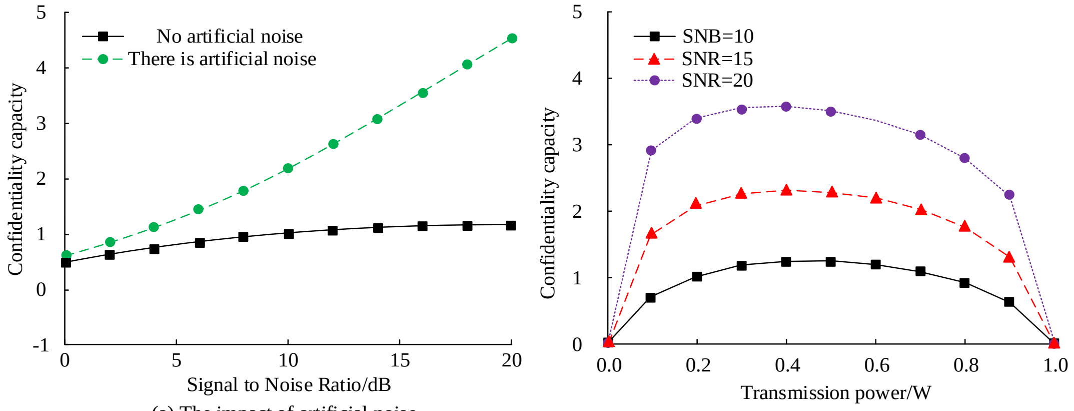
The system used in the study was Windows 10 64 bit, and the device processor was Inter (R) Core (TH) i5-12440. The device had 16GB of memory and the simulation experiment platform was MATLAB. The relevant parameters of the model are denoted in Table II.

TABLE II. MODEL PARAMETER

Parameter	Value	Unit	Parameter	Value	Unit
Total power	1	W	Cluster A radius	5	m
Antenna gain parameters	0.003	/	Cluster B radius	5	m
Main channel path loss index	2	/	Distance between the centers of cluster A and cluster B	15	m
Interference channel path loss index	2.5	/	Minimum node spacing	2	m
Communication bandwidth	5	MHz	Interference node density	0.05	Pieces/m2
Gaussian white noise unilateral spectral density	-174	dB	Eavesdropping node density	0.001	Pieces/m2

B. Analysis of the Effectiveness of Eavesdropping Interference Schemes

To evaluate the effectiveness of the proposed wireless communication network eavesdropping collaborative interference method, the study analyzed the variation of confidentiality capacity with the main channel SNR in the presence or absence of artificial noise, as well as the variation of confidentiality capacity with transmission power under different SNRs. The results are shown in Fig. 7.

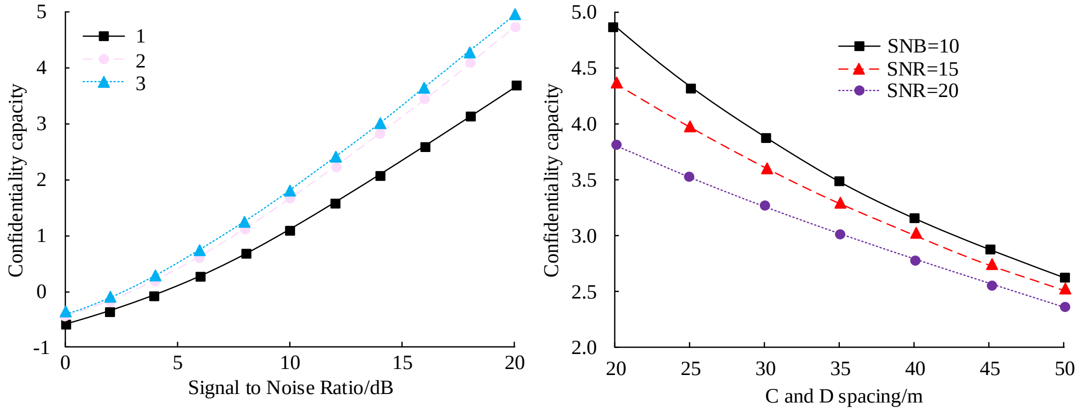


(a) The impact of artificial noise on confidentiality capacity

(b) The impact of transmission power on confidentiality capacity

Fig. 7. The recy capacity with the main channel parameter.

Fig. 7 (a) shows the change of confidentiality capacity with and without artificial noise as a function of the main channel SNR. The system confidentiality capacity remained between 0 and 1 without the addition of artificial noise. After adding artificial noise, the system's confidentiality capacity increased rapidly with the increase of the main channel SNR. When the main channel SNR was 0, the system's confidentiality capacity was about 0.5. When the main channel SNR increased to 20dB, the system's confidentiality capacity increased to above 4.0. Fig. 7 (b) shows the variation of system confidentiality capacity with transmission power under different SNRs. Regardless of the SNR, when the transmission power was 0.4W, the system's confidentiality capacity always reached its maximum value. The study analyzed the relationship between system confidentiality capacity and interfering nodes, as shown in Fig. 8.

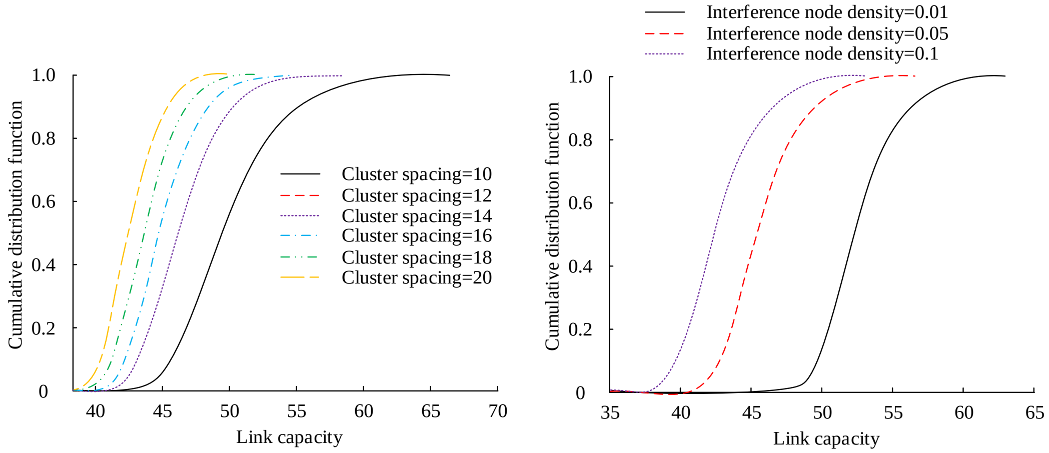


(a) The impact of artificial noise on confidentiality capacity

(b) The variation of confidentiality capacity with the distance between C and D

Fig. 8. Association between system confidentiality capacity and interference nodes.

Fig. 8 (a) shows the variation of system confidentiality capacity with SNR under different numbers of interfering nodes. When the number of interfering nodes was 1, the maximum confidentiality capacity of the system was about 4.7. After increasing the number of interfering nodes by one, the maximum confidentiality capacity of the system was about 5.8. Fig. 8 (b) shows the variation of system confidentiality capacity with the distance between C and D under different interfering nodes and F distances. When the distance between interfering nodes and F was fixed, the system's confidentiality capacity would increase with the increase of the distance between C and D. When the distance between nodes C and D was fixed, the system's confidentiality capacity would decrease as the distance between interfering nodes and F increased. Link capacity is a basic indicator for inter-cluster interference analysis. To assess the effectiveness of the proposed wireless communication eavesdropping interference scheme combining clustering scenarios and PLS, the variation of link capacity with cumulative distribution function was studied and analyzed under different cluster spacing and interference node density. The results are shown in Fig. 9.

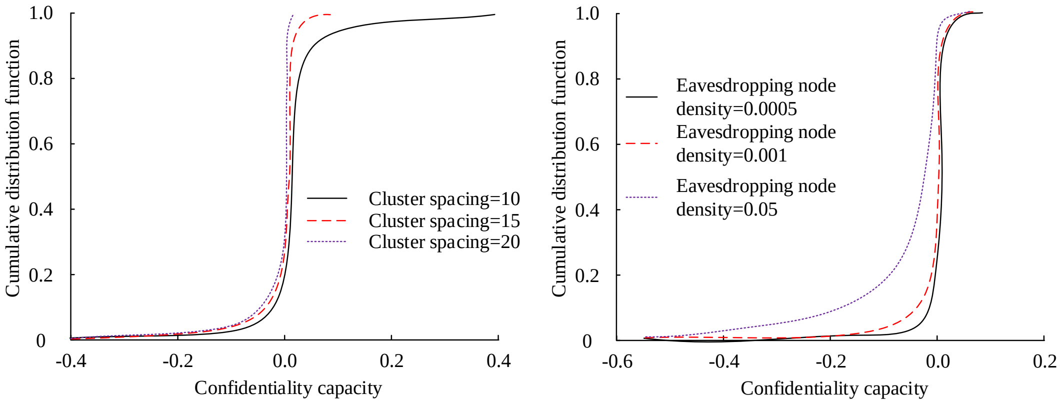


(a) Link capacity distribution under different cluster spacing

(b) The Impact of Interference Node Density on Link Capacity

Fig. 9. Effect of cluster spacing and interference node density on link capacity.

Fig. 9 (a) shows the distribution of link capacity under different cluster spacing. As can be seen, the cumulative distribution function value was 0.6, and when the cluster spacing was 10, the link capacity was about 50. When the cluster spacing was 20, the link capacity was about 30. As the cluster spacing increased, the overall link capacity decreased. As the distance between clusters increased, the loss of the signal during propagation also increased, resulting in a decrease in the energy received by the receiver, thereby reducing the overall communication performance. Fig. 9 (b) shows the distribution of link capacity under different interference node densities. As can be seen, when the cumulative distribution function value was 0.6 and the interference node density was 0.01Pieces/m2, the system link capacity was 54. When the interference node density was 0.1Pieces/m2, the system link capacity was 42. The link capacity would decrease as the density of interfering nodes increased, leading to a decrease in system communication performance. The study analyzed the variation of confidentiality capacity with the cumulative distribution function under different cluster spacing and eavesdropping node densities, as indicated in Fig. 10.

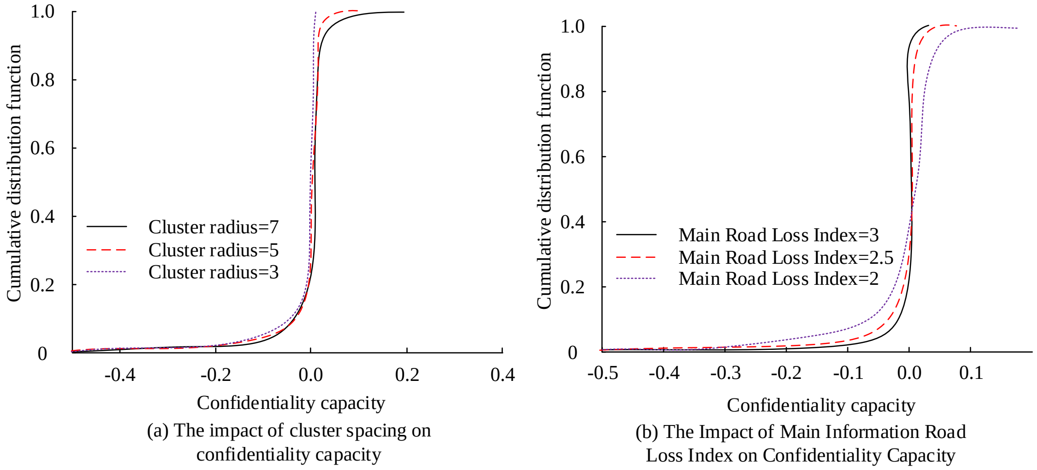


(a) The impact of cluster spacing on confidentiality capacity

(b) The impact of eavesdropping node density on confidentiality capacity

Fig. 10. Effect of cluster spacing and eavesdropping node density on confidentiality capacity.

Fig. 10 (a) shows the distribution of confidentiality capacity under different cluster spacing. As the distance between clusters increased, the signal from sender C to receiver D must be transmitted over a longer distance, which inevitably led to signal attenuation. Signal attenuation means a decrease in signal strength at receiver D, which requires a higher SNR to ensure the same communication quality. However, in actual wireless network environments, it is often difficult to compensate for signal attenuation caused by distance increase due to factors such as system transmission power and environmental noise. Fig. 10 (b) shows the distribution of confidentiality capacity under different eavesdropping node densities. As the density of eavesdropping nodes decreased, the confidentiality capacity of the system was also constantly increasing. The lower the density of eavesdropping nodes and the higher the confidentiality capacity, the better the communication performance of the system. Finally, the study also analyzed the impact of clustering radius, path loss index on the main channel, and confidentiality capacity, as denoted in Fig. 11.



(a) The impact of cluster spacing on confidentiality capacity

(b) The Impact of Main Information Road Loss Index on Confidentiality Capacity

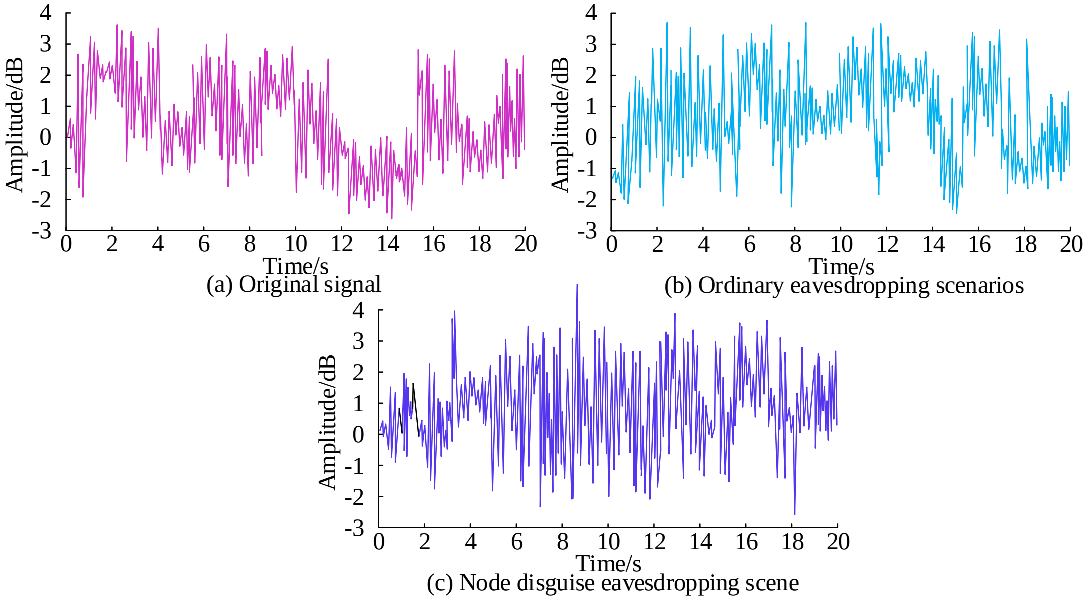
Fig. 11. Effect of branch cluster radius and main channel path loss index on secrecy capacity.

Fig. 11 (a) shows the distribution of confidentiality capacity under different clustering radii. The clustering radius had a relatively small impact on the system's confidentiality capacity. Fig. 11 (b) shows the distribution of confidentiality capacity under different main channel path loss indices. As the main channel path loss index decreased, the confidentiality capacity showed a slight increase trend. To further verify the influence of different parameter settings on confidentiality capacity, the study performed ANOVA on main channel SNR, transmission power and number of interfering nodes, and the results are shown in Table III.

TABLE III. ANALYSIS OF VARIANCE OF MAIN CHANNEL SNR, TRANSMISSION POWER AND INTERFERENCE NODES

Factor	Df1	Df2	F	P
Main channel SNR	3	30	25.6	<0.001
Transmission power	2	20	13.4	<0.001
Number of interfering nodes	4	35	8.9	<0.001

ANO analysis of variance showed that the main channel SNR (*F* (3,30) =25.6, *P* <0.001), transmission power (*F* (2,20) =13.4, *P* <0.001) and the number of interfering nodes (*F* (4,35) =8.9, *P* <0.001) significantly influenced the confidentiality capacity. To further study the eavesdropping interference effect of the design method, the results was compared with the node camouflage eavesdropping scenario, and the results are shown in Fig. 12.



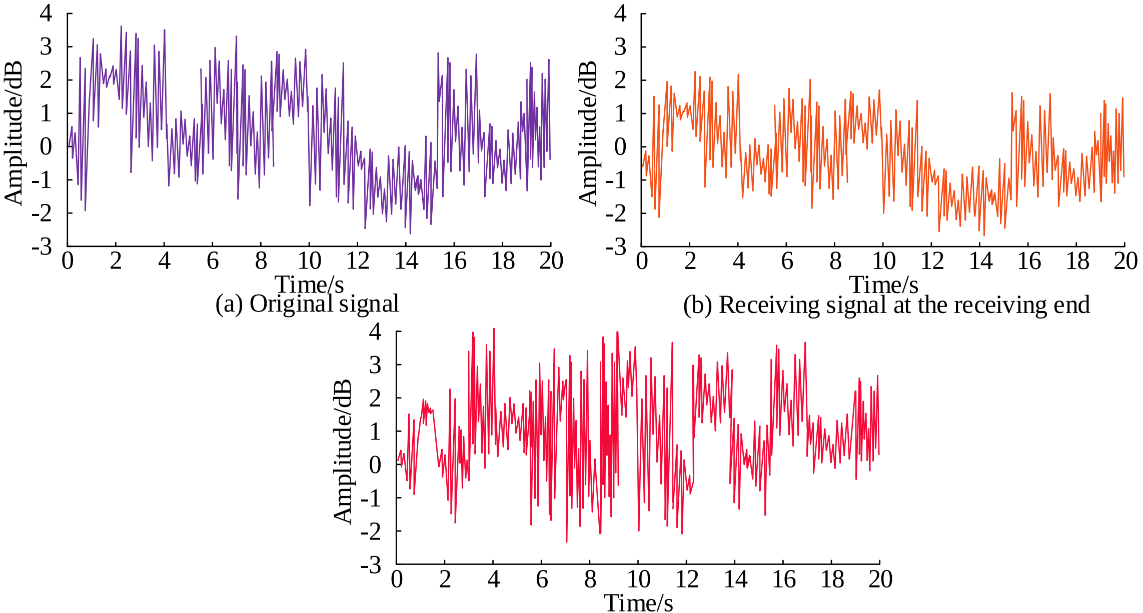
(a) Original signal

(b) Ordinary eavesdropping scenarios

(c) Node disguise eavesdropping scene

Fig. 12. The eavesdropping and interference effects in different scenarios.

Fig. 12 (a) shows the original signal at the transmitting end, Fig. 12 (b) shows the received signal of the eavesdropping channel in a normal eavesdropping scenario, and Fig. 12 (c) shows the received signal of the eavesdropping channel in a node disguised eavesdropping scenario. Regardless of the eavesdropping interference in any scenario, the designed eavesdropping interference scheme could always effectively conceal the true information of the transmitting signal. To further verify the practical application of the wireless communication network, the quality of the signal received by the receiver and the effective information in the eavesdropping information were compared. The results are shown in Fig. 13.



(a) Original signal

(b) Receiving signal at the receiving end

(c) Eavesdroppers receive signals

Fig. 13. The of interference under artificial noise.

Fig. 13 (a) is the original signal, Fig. 13 (b) is the receiving signal of the receiver, and Fig. 13 (c) is the receiving signal of the eavesdropping party. It can be seen that the waveform of the received signal at the receiving end was basically the same as the original signal waveform, but the signal amplitude was weakened, and the overall information was kept intact. The waveform of the signal received by the eavesdropper was quite different from the original signal, and only the individual bands overlapped somewhat, and the original signal information was basically not reserved. Under the interference of artificial noise, the receiving end could still receive the original signal well, and effectively avoid the risk of information leakage. When deploying the artificial noise, if the noise signal characteristics were similar to the emission signal characteristics, the eavesdropping interference effect would be poor. The artificial noise remained unchanged for a long time, which would also lead to the poor eavesdropping interference effect. Therefore, when deploying artificial noise, it is necessary to adjust the artificial noise signal at different time to improve the eavesdropping interference effect of artificial noise. To further verify the effectiveness of network eavesdropping and interference technology based on clustering scenarios and PLS, an experiment was designed to compare the anti-eavesdropping effect of this method with the artificial noise assisted secure wireless communication technology proposed in study [19]. The results are shown in Table IV.

In Table IV, when using the research-designed anti-eavesdropping method, the effective signal proportion in the signal received by the eavesdropper did not exceed 1%. However, when using the method proposed in study [19], the effective signal proportion in the signal received by the eavesdropper could reach up to 2.67%. Research on anti-eavesdropping technology can better protect privacy and security.

TABLE IV. COMPARISON OF THE ANTI-EAVESDROPPING EFFECT

Signal number	Proportion of effective information (%)	
	Proposed method	Reference [19]
1	0.56	1.35
2	0.48	2.67
3	0.95	1.94
4	0.67	2.54
5	0.89	2.13

V. DISCUSSION

A collaborative eavesdropping interference scheme based on PLS and clustering scenarios was proposed to improve the communication security of wireless communication networks. By adding artificial noise to the transmitted signal, eavesdropping signals could be successfully interfered with, enhancing the advantages of the main channel and achieving eavesdropping interference in wireless communication networks. The research results indicated that as the SNR of the main channel increased, the confidentiality capacity of the system significantly improved, increasing from 0.5 to over 4.0. In addition, as the number of interfering nodes increased, the confidentiality capacity also increased, similar to the research results of Pang X et al. [21]. The proposed method not only provided new research directions and possible solutions for PLS, but also provided important reference value for clustering scenarios in wireless sensor networks. By fully utilizing the advantages of network topology without the need for additional equipment or hardware, PLS implementation was ensured, providing a new perspective and solution strategy for wireless communication network security issues. The different channel models and environmental noise considered in the experiment may not cover all the complex details in practical scenarios, and future research needs to be further deepened. For example, it is possible to explore universal interference strategies that can adapt to various channel conditions and noisy environments, and optimize the deployment methods of interfering nodes. In the future, further research and optimization will be conducted on the generation mechanism of artificial noise to adapt to the dynamically changing network environment. Secondly, it should explore interference strategies under different channel conditions to improve the adaptability and robustness of the system. Finally, considering the compatibility issues in actual deployment, it will investigate how to seamlessly integrate PLS technology with existing wireless communication networks.

VI. Conclusion

To improve the communication security of wireless sensor networks, a research proposal was proposed to combine PLS with cluster routing. A collaborative eavesdropping interference scheme combining PLS and cluster scenarios was designed to protect the communication security of wireless sensor networks. This scheme added artificial noise to the transmitted information to achieve eavesdropping interference. The results showed that in an environment without artificial noise, the confidentiality capacity ranged from 0 to 1. After using artificial noise, as the SNR of the main channel increased from 0 to 20dB, the confidentiality capacity could increase from 0.5 to over 4.0. When the cumulative distribution function value was 0.6, the cluster spacing increased from 10 to 20, and the link capacity decreased from 50 to 30. The density of interfering nodes was increased from 0.01Pieces/m² to 0.1Pieces/m². At that time, the link capacity decreased from 54 to 42. When the path loss index decreased, the confidentiality capacity slightly increased, while the clustering radius had little effect on the confidentiality capacity. The addition of artificial noise effectively improved the confidentiality capacity of the system, which was highly sensitive to the SNR of the main channel and achieved the optimal confidentiality capacity at a certain transmission power level. The increase in the number of interfering nodes could significantly affect the confidentiality capacity of the system, but this effect tended to saturate after the number of nodes reached a certain threshold. In terms of spatial distribution, an increase in cluster spacing and interference node density could lead to a decrease in link capacity, thereby affecting communication performance. When conducting simulation analysis on eavesdropping interference schemes, the consideration of different channel models and environmental noise was insufficient to cover all the complex details in actual scenarios, and further deepening is needed. In the future, it can explore universal interference strategies that can adapt to various channel conditions and noisy environments, and optimize the deployment methods of interference nodes. The use of PLS and cluster routing to achieve eavesdropping interference in communication signals greatly protects user personal privacy and improves wireless network communication security, which can effectively promote the development of the Internet of Things.

Acknowledgment

The research is supported by Education Science 14th Five-Year Plan Project of Fujian Province, Research on the construction of training system of digital literacy education for college students (No. Fjxczx23-302).

References

[1] Huang C, Hu S, Alexandropoulos G C, Zappone A, Yuen C, Zhang R, Debbah M. Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends. IEEE Wireless Communications, 2020, 27(5): 118-125.

[2] Arfaoui M A, Soltani M D, Tavakkolnia I, Ghrayeb A, Safari M, Assi C M, Haas H. Physical layer security for visible light communication systems: A survey. IEEE Communications Surveys & Tutorials, 2020, 22(3): 1887-1908.

[3] Pirayesh H, Zeng H. Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2022, 24(2): 767-809.

[4] Fang S, Chen G, Li Y. Joint optimization for secure intelligent reflecting surface assisted UAV networks. IEEE Wireless Communications Letters, 2020, 10(2): 276-280.

[5] Wang C X, Di Renzo M, Stanczak S, Wang S, Larsson E G. Artificial intelligence enabled wireless networking for 5G and beyond: Recent advances and future challenges. IEEE Wireless Communications, 2020, 27(1): 16-23.

[6] Du J, Jiang C, Wang J, Ren Y, Debbah M. Machine learning for 6G wireless networks: Carrying forward enhanced bandwidth, massive access, and ultrareliable/low-latency service. IEEE Vehicular Technology Magazine, 2020, 15(4): 122-134.

[7] Wei Z, Liu F, Masouros C, Su N, Petropulu A P. Toward multi-functional 6G wireless networks: Integrating sensing, communication, and security. IEEE Communications Magazine, 2022, 60(4): 65-71.

[8] Wang C, Wang D, Tu Y, Xu G, Wang H. Understanding node capture attacks in user authentication schemes for wireless sensor networks. IEEE Transactions on Dependable and Secure Computing, 2020, 19(1): 507-523.

[9] Naghibi M, Barati H. SHSDA: secure hybrid structure data aggregation method in wireless sensor networks. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(12): 10769-10788.

[10] Wu F, Li X, Xu L, Vijayakumar P, Kumar N. A novel three-factor authentication protocol for wireless sensor networks with IoT notion. IEEE Systems Journal, 2020, 15(1): 1120-1129.

[11] Jia X C. Resource-efficient and secure distributed state estimation over wireless sensor networks: A survey. International Journal of Systems Science, 2021, 52(16): 3368-3389.

[12] Hu S, Chen X, Ni W, Hossain E, Wang X. Distributed machine learning for wireless communication networks: Techniques, architectures, and applications. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1458-1493.

[13] Li X, Zheng Y, Khan W U, Zeng M, Li D, Ragesh G K, Li L. Physical layer security of cognitive ambient backscatter communications for green Internet-of-Things. IEEE Transactions on Green Communications and Networking, 2021, 5(3): 1066-1076.

[14] Yuan X, Zhang Y J A, Shi Y, Yan W, Liu H. Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities. IEEE Wireless Communications, 2021, 28(2): 136-143.

[15] Pirayesh H, Zeng H. Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2022, 24(2): 767-809.

[16] Yu X, Xu D, Sun Y, Ng D W K, Schober R. Robust and secure wireless communications via intelligent reflecting surfaces. IEEE Journal on Selected Areas in Communications, 2020, 38(11): 2637-2652.

[17] Matthaiou M, Yurduseven O, Ngo H Q, Morales-Jimenez D, Cotton S L, Fusco V F. The road to 6G: Ten physical layer challenges for communications engineers. IEEE Communications Magazine, 2021, 59(1): 64-69.

[18] Polese M, Jornet J M, Melodia T, Zorzi M. Toward end-to-end, full-stack 6G terahertz networks. IEEE Communications Magazine, 2020, 58(11): 48-54.

[19] Hong S, Pan C, Ren H, Wang K, Nallanathan A. Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface. IEEE Transactions on Communications, 2020, 68(12): 7851-7866.

[20] Bandewad G, Datta K P, Gawali B W, Pawar S N. Review on Discrimination of Hazardous Gases by Smart Sensing Technology. Artificial Intelligence and Applications, 2023, 1(2): 86-97.

[21] Pang X, Sheng M, Zhao N, Tang J, Niyato D, Wong K K. When UAV meets IRS: Expanding air-ground networks via passive reflection. IEEE Wireless Communications, 2021, 28(5): 164-170.

Development of a Hybrid Quantum Key Distribution Concept for Multi-User Networks

Begimbayeva Y[1], Zhaxalykov T[2]*, Makarov M[3], Ussatova O[4], Tynymbayev S[5], Temirbekova Zh[6]
KazNRTU Named after K. I. Satbayev, Almaty, Kazakhstan[1, 2, 3, 4]
Kazakh British Technical University, Almaty, Kazakhstan[1, 2, 3]
Department of Cybersecurity, Energo University, Almaty, Kazakhstan[1, 4]
Faculty Information Technology, Kazakh National University Named After Al-Farabi (KazNU), Almaty, Kazakhstan[5]
Faculty of Computer Technology and Cybersecurity, International IT University (IITU), Almaty, Kazakhstan[6]

***Abstract*—This paper investigates the increasing concerns related to the vulnerability of contemporary security solutions in the face of quantum-based attacks, which pose significant challenges to existing cryptographic methods. Most current Quantum Key Distribution (QKD) protocols are designed with a focus on point-to-point communication, limiting their application in broader network environments where multiple users need to exchange information securely. To address this limitation, a thorough analysis of twin-field-based algorithms is conducted, emphasizing their distinct characteristics and evaluating their performance in practical scenarios in Sections II, III, and IV. By synthesizing insights from these analyses, integrating cutting-edge advancements in Quantum Communication technologies, and drawing on proven methodologies from established point-to-point protocols, this study introduces a novel concept for a Hybrid Twin-Field QKD protocol in Section IV. This network-oriented approach is designed to facilitate secure communication in networks involving multiple users, offering a practical and scalable solution. The proposed protocol aims to reduce resource consumption while maintaining high-security standards, thereby making it a viable option for real-world quantum communication networks. This work contributes to the development of more resilient and efficient quantum networks capable of withstanding future quantum-based threats.**

Keywords*—Quantum key distribution; quantum communication; multi-user networks; network security; quantum-based attacks; cryptography; point-to-point protocols; resource efficiency; cryptography; information security

I. Introduction

The increasing concern regarding the physical vulnerability of fiber networks has become a significant issue, as traditional security mechanisms are increasingly bypassed by sophisticated attackers. This escalating threat underscores the necessity for the development of innovative quantum-based security solutions. Notably, the global metric for the 'Estimated Cost of Cybercrime' within the cybersecurity sector is projected to rise steadily from 2023 to 2028, with an anticipated increase of 5.7 trillion U.S. dollars, representing a 69.94% growth. By 2028, after eleven consecutive years of growth, this figure is expected to reach a new high of 13.82 trillion U.S. dollars [1], emphasizing the urgent need for advanced cybersecurity measures. Furthermore, the ongoing advancements and strategic roadmaps of technology leaders, such as IBM [2], suggest rapid developments in the computational power of quantum computers, posing significant threats to existing secure communication protocols like RSA [3] and AES [4]. The widespread reliance on these algorithms, particularly among critical businesses essential to the functioning of foundational societal ecosystems, exacerbates the risk posed by emerging quantum threats.

Recent years have seen significant progress in the field of cryptography, with researchers exploring new mathematical foundations and encryption techniques to enhance security [5] [6] [7] [8] [9] [10].

To mitigate these risks, the implementation of quantum cryptography offers a promising solution. Quantum cryptography provides secure communication channels that are resilient to both classical and quantum attacks, leveraging two fundamental principles of quantum mechanics: Quantum Entanglement, which enables the encoding and sharing of information across vast distances while monitoring for any unauthorized interference, and the No-Cloning Theorem, which ensures protection against potential eavesdroppers attempting to replicate unique quantum states. One effective method for achieving such security is through Quantum Key Distribution (QKD) protocols, which facilitate the secure generation and distribution of secret keys among communication participants.

However, the majority of existing QKD protocols are limited to point-to-point applications or are heavily reliant on specific infrastructures, leaving much of the global network infrastructure vulnerable. This paper seeks to address this challenge by proposing a novel concept for a network-oriented QKD protocol.

- Research problem: Currently available protocols are only suitable in a point-to-point scenario.
- Research questions: a) Is it possible to construct a different protocol that would be able to support network communication? b) Is it possible to make it applicable to the current network infrastructure?
- Research objectives: a) to review the related literature b) to find suitable protocols for the optic fiber-based network communication c) explain the proposed approach mathematically.

- Research significance and contribution: a novel QKD network-oriented approach applicable to the current optic fiber infrastructure.

II. Literature Review

A. Historical Origins and the Emergence of First QKDs

The early 1970s began with the initial development of Quantum Key Distribution (QKD) protocols. By 1984, the scientific community was introduced to the novel polarization-based algorithm for key distribution [11] developed by C. H. Bennett and G. Brassard, marking a significant milestone. In this work, Bennett and Brassard proposed a key distribution protocol based on the polarization property of a quantum state as well as a change of measurement bases. Although its protocol as-is can be utilized over the current infrastructure, it seriously lacks in terms of security against such attacks [12] as IRUD attacks, Beam-Splitting attacks, Denial of Service attacks, Man-In-The-Middle, IRA attacks, etc. Therefore, making it not a standalone QKD solution but a potential building block for a bigger picture.

B. Entanglement-based QKDs

This discovery was closely followed by another, in 1992, with the presentation of the first entanglement-based algorithm [13] developed by A. K. Ekert. In this work, Ekert utilized the property of entanglement in order to address the possibility of an eavesdropping attack. However, this protocol as-is also quite vulnerable [12] to IRUD attacks, Beam-Splitting attacks, and Denial of Service attacks. Overall, these innovative algorithms utilized fundamental concepts of quantum physics such as the Entanglement Effect, Quantum Teleportation, Polarization, etc. These foundational algorithms have paved the way for all subsequent research in the field.

C. Review of Measurement Device Independent QKDs

The next logical step in the development of this branch was the new protocols that further advanced the complexity of security-assuring physical phenomena, such as BBM92 [14], SARG04 [15] [16], KMB09 [15], AK15 [16], etc. As well as continuous testing and improvement of already existing ones. For instance, since the emergence of E91 as a theoretical concept, there have been many tests that piece-by-piece proved the concept [17] [18] [19] [20], yet still failed to prove its applicability in field test or real-world applications due to poor unstable key-generation rates, relying on a theoretical piece of equipment such as quantum repeaters, limited duration of CHSH violation, or poor handling of noise. The same was done, albeit more successfully, for BB84 [21][22][23]. Although, BB84 is still suffers from the weak coherence of quantum states during transmission, which is limiting its operational range significantly. It also suffers from a limited key generation rate as-is, though there is a possibility for improvement. However, while E91 has hardly ever seen practical field applications, BB84 has already been tested in real-world applications [24] [25] and is already commercially available. After that, the next big step in the development of QKD protocols was Measurement-Independent QKD (MI-QKD) that are removing all detector side-channel attacks as well as Device-Independent QKD (DI-QKD), which security does not rely on trusting that the quantum devices used are truthful. Ultimately, these two sub-branches merged into one (MDI-QKD).

D. Twin-Field-based MDI-QKDs

One representative of this sub-branch is a Twin-Field group of QKD protocols [26] that provide a much higher key rate and greater distance compared to previous strategies (such as adding extra loss or not using any compensation). [27]. Some examples of Twin-Field QKD Protocols are include but not limited to: Sending-Not-Sending (SNS) [28], CAL19 [29], or Phase-Matching Protocol [30], which demonstrates the potential to overcome the key-rate limit and achieves a quadratic improvement over phase-encoding MDI-QKD [30]. For instance, the SNS protocol claimed to reach a distance limit of up to 800 km without misalignment error, while authors of CAL19 managed to find a solution to the key-rate drop issue of the original TF-QKD by Lucamarini et al. [26] and improve the key-rate by an order of magnitude. All of these protocols not only provided ways of robust security against common threats but also addressed some of the crucial issues on the way toward actually functioning Quantum Network [27].

E. Authentication

While all of these algorithms and approaches can be effective to various degrees and the question of a central node becoming trusted is still standing, one has to consider an approach for another big question that could render previously mentioned algorithms useless – the authentication phase. Currently, few quantum authentication algorithms would apply to this setup. Firstly, one should focus on those algorithms that do not utilize entanglement or use it in a limited capacity, since a system that requires necessary equipment for entanglement would be considerably expensive.

A good example is the work of Kanamori et al. [31]. Instead of solely relying on entanglement or a trusted center, authors chose to capitalize on the superposition. One big advantage is that this particular algorithm can re-use the TFQKD (1 phase) for the initial authentication. Another advantage is that it can be utilized even without a classical channel of communication, which provides additional security due to the dispersed approach. However, it would be cumbersome to re-use this algorithm due to the need for the generation of new keys. Another great example is the work of Zhang et al. [32]. This approach shares many advantages with the previous one, but it has one that might tip the scales to its side - it can be re-used later without the re-generation of the key.

Although the algorithms that require devices related to entanglement can make the whole system considerably expensive, it is still required to review those that fit the design of this setup. For instance, the work of Lin et al. [33] could be utilized because it does not require a trusted center. Additionally, a lot of the crucial mechanisms that are necessary are also pretty straightforward, such as - a combination of CNOT gates, different measurement bases, etc. This approach also does not rely on a classical communication channel, which is a plus.

Despite the abundance of available algorithms, further security analysis is required.

III. Similar Works

While the idea of Lucamarini et al. [26] is still - comparatively - fresh, there are many teams worldwide already who share the same excitement and the desire for a more secure, far- reaching, multi-node QKD protocol. For example, the workof Cao et al. [34] [35] attempts to improve on the protocol provided by Lucamarini et al. by providing it with additional layers of randomization and detection. However, it is still a standalone protocol that does not cover all of the inherent security concerns of multi-node communication. One such example could be the insider attack, both from a compromised center and/or nodes. As for the work of Metwaly et al. [36], while it does have an all-encompassing approach to ensure the security of the network as well as providing a way of scaling this approach for a network of networks, it is still very theoretical and lacks concrete examples of how certain stages can be achieved, if at all. A good example of the same idea but with better authentication ingrained would be the work of Sellami et al. [37]. In this work, a fairly straightforward approach to authentication was described. Still, the question of a trusted center stands.

IV. Methods

A. Twin-Field Quantum Key Distribution

Twin-Field Quantum Key Distribution - is a protocol that is one of many protocols (more specifically MDI-QKD protocols) that sup- ports the delivery or distribution of secret key fragments or complete secret keys between certain parties via the utilizationof laws of quantum mechanics. More specifically, the classicalversion of this protocol [26] utilizes the notion of wave-particle interference between two parties Alice and Bob who utilize a remote measuring device, which is called Charlie or Eve. Eachof the participants utilizes what is called Weak Coherent State [38] in the X basis as well as Decoy State [39] in the Z basis both have assigned randomized phases and intensities.

Twin-Field Quantum Key Distribution (TF-QKD) is a protocol within the broader category of Measurement-Device-Independent Quantum Key Distribution (MDI-QKD) protocols, designed to facilitate the secure delivery or distribution of secret key fragments or complete secret keys between parties using the principles of quantum mechanics. The classical version of this protocol [26] (General Scheme is shown in Fig. 1.) employs the concept of wave-particle interference between two parties, commonly referred to as Alice and Bob, who interact through a remote measurement device, often termed Charlie or Eve. Each participant utilizes a Weak Coherent State [38] in the X basis and a Decoy State [39] in the Z basis, both of which are characterized by randomized phases and intensities.

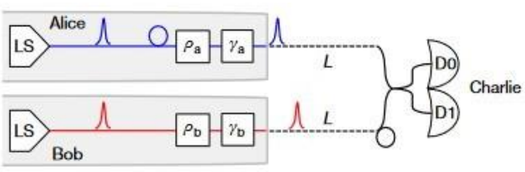


Fig. 1. Twin field QKD general scheme [26].

After the randomization of phases and intensities, each participant transmits respective Weak Coherent States (WCS) to Eve or Charlie, who performs the measurement and subsequently announces the acquired result. The announcement indicates whether the measurement detected photons with matching logical values (00 and 11) or differing values (10 and 01). Despite Eve being the entity that conducts the measurement and reports the results, Eve remains unaware of the actual key values (whether the bits are 1 and 1 or 0 and 0); Eve only knows the parity of the results. This particular QKD protocol ensures the centralized delivery of the "network portion" of the key to all hosts while providing robust security against external threats such as eavesdropping and Man-in-the-Middle (MITM) attacks.

B. KMB09

KMB09 is a protocol that, despite some skepticism, is considered part of the broader category of Measurement-Device-Independent Quantum Key Distribution (MI-QKD) protocols. The protocol relies on the mechanism of encoding a qubit into at least four different states (for simplicity, N=2 is considered) for Alice using two bases, E and F, as shown in Fig. 2. In the initial step, Alice randomly selects a basis and an index for encoding a photon and transmits the encoded photon through a quantum channel to Bob. Bob then measures the incoming photons using a randomly chosen basis. For Bob's measurement to be meaningful, Alice must disclose some information about the chosen bases publicly through a classical communication channel, such as fiber-optic. Specifically, Alice needs to reveal the selected index, either 1 or 2. However, this disclosure does not allow Eve or any other malicious party to gain knowledge about the key, as even with knowledge of the index, Eve cannot determine which basis Alice chose.

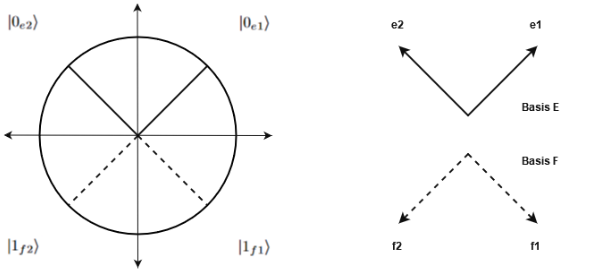


Fig. 2. KMB09 bases.

For example, if basis E is used to encode 0 and basis F is used to encode 1, the outcome of Bob's measurement, in conjunction with the non-parity of indices chosen by both Alice and Bob, determines whether the result is 1 or 0. If the indices match, a "no signal" message is announced to enhance the security of the transmission. This step ensures that the transmission remains secure even in the presence of potential eavesdropping.

As a result, and in alignment with findings from the original research article [40] and a recent overview paper [15], this protocol ensures the secure exchange of user or node-specific key portions between network participants, effectively mitigating the risk of intercept-resend attacks and similar types of security threats.

C. Proposed Method

This section details the functioning of the proposed hybrid concept within a network infrastructure that accommodates multiple users. The core objective of this concept is to secure communication among a verified number of nodes or clients within a centralized, untrusted network, as illustrated in Fig. 3.

To accomplish this, the approach combines the strengths of Measurement-Device-Independent Quantum Key Distribution (MDI-QKD), Measurement-Independent Quantum Key Distribution (MI-QKD), Continuous Variable QKD (CV-QKD), and Discrete Variable QKD (DV-QKD).

In this configuration, the untrusted center and the primary measuring device can be represented by entities such as Charlie or Eve, as the specific identity is inconsequential. The current iteration of this hybrid protocol is designed for integration with classical infrastructure. Consequently, the quantum channels utilized are standard fiber-optic cables.

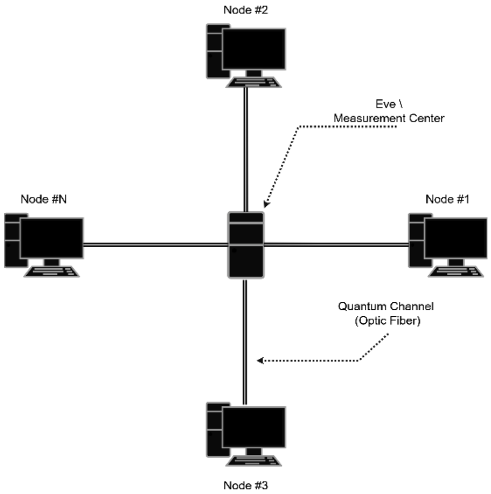


Fig. 3. General setup.

First, it is essential to verify the identities of all nodes within the network and eliminate any impostors. This can be achieved by employing a Quantum Digital Signature (QDS) protocol, such as the one described in study [41]. Once this verification process is complete, the first phase of the protocol can begin.

In the first phase, the Twin-Field QKD protocol is employed, wherein Weak Coherent States (WCSs) are transmitted from each authenticated node to the untrusted central node, Eve. Eve then measures the combined interference of all the sent states. The resulting values are not strictly 1 or 0 but rather a fluctuation between them. These fluctuations can be resolved using the Sigmoid Function, with the results publicly announced. By following this process, all nodes within a specific timeframe will obtain the "network portion" of the key.

In this step, each node transmits its respective randomized Weak Coherent States to initiate the creation of the "network portion" of the key at the measuring device, Eve (Fig. 4). This process should be repeated K times until a sufficient number of bits is accumulated within the "network portion" of the key.

In detail, each authenticated node U_i generates weak coherent states $|a_i\rangle$, where a_i represents the amplitude of the coherent state. The coherent state $|a_i\rangle$ can be expressed by following Eq. (1):

$$a_i = e^{\frac{\left(-|a_i|^2\right)}{2}} \sum_n^\infty \frac{a_i^n}{\sqrt{n!}} |n\rangle \tag{1}$$

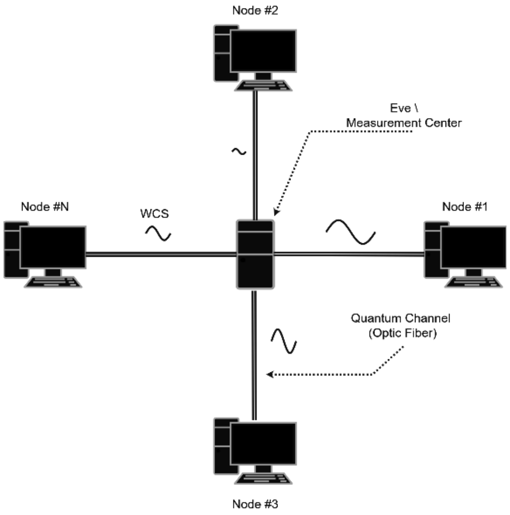


Fig. 4. Simplified stage 1.

where$|n\rangle$ represents the state with n photons. These states are sent to the central node E (Eve). The central node E measures the interference of all coherent states $|a_i\rangle$ sent by the different nodes. The total state at the central node can be described by a superposition of coherent states, refer to the Eq. (2):

$$|\phi\rangle = \sum_j |a_i\rangle \tag{2}$$

where $|a_i\rangle$ is the coherent state sent by the node U_j. The measured interference values I will fluctuate between 0 and 1. To convert these fluctuations into a more convenient format, the sigmoid function is applied. For the example refere to the Eq. (3):

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

where x is the measured interference value. The result of $\sigma(1)$ provides a probabilistic estimate, which is then publicly announced to all nodes.

Once the measurement results are announced, each node can utilize this data to generate the "network portion" of the key. Let the measured values for node U_i and the central node E be denoted as K_i and K_E, respectively. Then, the key fragment for node U_i can be described as the following function (4):

$$U_i = f(\sigma(I), metadata) \tag{4}$$

where f is a function that defines how the measured data is transformed into key values.

Following the distribution of the "network portion" (NP), the next phase involves the generation and organization of "pair portions" (PP). This phase requires the application of the KMB09 protocol for individual pairing and key exchange. Each node or client must initiate a pairing process with every other node in the network, resulting in a total of $n - 1$ pairings per node. Consequently, the total number of unique keys generated will be $(n * (n - 1))/2$. The uniqueness of these pairwise keys is critical for ensuring security, as it provides protection not only against external threats but also from potential internal eavesdroppers.

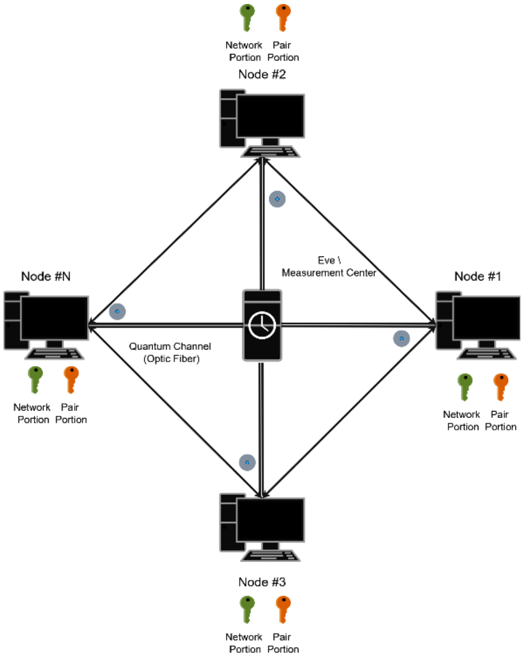


Fig. 5. Simplified stage 2.

In this setup (Fig. 5), each node transmits its randomly base-encoded photons to other nodes via a central untrusted measuring device, Eve. In this scenario, Eve acts purely as an intermediary, directing the photons to the appropriate quantum channels between the nodes intended for pairing. As a result, Eve does not obtain any information about the pair key, even if Eve attempts to intercept and resend a photon.

Thus, both the "network portion" and the "pair portion" have been successfully established. These keys can now be combined to generate a pair-unique master key, which facilitates the initiation of encrypted communication between selected nodes. To further elucidate the second part of the protocol, a mathematical analysis will be provided.

Consider a network consisting of n nodes. Each node U_i must establish a secure connection with each of the other nodes U_j, where $i \neq j$. For each node U_j, it is necessary to establish pairwise connections with the remaining $n-1$ nodes. As a result, there will be $\frac{n(n-1)}{n}$ unique pairwise keys. This quantity is determined by the formula for the number of combinations provided below (5):

$$\binom{n}{k} = \frac{n(n-1)}{2} \tag{5}$$

The KMB09 protocol is employed to generate and exchange pairwise keys between nodes. This protocol is grounded in quantum mechanics and includes the following steps:

- Initialization: Nodes U_i and U_j initiate the process by exchanging quantum states ϕ_i and ϕ_j, respectively.

- Measurement: Each node conducts measurements on the quantum state received from the other node.

- Key Extraction: Based on the measurements obtained, each node derives the key information K_{ij} corresponding to the secure communication between nodes U_i and U_j .

Let the key generated between nodes U_i and U_j be denoted as K_{ij}. Mathematically, this can be expressed as a key generation function provided below (6):

$$K_{ij} = f\left(|\phi_i\rangle, |\phi_j\rangle\right) \tag{6}$$

where f is a function that defines the algorithm for deriving a key based on the exchange of quantum states. After generating the pairwise keys K_{ij} for each pair of nodes, each node possesses:

- The network portion of the key $K_{network}$, which was generated during the first phase.

- The paired portion of the key K_{ij}, which was generated during the second phase for each node U_j .

To generate a unique master key for a pair of nodes U_i and U_j , it is necessary to combine their respective key components. Let $K_{master,ij}$ denote the master key for nodes U_i and U_j as in the example provided below (7):

$$K_{master,ij} = g\left(K_{master}, K_{ij}\right) \tag{7}$$

where g is a function that defines the method for combining the network portion and the pair portion of the keys.

Typically, this process involves applying an XOR operation or another concatenation function (8):

$$K_{master,ij} = K_{master} \oplus K_{ij} \tag{8}$$

where $\oplus$ represents the bitwise XOR operation.

To evaluate the scalability of the protocol, a graphical representation of the network is employed (Fig. 5). In a network consisting of *N* nodes, each node can exchange keys with every other node. This configuration can be visually depicted as a complete graph, where the nodes are represented as vertices and the connections between them are illustrated as edges.

V. Results and Discussion

In this paper, after careful comparison and analysis of existing methods, algorithms, and protocols a novel approach was proposed. This Hybrid Twin-Field QKD approach presents an opportunity to securely generate and share a secret key, communicate between specific nodes secured from internal eavesdroppers by the KMB09 protocol, and communicate within a network secured from outside interferences and eavesdroppers by Twin-Field QKD.

While the proposed hybrid QKD protocol is theoretically feasible, its practical implementation is currently limited, as only the individual components have been demonstrated to be achievable in real-world settings. Moreover, although Twin-Field QKD theoretically supports communication distances of up to 600 km or even 800 km, the protocol's overall range is constrained by the shortest distance supported by KMB09. This limitation highlights an important objective for future work:

extending the effective communication range of the hybrid protocol.

Additionally, there is a need for a more comprehensive analysis of the internal and external security aspects of the proposed concept, including metrics such as Quantum Bit Error Rate (QBER) [42]. Further research is also necessary in related areas, including Quantum Digital Signature (QDS) protocols, quantum authentication protocols in general, and improvements to the performance of KMB09. These avenues of investigation are crucial for enhancing the robustness and practicality of the hybrid QKD protocol.

VI. Conclusion

In conclusion, this paper has introduced and thoroughly analyzed a hybrid Twin-Field Quantum Key Distribution (QKD) protocol tailored for multi-user quantum networks. The proposed protocol addresses the increasing need for secure communication within untrusted, centralized networks, leveraging the strengths of both classical and quantum cryptographic techniques. By combining elements from various Quantum Key Distribution Protocols (QKDPs), the hybrid approach enhances the scalability and security of key distribution among multiple nodes.

The paper has provided a detailed examination of the global security landscape, highlighting the evolving challenges posed by quantum computing and the limitations of traditional cryptographic methods. Through a historical overview of QKDPs, the research identified key areas for improvement and integrated these insights into the proposed protocol.

The hybrid Twin-Field QKD protocol offers a robust solution for secure key distribution in complex network environments, ensuring protection against both external and internal threats. As quantum technologies continue to advance, this protocol represents a significant step toward realizing secure, scalable quantum communication networks. Future work may focus on further optimizing the protocol's efficiency and exploring its practical implementation in real-world quantum networks.

Acknowledgment

Research work was carried out within the framework of the project AP19675961 “Development and research of keys distribution protocols based on quantum properties”, which is being implemented at the Non-profit joint-stock company "Kazakh National Research Technical University named after K.I. Satbayev".

References

[1] Estimated cost of cybercrime worldwide 2018-2029. [Online]. Available: https://www.statista.com/forecasts/1280009/ cost-cybercrime-worldwide.

[2] The future of computing is quantum-centric. [Online]. Available: https://www.ibm.com/roadmaps/quantum/.

[3] V. Bhatia and K. Ramkumar, “An efficient quantum computing technique for cracking rsa using shor’s algorithm,” in 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020, pp. 89–94.

[4] S. Jaques, M. Naehrig, M. Roetteler, and F. Virdia, “Implementing grover oracles for quantum key search on aes and lowmc,” in Advances in Cryptology – EUROCRYPT 2020, A. Canteaut and Y. Ishai, Eds. Cham: Springer International Publishing, 2020, pp. 280–310.

[5] Biyashev, R.G., Kalimoldayev M.N., Nyssanbayeva, S.E., Kapalova N.A., Dyusenbayev, D.S., Algazy K.T., Development and analysis of the encryption algorithm in nonpositional polynomial notations // Eurasian Journal of Mathematical and Computer Applications. – 2018. - № 6(2). - P.19-33. DOI: 10.32523/2306-6172-2018-6-2-19-33.

[6] R.G. Biyashev, N.A. Kapalova, D.S. Duysenbayev, K.T. Algazy, Waldemar Wojcik, Andrzej Smolarz Development and Analysis of Symmetric Encryption Algorithm Qamal Based on a Substitution-permutation Network, International journal of electronics and telecommunications, № 1, 2021, P. 127-132 DOI: 10.24425/ijet.2021.135954.

[7] R. G. Biyashev, S. E. Nyssanbayeva, and Ye. Y. Begimbayeva Development of the model of protected cross-border information interaction // Open Engineering. – 2016. – № 6. – P. 199 – 205, DOI: https://doi.org/10.1515/eng-2016-0025.

[8] Maksat N. Kalimoldayev, Rustem G. Biyashev, Saule E. Nyssanbayeva, Yenlik Ye. Begimbayeva Modification of the digital signature, developed on the nonpositional polynomial notations // Eurasian Journal of Mathematical and Computer Applications. – 2016. – Vol. 4, Is. 2. – P. 33 – 38, DOI: 10.32523/2306-6172-2016-4-2-33-38.

[9] Y. Begimbayeva, T. Zhaxalykov and O. Ussatova, "Investigation of Strength of E91 Quantum Key Distribution Protocol," 2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS), Novosibirsk, Moscow, Russian Federation, 2023, pp. 10-13, doi: 10.1109/OPCS59592.2023.10275771.

[10] Ussatova, O., Makilenov, S., Mukaddas, A., Amanzholova, S., Begimbayeva, Y., & Ussatov, N. (2023). Enhancing healthcare data security: a two-step authentication scheme with cloud technology and blockchain. Eastern-European Journal of Enterprise Technologies, 6(2 (126), 6–16. https://doi.org/10.15587/1729-4061.2023.289325.

[11] C. H. Bennett and G. Brassard, “Quantum cryptography: Public key distribution and coin tossing,” Mar 2020. [Online]. Available: https://arxiv.org/abs/2003.06557v1

[12] A. Abushgra and K. Elleithy, “Qkdp’s comparison based upon quantum cryptography rules,” in 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016, pp. 1–5.

[13] A. K. Ekert, “Quantum cryptography based on bell’s theorem,” Phys. Rev. Lett., vol. 67, pp. 661–663, Aug 1991. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.67.661

[14] C. H. Bennett, G. Brassard, and N. D. Mermin, “Quantum cryptography without bell’s theorem,” Phys. Rev. Lett., vol. 68, pp. 557–559, Feb 1992. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.68.557

[15] M. Lopes and N. Sarwade, “On the performance of quantum cryptographic protocols sarg04 and kmb09,” in 2015 International Conference on Communication, Information Computing Technology (ICCICT), 2015, pp. 1–6.

[16] A. A. Abushgra, “Sarg04 and ak15 protocols based on the run-time execution and qber,” in 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP), 2021, pp. 176–180.

[17] R. Ursin, F. Tiefenbacher, T. Schmitt-Manderbach, H. Weier, T. Scheidl, M. Lindenthal, B. Blauensteiner, T. Jennewein, J. Perdigues, P. Trojek, B. O¨ mer, M. Fu¨rst, M. Meyenburg, J. Rarity, Z. Sodnik, C. Barbieri, H. Weinfurter, and A. Zeilinger, “Entanglement-based quantum communication over 144km,” Nature Physics, vol. 3, no. 7, p. 481–486, Jun 2007. [Online]. Available: http://dx.doi.org/10.1038/nphys629

[18] A. Ling, M. P. Peloso, I. Marcikic, V. Scarani, A. Lamas-Linares, and C. Kurtsiefer, “Experimental quantum key distribution based on a bell test,” Physical Review A, vol. 78, p. 020301, 8 2008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.78.020301

[19] M. Fujiwara, K. ichiro Yoshino, Y. Nambu, T. Yamashita, S. Miki, H. Terai, Z. Wang, M. Toyoshima, A. Tomita, and M. Sasaki, “Modified e91 protocol demonstration with hybrid entanglement photon source,” Optics Express, vol. 22, p. 13616, 6 2014. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?uri=oe-22-11-13616

[20] J. Yin, Y. Cao, and et al., “Satellite-based entanglement distribution over 1200 kilometers,” Science, vol. 356, no. 6343, p. 1140–1144, Jun 2017. [Online]. Available: http://dx.doi.org/10.1126/science.aan3211

[21] B. Kebapci, V. E. Levent, S. Ergin, G. Mutlu, I. Baglica, A. Tosun, P. Paglierani, K. Pelekanakis, R. Petroccia, J. Alves, and M. Uysal, “Fpga-based implementation of an underwater quantum key distribution system with bb84 protocol,” IEEE Photonics Journal, vol. 15, no. 4, pp. 1–10, 2023.

[22] C. Lee, I. Sohn, and W. Lee, “Eavesdropping detection in bb84 quantum key distribution protocols,” IEEE Transactions on Network and Service Management, vol. 19, no. 3, pp. 2689–2701, 2022.

[23] M. Stipcˇevic´, “Enhancing the security of the bb84 quantum key distribution protocol against detector-blinding attacks via the use of an active quantum entropy source in the receiving station,” Entropy, vol. 25, no. 11, 2023. [Online]. Available: https://www.mdpi.com/ 1099-4300/25/11/1518

[24] J. F. Dynes, A. Wonfor, and et al., “Cambridge quantum network,” npj Quantum Information, vol. 5, no. 1, Nov 2019. [Online]. Available: http://dx.doi.org/10.1038/s41534-019-0221-4

[25] M. Sasaki, M. Fujiwara, and et al., “Field test of quantum key distribution in the tokyo qkd network,” Mar 2011. [Online]. Available: https://arxiv.org/abs/1103.3566v1

[26] M. Lucamarini, Z. L. Yuan, J. F. Dynes, and A. J. Shields, “Overcoming the rate–distance limit of quantum key distribution without quantum repeaters,” Nature, vol. 557, no. 7705, p. 400–403, May 2018. [Online]. Available: http://dx.doi.org/10.1038/s41586-018-0066-6

[27] X. Zhong, W. Wang, L. Qian, and H.-K. Lo, “Proof-of-principle experimental demonstration of twin-field quantum key distribution over optical channels with asymmetric losses,” npj Quantum Information, vol. 7, no. 1, Jan 2021. [Online]. Available: http://dx.doi.org/10.1038/ s41534-020-00343-5

[28] Z.-W. Yu, X.-L. Hu, C. Jiang, H. Xu, and X.-B. Wang, “Sending-or- not-sending twin-field quantum key distribution in practice,” Scientific Reports, vol. 9, no. 1, p. 3080, Feb 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-39225-y

[29] M. Curty, K. Azuma, and H.-K. Lo, “Simple security proof of twin-field type quantum key distribution protocol,” npj Quantum Information, vol. 5, no. 1, Jul 2019. [Online]. Available: http: //dx.doi.org/10.1038/s41534-019-0175-6

[30] X. Ma, P. Zeng, and H. Zhou, “Phase-matching quantum key distribution,” Phys. Rev. X, vol. 8, p. 031043, Aug 2018. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.8.031043

[31] Y. Kanamori, S.-M. Yoo, D. A. Gregory, and F. T. Sheldon, “Authentication protocol using quantum superposition states,” International Journal of Network Security, vol. 9, no. 2, p. 101–108, Jan. 2009. [Online]. Available: http: //ijns.jalaxy.com.tw/contents/ijns-v9-n2/ijns-2009-v9-n2-p101-108.pdf

[32] D. Zhang and X. Li, “Quantum authentication using orthogonal product states,” in Third International Conference on Natural Computation (ICNC 2007), vol. 4, 2007, pp. 608–612.

[33] T.-S. Lin, I.-M. Tsai, H.-W. Wang, and S.-Y. Kuo, “Quantum authentication and secure communication protocols,” in 2006 Sixth IEEE Conference on Nanotechnology, vol. 2, 2006, pp. 863–866.

[34] X.-Y. Cao, Y.-S. Lu, Z. Li, J. Gu, H.-L. Yin, and Z.-B. Chen, “High key rate quantum conference key agreement with unconditional security,” IEEE Access, vol. 9, p. 128870–128876, Jan. 2021. [Online]. Available: https://doi.org/10.1109/access.2021.3113939

[35] X.-Y. Cao, J. Gu, Y.-S. Lu, H.-L. Yin, and Z.-B. Chen, “Coherent one-way quantum conference key agreement based on twin field,” New Journal of Physics, vol. 23, no. 4, p. 043002, Apr. 2021. [Online]. Available: https://doi.org/10.1088/1367-2630/abef98

[36] A. Metwaly, M. Z. Rashad, F. A. Omara, and A. A. Megahed, “Architecture of point to multipoint qkd communication systems (qkdp2mp),” in 2012 8th International Conference on Informatics and Systems (INFOS), 2012, pp. NW–25–NW–31.

[37] S. Ali, O. Mahmoud, and A. A. Hasan, “Multicast network security using quantum key distribution (qkd),” Jul. 2012. [Online]. Available: https://doi.org/10.1109/iccce.2012.6271355

[38] T. F. da Silva, G. C. do Amaral, D. Vitoreti, G. P. T. ao, and J. P. von der Weid, “Spectral characterization of weak coherent state sources based on two-photon interference,” J. Opt. Soc. Am. B, vol. 32, no. 4, pp. 545–549, Apr 2015. [Online]. Available: https://opg.optica.org/josab/abstract.cfm?URI=josab-32-4-545

[39] W.-Y. Hwang, “Quantum key distribution with high loss: Toward global secure communication,” Phys. Rev. Lett., vol. 91, p. 057901, Aug 2003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett. 91.057901

[40] M. M. Khan, M. Murphy, and A. Beige, “High error-rate quantum key distribution for long-distance communication,” New Journal of Physics, vol. 11, no. 6, p. 063043, jun 2009. [Online]. Available: https://dx.doi.org/10.1088/1367-2630/11/6/063043

[41] C.-H. Zhang, X. Zhou, C.-M. Zhang, J. Li, and Q. Wang, “Twin-field quantum digital signatures,” Opt. Lett., vol. 46, no. 15, pp. 3757–3760, Aug 2021. [Online]. Available: https://opg.optica.org/ol/abstract.cfm?URI=ol-46-15-3757

[42] M. Niemiec and A. R. Pach, “The measure of security in quantum cryptography,” Dec. 2012. [Online]. Available: https://doi.org/10.1109/ glocom.2012.6503238.

The Role of Artificial Intelligence in Enhancing Business Intelligence Capabilities for E-Commerce Platforms

Sinek Mehuli Br Perangin-Angin[1]*, David Jumpa Malem Sembiring[2], Asprina Br Surbakti[3], Soleh Darmansyah[4]
Institut Teknologi Dan Bisnis Indonesia, Medan, Indonesia[1, 2, 3]
Faculty of Engineering, Universitas Medan Area, Medan, Indonesia[4]

***Abstract*—This research focuses on the application of BERT (Bidirectional Encoder Representations from Transformers) and Graph Neural Networks (GNNs) to improve business intelligence (BI) capabilities on e-commerce platforms. The main aim of the research is to develop automation methods for the classification of customer interactions and to create a more effective product recommendation system. In this study, BERT was used to analyze and classify customer interaction texts, including questions, complaints, and reviews, with accuracy reaching 97% and sentiment analysis accuracy of 93%. GNNs are applied to model complex relationships between customers and products based on transaction data, then used to provide product recommendations. The evaluation results show that the GNNs model achieved a mean average precision (MAP) of 0.92 and a normalized discounted cumulative gain (NDCG) of 0.88, indicating high relevance and accuracy in product recommendations. This research concludes that the integration of BERT and GNNs improves operational efficiency through classification automation but also provides added value in marketing strategies with better personalization of recommendations.**

Keywords*—Bidirectional Encoder Representations from Transformers (BERT); Graph Neural Networks (GNNs); business intelligence; e-commerce; product recommendation

I. Introduction

In recent years, the e-commerce industry has experienced a significant surge in growth, resulting in the generation of vast and intricate amounts of data [1]. This dataset contains comprehensive information regarding customer transactions, shopping patterns, customer engagements, and more relevant data. In order to stay competitive, e-commerce platforms must create advanced business intelligence (BI) systems to evaluate this data and extract practical insights. In this situation, it is critical to enhance business intelligence (BI) skills using artificial intelligence (AI) technology [2], [3].

Bidirectional Encoder Representations from Transformers (BERT) and Graph Neural Networks (GNNs) are two advanced AI algorithms that have the potential to greatly improve business intelligence capabilities for e-commerce systems [4-6]. Google created the sophisticated transformer model known as BERT. This paradigm has transformed the field of natural language processing (NLP) because of its ability to comprehend bidirectional context in text. Natural language processing tasks such as natural language interpretation, text classification, and information retrieval can all benefit from the versatile application of BERT [7], [8]. BERT can enhance the study of customer interaction in e-commerce. By leveraging its advanced capabilities in comprehending conversations and context, BERT has the potential to enhance customer service by delivering more prompt and tailored assistance [9]. Graph Neural Networks (GNNs) represent a distinct class of neural networks designed to process and analyze data structured in a graph format. This algorithm is highly efficient at analyzing intricate linkages and interactions among diverse elements. By representing products, customers, and transactions as graphs, we can employ Graph Neural Networks (GNNs) to detect intricate patterns and correlations in e-commerce. This is highly beneficial for constructing more precise and pertinent product suggestion systems by utilizing data on the correlation between products and client preferences [10-13].

The objective of this study is to investigate the utilization of BERT and GNNs in enhancing business intelligence capabilities on e-commerce platforms [14]. This research aims to analyze customer interactions and make product recommendations, with the goal of gaining new insights into how artificial intelligence (AI) can enhance customer experience and boost business performance in the e-commerce industry [15]. It also seeks to evaluate the effectiveness of different algorithms in real-world situations, offering guidance for implementing improved business intelligence (BI) in the future [16].

II. Research Methodology

A. Methodology

The research commences by examining consumer interactions utilizing the BERT (Bidirectional Encoder Representations from Transformers) paradigm [17]. The initial stage is gathering conversational text data, reviews, and client inquiries from e-commerce platforms. Subsequently, this data undergoes pre-processing to eliminate any unwanted interference and make it ready for subsequent analysis. BERT models undergo fine-tuning with these datasets to perform tasks like text categorization and comprehension of conversational context, resulting in more profound characteristics and a better grasp of consumer preferences and requirements [18].

Following the study using BERT, the research proceeds by constructing an association graph between customers and items with graph neural networks (GNNs) [19]. Graphs are constructed using transaction data and BERT analysis outcomes to depict intricate connections among entities. Graph Neural Networks (GNNs) are subsequently utilized to represent and examine relationship patterns inside these graphs, with a specific emphasis on producing precise and pertinent product recommendations [20], [21].

Performance evaluations are conducted by utilizing metrics like precision, recall, F1-score, and NDCG (normalized discounted cumulative gain) to assess the efficiency of the recommendation model [22]. The outcomes of these two phases are merged to offer a more all-encompassing business intelligence solution and enhance the e-commerce platform's abilities in comprehending and catering to its customers can be seen in Fig. 1.

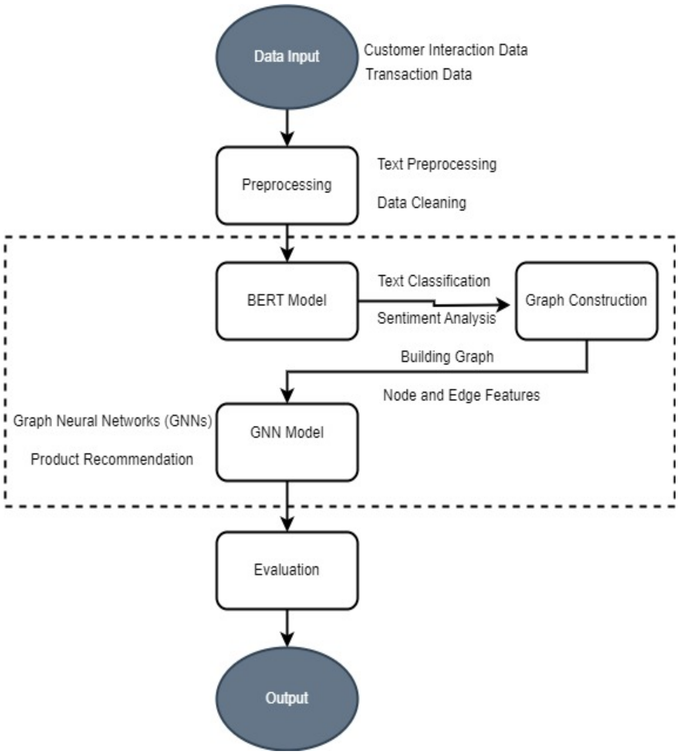


Fig. 1. Research methodology.

B. Problem Solving Approach

1) Artificial Intelligence (AI) in Business Intelligence (BI) for e-commerce: Artificial intelligence (AI) plays a crucial role in advancing business intelligence (BI) systems [23]. Within the realm of electronic commerce, business intelligence (BI) converts unprocessed data, including sales transactions, customer activity, and user interactions, into significant insights that influence company decision-making [24]. By integrating artificial intelligence (AI), business intelligence (BI) systems may analyze vast quantities of data, uncover concealed patterns, and offer highly precise, predicted insights. This integration significantly improves an e-commerce platform's capacity to comprehend and cater to its clients in a more efficient manner [25].

2) Bidirectional Encoder Representations from Transformers (BERT): BERT possesses the capacity to comprehend bidirectional context in text, enabling it to analyze words in sentence context from both the left-to-right and right-to-left orientations. These tasks, such as text classification, information retrieval, and natural language understanding, are particularly crucial [26]. In the realm of e-commerce, BERT is employed to scrutinize client interactions, including product evaluations and dialogues with customer care, to extract profound insights into customer requirements and inclinations. Utilizing this comprehension of consumer interactions helps enhance customer service and customization. Below are the sequential instructions for utilizing BERT. The following main components underlie BERT [27-30]:

Phase 1→ Input Embeddings

$$mbedding\ (x_i) = TokenEmbedding\ (x_i) + ESegmentEmbedding\ (x_i) + PositionEmbedding\ (x_i) \quad (1)$$

Phase 2 → Self-Attention Mechanism

BERT's key component self-attention allows the model to focus on different parts of the input sequence when processing each token.

Scaled Dot-Product Attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Where:

Q query matrix

K key matrix

V value matrix

d_k dimensions of the key

Multi-Head Attention

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^O \quad (3)$$

Where:

$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

W_i^Q, W_i^K, W_i^V weight matrix for the i th head

W^O output weight matrix

Phase 3 → Feed-Forward Neural Network

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \quad (4)$$

Phase 4 → Layer Normalization

$$LayerNorm(x) = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}}.\gamma + \beta \quad (5)$$

Where:

$E[x]$ the average value of the input

$Var[x]$ variance of the input

ϵ small constant to prevent division by zero

γ and β parameters that can be studied

Phase 5 → Loss Function

$$Loss = -\sum i \in masked\ positions\ LogP(x_i|x_{<i}, x_{>i}) \quad (6)$$

Phase 6 →Training Objective

BERT is trained in masked language modeling (MLM) and next sentence prediction (NSP).

Masked Language Modelling (MLM)

$$MLM\ Loss = -\sum i \in masked\ positions\ LogP\ (x_i|x_{masked}) \quad (7)$$

Next Sentence Prediction (NSP)

The model is trained to predict whether two input sentences appear sequentially in the original text.

$$NSP\ Loss = -(y \log P\ (IsNext) + (1-y)\log P(NotNext)) \quad (8)$$

Where y Binary labels indicate whether two sentences are sequential or not.

3) Graph Neural Networks (GNNs): Graph neural networks (GNNs) proficiently examine intricate connections and interactions among diverse elements. Graphs consist of nodes that can represent various entities, such as consumers and items [31]. The connections between these entities, such as buy transactions, are represented by edges. Graph Neural Networks (GNNs) have the ability to acquire knowledge about these graphs and are employed for tasks such as predicting connections (e.g., suggesting products) and categorizing nodes. E-commerce platforms can enhance the precision and relevance of their product recommendation algorithms by leveraging graph neural networks (GNNs) to analyze intricate relationship patterns between products and customers [32-34].

Phase 1 → Graph Representation

Graf G defined as a pair (V, E), where V is the set of nodes and E is a set of edges.

Phase 2 → Basic Notation

h_u Features of neighboring nodes u

W Learnable weight matrix

σ Non-linear activation function

$N(v)$ Set of neighbors of node v

Phase 3 → Propagation Rule

General Message-Passing Framework

$$h_v^k = \sigma(W^k.AGGREGATE^k(\{h_u^{k-1}: u\epsilon N(v)\})) \quad (9)$$

Phase 4 → Output Layers

After the propagation layer, the node representation is updated and ready to be used in predictions

C. Data Classification and Analysis

The primary data consists of customer interactions, product transactions, product details, and relationship networks connecting customers and products. We evaluate customer contact data, including chats, reviews, and feedback, using the BERT model to gain a comprehensive understanding of customer preferences and wants [35]. We utilize data on product transactions, including client purchases, to construct graphs that illustrate the correlation between customers and products. Product information, such as category and price, enhances the feature nodes in the network. Graph neural networks (GNNs) subsequently examine the graph representing the link between customers and products to identify intricate relationship patterns and deliver precise product suggestions. The analysis procedure entails extracting features, gathering information from neighboring nodes in the network, and using a non-linear activation function to generate node representations [36] that can be seen in Table I.

TABLE I. NODE REPRESENTATIONS

Data Type	Description	Analysis Method	Objective
Customer Interaction	Customer conversations, reviews and feedback.	BERT	Understand customer preferences and needs
Product Transactions	Information about product purchases by customers.	Transaction Analysis and GNSs	Building customer-product relationship graphs and pattern analysis.
Product Information	Product details include category and price.	Data Enrichment & GNNs	Node features in a graph.
Relationship Graph	Graph representation of the relationship between customers and products based on transactions.	Graph Neural Networks (GNNs)	Detect relationship patterns and product recommendations.

III. RESULT AND DISCUSSION

Data in Tables II, III and IV include customer interactions, product transactions, product information, as well as additional features used in the customer-product relationship graph and data in Tables V, VI and VII include information for analysis, from customer interactions to product transactions and relationship graphs. Each parameter in the data, such as Text Length, Transaction Frequency, Edge Weight, and Node Features, provides context for understanding how customers interact with products on an e-commerce platform. This data provides the basis for in-depth analysis and implementation of algorithms to improve Business Intelligence on e-commerce platforms.

TABLE II. CUSTOMER INTERACTION DATA

Interaction ID	Customer ID	Text
1	101	"What is the return policy for this item?"
2	102	"Do you have this product in size M?"
3	103	"How long does shipping take?"
4	104	"Can I change my order after placing it?"
5	105	"Are there any discounts available?"
6	106	"I received a damaged item, what should I do?"
7	107	"Is this product available in other colours?"
8	108	"I need help with tracking my order."
9	109	"Can I get a refund for this purchase?"
10	110	"What is the warranty period for this product?"

TABLE III. PRODUCT TRANSACTION DATA

Transaction ID	Customer ID	Product ID	Quantity	Transaction Frequency	Purchase Variety
1	101	1001	2	3	2
2	102	1002	1	1	1
3	103	1003	3	4	2
4	104	1004	1	2	1
5	105	1005	2	3	1
6	101	1006	1	3	2
7	102	1007	1	1	1
8	103	1008	1	4	2
9	104	1009	1	2	1
10	105	1010	2	3	1

TABLE IV. PRODUCT INFORMATION DATA

Product ID	Product Name	Category	Price	Stock Availability	Product Ratings
1001	T-shirt	Clothing	20	In Stock	4.5
1002	Running Shoes	Footwear	50	In Stock	4.7
1003	Leather Jacket	Clothing	100	Out of Stock	4.6
1004	Smartwatch	Electronics	150	In Stock	4.8
1005	Wireless Earbuds	Electronics	30	In Stock	4.4
1006	Backpack	Accessories	40	In Stock	4.3
1007	Sunglasses	Accessories	25	In Stock	4.5
1008	Laptop	Electronics	500	In Stock	4.9
1009	Office Chair	Furniture	200	Out of Stock	4.6
1010	Water Bottle	Accessories	15	In Stock	4.2

TABLE V. NODES (CUSTOMERS AND PRODUCTS)

Node ID	Type	Age	Gender	Category	Price	Product Ratings
101	Customer	25	Male	-	-	-
102	Customer	30	Female	-	-	-
103	Customer	22	Male	-	-	-
104	Customer	28	Female	-	-	-
105	Customer	35	Male	-	-	-
1001	Product	-	-	Clothing	20	4.5
1002	Product	-	-	Footwear	50	4.7
1003	Product	-	-	Clothing	100	4.6
1004	Product	-	-	Electronics	150	4.8
1005	Product	-	-	Electronics	30	4.4

TABLE VI. EDGES (TRANSACTIONS)

Source Node	Target Node	Transaction ID	Quantity	Edge Weight
101	1001	1	2	40
102	1002	2	1	50
103	1003	3	3	300
104	1004	4	1	150
105	1005	5	2	60

TABLE VII. NODE FEATURES

Node ID	Feature 1	Feature 2
101	Age: 25	Gender: Male
102	Age: 30	Gender: Female
103	Age: 22	Gender: Male
104	Age: 28	Gender: Female
105	Age: 35	Gender: Male
1001	Category: Clothing	Price: 20
1002	Category: Footwear	Price: 50
1003	Category: Clothing	Price: 100
1004	Category: Electronics	Price: 150
1005	Category: Electronics	Price: 30

A. Data Processing Results with BERT

Most customers ask about return policies, product availability, and shipping information. This indicates areas need to be clarified on e-commerce sites to reduce customer service burden. Complaints about damaged goods and requests for refunds were most common, indicating a need to improve delivery quality and return policies. BERT successfully classifies customer interaction texts into relevant categories with a high level of confidence. This allows e-commerce platforms to automatically handle various types of customer inquiries or complaints more efficiently, shown in Table VIII.

TABLE VIII. CUSTOMER INTERACTION ANALYSIS (TEXT CLASSIFICATION)

Interaction ID	Customer ID	Text	Predicted Class	Confidence Score
1	101	"What is the return policy for this item?"	Inquiry: Return Policy	0.98
2	102	"Do you have this product in size M?"	Inquiry: Product Availability	0.95
3	103	"How long does shipping take?"	Inquiry: Shipping Information	0.97
4	104	"Can I change my order after placing it?"	Inquiry: Order Modification	0.96
5	105	"Are there any discounts available?"	Inquiry: Discounts	0.93
6	106	"I received a damaged item, what should I do?"	Complaint: Damaged Item	0.99
7	107	"Is this product available in other colours?"	Inquiry: Product Availability	0.95
8	108	"I need help with tracking my order."	Inquiry: Order Tracking	0.94
9	109	"Can I get a refund for this purchase?"	Request: Refund	0.97
10	110	"What is the warranty period for this product?"	Inquiry: Warranty Information	0.96

B. *Data Processing Results with GNNs*

GNNs identify purchasing patterns and provide relevant product recommendations for each customer. These recommendations are based on historical relationships between products frequently purchased together by other customers with similar purchasing patterns as in the Table IX.

TABLE IX. CUSTOMER SEGMENTATION ANALYSIS

Segment	Characteristics	Actionable Insights
High-Value Customers	Customers who buy products at high prices	Focus on personalizing premium product offerings
Frequent Buyers	Customers with high purchasing frequency	Loyalty offers or discounts based on volume
Category Loyalists	Customers who tend to buy from one category	Promote related products in the same category

IV. VALIDATION

A. *Performance Evaluation of BERT*

BERT is used for customer interaction text classification. Performance evaluation is measured using several metrics [37], [38].

$$Accuracy = (\frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}) \tag{10}$$

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Positives\ (FP)} \tag{11}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Negatives\ (FN)} \tag{12}$$

$$F1-Score = 2\ \times \frac{Precision\ \times Recall}{Precision+Recall} \tag{13}$$

Table X shows BERT provides excellent results in text classification and sentiment analysis, with high scores in all key metrics. This reflects BERT's ability to effectively understand and process customer interaction texts for customer service automation and data-driven decision making in e-commerce platforms.

TABLE X. BERT PERFORMANCE EVALUATION

Evaluation Metrics	Value	Interpretation
Accuracy	97%	97% of BERT's total predictions were correct, demonstrating a high level of accuracy in customer interaction text classification.
Precision	0.95	95% of the positive predictions made by BERT were true positive, indicating that this model has a low false positive rate.
Recall	0.96	BERT successfully identified 96% of all existing positive cases, indicating that the model is very good at detecting the correct classes.
F1-Score	0.96	A harmonious combination of Precision and Recall, showing a good balance between the two.
Sentiment Analysis Accuracy	93%	BERT was able to classify sentiment with 93% accuracy, demonstrating reliability in determining whether text is positive, negative, or neutral.

B. *Performance Evaluation of GNNs*

GNNs are used for tasks such as product recommendation and customer-product relationship graph analysis. Its performance evaluation is measured using metrics such as Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) [39]:

MAP

$$MAP = \frac{1}{2}\sum_{i=1}^{n} AP(i) \tag{14}$$

NDCG

$$DCG_p = \sum_{i=1}^{p} \frac{2_i^{rel}-1}{\log_2(i+1)} \tag{15}$$

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{16}$$

Precision@K

$$Precision@K = \frac{Number\ of\ relevant\ items\ in\ K\ recommendations}{K} \tag{17}$$

TABLE XI. GNNS PERFORMANCE EVALUATION

Evaluation Metrics	Value	Interpretation
Mean Average Precision (MAP)	0.92	92% of the products recommended by GNNs are relevant and match customer preferences.
Normalized Discounted Cumulative Gain (NDCG)	0.88	Relevant recommendations tend to appear at the top of the recommendation list, increasing the likelihood of a product being chosen by a customer.
Precision@5	0.93	93% of the top five recommendations are relevant to customers.
Precision@10	0.89	89% of the top ten recommendations are relevant to customers.
Coverage	85%	GNNs cover 85% of the products in the catalogue, ensuring diverse and varied recommendations.
Hit Rate	0.90	90% of all customers find at least one relevant product in their recommendation list.

V. CONCLUSION

This research reveals how the use of BERT (Bidirectional Encoder Representations from Transformers) and Graph Neural Networks (GNNs) can improve business intelligence (BI) capabilities on e-commerce platforms. Through in-depth analysis, BERT was proven to be effective in classifying customer interaction texts with an accuracy rate of 97% and was able to perform sentiment analysis with an accuracy rate of 93%. This enables automation in the management of customer inquiries and complaints, directly increasing customer service efficiency. GNNs show good performance in providing relevant product recommendations, with a mean average precision (MAP) of 0.92 and a normalized discounted cumulative gain (NDCG) of 0.88. GNNs also managed to cover 85% of the products in the catalog, ensuring that the recommendations provided were varied. Customer segmentation generated by GNNs allows the identification of segments such as high-value customers and frequent buyers, which can be targeted with more precise marketing strategies, increasing campaign effectiveness and customer loyalty.

REFERENCES

[1] Alsmadi, A. Shuhaiber, M. Al-Okaily, A. Al-Gasaymeh, and N. Alrawashdeh, “Big data analytics and innovation in e-commerce: current insights and future directions,” Journal of Financial Services Marketing, May 2023, doi: 10.1057/s41264-023-00235-7.

[2] R. B. Y. Syah, H. Satria, M. Elveny, and M. K. M. Nasution, “Complexity prediction model: a model for multi-object complexity in consideration to business uncertainty problems,” Bulletin of Electrical Engineering and Informatics, vol. 12, no. 6, pp. 3697–3705, Dec. 2023, doi: 10.11591/eei.v12i6.5380.

[3] Y. Yang, N. Chen, and H. Chen, “The Digital Platform, Enterprise Digital Transformation, and Enterprise Performance of Cross-Border E-Commerce—From the Perspective of Digital Transformation and Data Elements,” Journal of Theoretical and Applied Electronic Commerce Research, vol. 18, no. 2, pp. 777–794, Mar. 2023, doi: 10.3390/jtaer18020040.

[4] Z. Yu et al., “Embedding text-rich graph neural networks with sequence and topical semantic structures,” Knowl Inf Syst, vol. 65, no. 2, pp. 613–640, Feb. 2023, doi: 10.1007/s10115-022-01768-4.

[5] S. Fan et al., “BOMGraph: Boosting Multi-scenario E-commerce Search with a Unified Graph Neural Network,” in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, New York, NY, USA: ACM, Oct. 2023, pp. 514–523. doi: 10.1145/3583780.3614794.

[6] T. Jiang, W. Sun, and M. Wang, “MSGAT-Based Sentiment Analysis for E-Commerce,” Information, vol. 14, no. 7, p. 416, Jul. 2023, doi: 10.3390/info14070416.

[7] I. Karabila, N. Darraz, A. EL-Ansari, N. Alami, and M. EL Mallahi, “BERT-enhanced sentiment analysis for personalized e-commerce recommendations,” Multimed Tools Appl, vol. 83, no. 19, pp. 56463–56488, Dec. 2023, doi: 10.1007/s11042-023-17689-5.

[8] Y. Xiong, N. Wei, K. Qiao, Z. Li, and Z. Li, “Exploring Consumption Intent in Live E-Commerce Barrage: A Text Feature-Based Approach Using BERT-BiLSTM Model,” IEEE Access, vol. 12, pp. 69288–69298, 2024, doi: 10.1109/ACCESS.2024.3399095.

[9] W. Chang and M. Zhu, “Sentiment analysis method of consumer comment text based on BERT and hierarchical attention in e-commerce big data environment,” Journal of Intelligent Systems, vol. 32, no. 1, Dec. 2023, doi: 10.1515/jisys-2023-0025.

[10] S. Zhou and N. S. Hudin, “Advancing e-commerce user purchase prediction: Integration of time-series attention with event-based timestamp encoding and Graph Neural Network-Enhanced user profiling,” PLoS One, vol. 19, no. 4, p. e0299087, Apr. 2024, doi: 10.1371/journal.pone.0299087.

[11] L. Zheng, Z. Li, J. Gao, Z. Li, J. Wu, and C. Zhou, “Domain Adaptation for Anomaly Detection on Heterogeneous Graphs in E-Commerce,” 2023, pp. 304–318. doi: 10.1007/978-3-031-28238-6_20.

[12] G. Xv et al., “E-commerce Search via Content Collaborative Graph Neural Network,” in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2023, pp. 2885–2897. doi: 10.1145/3580305.3599320.

[13] Z. Wen, “Generalizing Graph Neural Network across Graphs and Time,” in Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, New York, NY, USA: ACM, Feb. 2023, pp. 1214–1215. doi: 10.1145/3539597.3572986.

[14] X. Zhang, F. Guo, T. Chen, L. Pan, G. Beliakov, and J. Wu, “A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research,” Journal of Theoretical and Applied Electronic Commerce Research, vol. 18, no. 4, pp. 2188–2216, Dec. 2023, doi: 10.3390/jtaer18040110.

[15] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, “A review of sentiment analysis: tasks, applications, and deep learning techniques,” Int J Data Sci Anal, Jul. 2024, doi: 10.1007/s41060-024-00594-x.

[16] X. Guo et al., “Intelligent online selling point extraction and generation for e-commerce recommendation,” AI Mag, vol. 44, no. 1, pp. 16–29, Mar. 2023, doi: 10.1002/aaai.12083.

[17] P. Radanliev, “Artificial intelligence: reflecting on the past and looking towards the next paradigm shift,” Journal of Experimental & Theoretical Artificial Intelligence, pp. 1–18, Feb. 2024, doi: 10.1080/0952813X.2024.2323042.

[18] T. Vo, “A Novel Semantic-Enhanced Text Graph Representation Learning Approach through Transformer Paradigm,” Cybern Syst, vol. 54, no. 4, pp. 499–525, May 2023, doi: 10.1080/01969722.2022.2067632.

[19] C. Gao et al., “A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions,” ACM Transactions on Recommender Systems, vol. 1, no. 1, pp. 1–51, Mar. 2023, doi: 10.1145/3568022.

[20] D. Wu, Q. Wang, and D. L. Olson, “Industry classification based on supply chain network information using Graph Neural Networks,” Appl Soft Comput, vol. 132, p. 109849, Jan. 2023, doi: 10.1016/j.asoc.2022.109849.

[21] E. E. Kosasih, F. Margaroli, S. Gelli, A. Aziz, N. Wildgoose, and A. Brintrup, “Towards knowledge graph reasoning for supply chain risk management using graph neural networks,” Int J Prod Res, vol. 62, no. 15, pp. 5596–5612, Aug. 2024, doi: 10.1080/00207543.2022.2100841.

[22] H. Brama, L. Dery, and T. Grinshpoun, “Evaluation of neural networks defenses and attacks using NDCG and reciprocal rank metrics,” Int J Inf Secur, vol. 22, no. 2, pp. 525–540, Apr. 2023, doi: 10.1007/s10207-022-00652-0.

[23] C. Wang et al., “An empirical evaluation of technology acceptance model for Artificial Intelligence in E-commerce,” Heliyon, vol. 9, no. 8, p. e18349, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18349.

[24] S. J, Ch. Gangadhar, R. K. Arora, P. N. Renjith, J. Bamini, and Y. devidas Chincholkar, “E-commerce customer churn prevention using machine learning-based business intelligence strategy,” Measurement: Sensors, vol. 27, p. 100728, Jun. 2023, doi: 10.1016/j.measen.2023.100728.

[25] M. Azmi, A. Mansour, and C. Azmi, “A Context-Aware Empowering Business with AI: Case of Chatbots in Business Intelligence Systems,” Procedia Comput Sci, vol. 224, pp. 479–484, 2023, doi: 10.1016/j.procs.2023.09.068.

[26] M. Elveny, R. B. Y. Syah, and M. K. M. Nasution, “An boosting business intelligent to customer lifetime value with robust M-estimation,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 13, no. 2, p. 1632, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1632-1639.

[27] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, “BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis,” Appl Soft Comput, vol. 151, p. 111112, Jan. 2024, doi: 10.1016/j.asoc.2023.111112.

[28] K. Taneja, J. Vashishtha, and S. Ratnoo, “Transformer Based Unsupervised Learning Approach for Imbalanced Text Sentiment Analysis of E-Commerce Reviews,” Procedia Comput Sci, vol. 235, pp. 2318–2331, 2024, doi: 10.1016/j.procs.2024.04.220.

[29] M. Elveny, M. K. M. Nasution, and R. B. Y. Syah, “A Hybrid Metaheuristic Model for Efficient Analytical Business Prediction,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140848.

[30] S. Gheewala, S. Xu, S. Yeom, and S. Maqsood, “Exploiting deep transformer models in textual review based recommender systems,” Expert Syst Appl, vol. 235, p. 121120, Jan. 2024, doi: 10.1016/j.eswa.2023.121120.

[31] K. Sharma et al., “A Survey of Graph Neural Networks for Social Recommender Systems,” ACM Comput Surv, vol. 56, no. 10, pp. 1–34, Oct. 2024, doi: 10.1145/3661821.

[32] X. Li, L. Sun, M. Ling, and Y. Peng, “A survey of graph neural network based recommendation in social networks,” Neurocomputing, vol. 549, p. 126441, Sep. 2023, doi: 10.1016/j.neucom.2023.126441.

[33] Y. Mahendra and B. Bolla, “Unveiling the power of knowledge graph embedding in knowledge aware deep recommender systems for e-commerce: A comparative study,” Procedia Comput Sci, vol. 235, pp. 1364–1375, 2024, doi: 10.1016/j.procs.2024.04.128.

[34] G. Stalidis et al., “Recommendation Systems for e-Shopping: Review of Techniques for Retail and Sustainable Marketing,” Sustainability, vol. 15, no. 23, p. 16151, Nov. 2023, doi: 10.3390/su152316151.

[35] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, “A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions,” J Big Data, vol. 11, no. 1, p. 18, Jan. 2024, doi: 10.1186/s40537-023-00876-4.

[36] Y. Chen et al., “SP-GNN: Learning structure and position information from graphs,” Neural Networks, vol. 161, pp. 505–514, Apr. 2023, doi: 10.1016/j.neunet.2023.01.051.

[37] A. Bello, S.-C. Ng, and M.-F. Leung, “A BERT Framework to Sentiment Analysis of Tweets,” Sensors, vol. 23, no. 1, p. 506, Jan. 2023, doi: 10.3390/s23010506.

[38] R. B. Y. Syah, R. Muliono, M. Akbar Siregar, and M. Elveny, “An efficiency metaheuristic model to predicting customers churn in the business market with machine learning-based,” IAES International Journal of Artificial Intelligence (IJ-AI), vol. 13, no. 2, p. 1547, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1547-1556.

[39] H. Yang and T. Gonçalves, “Field features: The impact in learning to rank approaches,” Appl Soft Comput, vol. 138, p. 110183, May 2023, doi: 10.1016/j.asoc.2023.110183.

Elevating Grape Detection Precision and Efficiency with a Novel Deep Learning Model

Xiaoli Geng*, Yaru Huang, Yangxu Wang
Department of Network Technology, Guangzhou Institute of Software Engineering, Conghua, Guangdong, China

***Abstract*—In the domain of modern agricultural automation, precise grape detection in orchards is pivotal for efficient harvesting operations. This study introduces the Grapes Enhanced Feature Detection Network (GEFDNet), leveraging deep learning and convolutional neural networks (CNN) to enhance target detection capabilities specifically for grape detection in orchard environments. GEFDNet integrates an innovative Enhanced Feature Fusion Module (EFFM) into an advanced YOLO architecture, employing a 16x downsampling Backbone for feature extraction. This approach significantly reduces computational complexity while capturing rich spatial hierarchies and accelerating model inference, which is crucial for real-time object detection. Additionally, an optimized dual-path detection structure with an attention mechanism in the Neck enhances the model's focus on targets and robustness against dense grape detection and complex background interference, a common challenge in computer vision applications. Experimental results demonstrate that GEFDNet achieves at least a 3.5% improvement in mean Average Precision (mAP@0.5), reaching 89.4%. It also has a 9.24% reduction in parameters and a 10.35 FPS increase in frame rate compared to YOLOv9. This advancement maintains high precision while improving operational efficiency, offering a promising solution for the development of automated harvesting technologies. The study is publicly available at: https://github.com/YangxuWangamI/GEFDNet.**

Keywords*—Computer vision; deep learning; Convolutional Neural Networks (CNN); real-time object detection; dual-path detection structure

I. Introduction

Grapes, as deciduous vines of the Vitis genus, are celebrated as the "Queen of Fruits." They are not only rich in nutrients but also possess significant medicinal value, making them one of the most popular fruits globally [1]. In the field of agricultural automation, precise grape detection is key to improving harvesting efficiency and fruit quality. Although manual harvesting is still the mainstream method, it is inefficient and labor-dependent [2], creating an urgent need for automated solutions. Existing vision detection systems face challenges in complex orchard environments, such as changes in lighting, occlusions, and fruit overlapping, which limit their performance. Therefore, a robust detection model is crucial for robots to achieve target perception in complex vineyard scenarios [3].

To enhance the recognition ability and efficiency of deep learning models in orchard grape detection, the goal of this study is to develop a fast, parameter-reduced, and low-miss detection model for dense and occluded grape detection in orchards, named Grapes Enhanced Feature Detection Network (GEFDNet). At the same time, YOLOv9 [4], as the latest generation of the YOLO series, has demonstrated its excellent accuracy and speed in various general object detection tasks through optimized network architecture and detection algorithms. Despite this, applying YOLOv9 directly to grape detection tasks in orchards still faces specific challenges. In response to these challenges, the GEFDNet model targets grapes, innovatively designing a 16x downsampling Backbone network and proposing a new high-efficiency scale fusion module called the Enhanced Feature Fusion Module (EFFM) module, aiming to capture target feature information at a finer granularity. Applied to the main trunk and detection neck networks, it significantly reduces the model's computational burden and parameter volume, enabling GEFDNet to better adapt to the complex and variable agricultural environment.

In the experiments, to objectively and comprehensively evaluate model performance, this study conducted comparative experiments with seven other advanced methods, especially an in-depth performance evaluation against the benchmark model YOLOv9. Performance analysis results show that GEFDNet has increased the mean Average Precision (mAP@0.5) on the test dataset by at least 3.5%. Through visual analysis, the model's advantages in dealing with challenging complex scenes were further revealed. In addition, compared to YOLOv9, GEFDNet has reduced the parameter volume by about 9.24% and increased the frame rate (FPS) by 10.35. This series of data highlights the efficiency of GEFDNet in object detection tasks.

The main contributions of this paper are as follows:

- The design of the EFFM module, which enhances the accuracy and efficiency of target detection in images by providing a powerful feature extraction and fusion mechanism for grape target detection tasks.
- The innovative design of a 16x downsampling Backbone network addresses the prolonged training times and weight redundancy issues associated with YOLOv9's detection neck and auxiliary branch structures.
- Optimization of the main detection neck design overcomes the difficulty of detecting occlusions and dense targets, reduces the model's computational burden and parameter count, and ensures high detection accuracy in resource-constrained environments.
- Comparative experiments with seven other popular detection models demonstrate GEFDNet's advantages in lightweight design, further verifying its effectiveness and feasibility.

*Corresponding Author

The layout of this paper is as follows: Section I (this section) introduces the prominent issues in the research field and the motivation behind the model design. Section II summarizes the research background and the challenges of existing technology. Section III provides a detailed introduction to the characteristics of the dataset and the principles of model design. Section IV includes the experimental process, model performance comparison, and result analysis. Section V discusses the research findings and proposes future research plans. Section VI summarizes the entire paper.

II. Background

With the rapid development of computing power and deep learning techniques [5], convolutional neural networks have attracted attention across various industrial sectors. Object detection models integrated with deep learning are being increasingly applied to agricultural studies, including fruit recognition [6, 7], disease detection [8, 9, 10], and yield estimation [11, 12]. Currently, deep learning-based fruit object detection models are mainly divided into two categories: one category is the region proposal-based two-stage detection models, such as Faster R-CNN [13] and Spatial Pyramid Pooling Network (SPP-Net) [14], which have high detection accuracy but are slower due to their two-stage nature. The other category is the regression-based single-stage detection models, such as SSD [15], YOLO [16, 17], and CenterNet [18], which maintain high detection accuracy while offering faster detection speeds and stronger real-time capabilities. Due to the high demand for detection speed in most tasks, especially in real-time scenarios, single-stage algorithms have more advantages in practical applications.

In recent years, research on deep learning models for grape detection has been continuously emerging. The latest research from Wu et al., 2024 [19], uses Adaptive Training Sample Selection (ATSS) as a label matching strategy to improve the quality of positive samples and address the challenge of detecting grape stems with similar colors. They utilize the Wise-IoU (Sequential Evidence for Intersection over Union) loss function with weighted interpolation to overcome the limitations of CIoU, which does not consider the geometric properties of targets, thus improving detection efficiency. Behera et al., 2023 [20], proposed an FR-CNN algorithm for plant fruit prediction using Intersection over Union (IoU), achieving an 89% accuracy rate in fruit yield estimation. Aguiar et al., 2021 [21], used deep learning models for grape cluster detection with an average accuracy of 66.96%. Pereira et al., 2019 [22], introduced a grape detection method based on the AlexNet neural network architecture, achieving a high average accuracy of 77.30%. Rong et al., 2024 [23], proposed a grape cluster detection method based on Spatial-to-Depth Convolution (STD-Conv) and Simple Attention Mechanism (SimAM), expanding the dataset through data augmentation technology, enabling the improved YOLOX model to achieve an 88.40% average accuracy in grape cluster detection. Marani et al., 2020 [24], proposed a vehicle-mounted RGB-D camera system for grape recognition using a deep learning framework. Sozzi et al., 2021 [25], used the YOLOv4 model for the detection and counting of grape clusters, achieving an accuracy rate of 48.90%. Li et al., 2021 [26], proposed an improved YOLOv4-tiny model, YOLO-Grape, to address the issue of unrecognizable accuracy caused by complex background scenes such as shadows and overlaps.

III. Materials and Methods

This section provides a detailed description of the datasets used in the experiments and elucidates the design principles, innovations, and activation functions of the GEFDNet model.

A. Datasets

To validate the effectiveness and adaptability of the proposed method, the experiments in this study are conducted using the Embrapa Wine Grape Instance Segmentation Dataset (Embrapa WGISD) [27]. This dataset was created for the application of object detection and instance segmentation techniques in image monitoring and field robot vision in vineyards, containing instance images of five different grape varieties. These images were captured under natural field conditions, encompassing various postures, lighting and focus conditions, as well as genetic and phenotypic variations such as shape, color, and compactness.

The images of the dataset were taken using a Canon EOS REBEL T3i DSLR camera and Motorola Z2 Play smartphone at the Guaspari Winery in São Paulo, Brazil. The image resolution was adjusted to a width of 2,048 pixels to balance image detail and processing time. The dataset was annotated with rectangular bounding boxes to identify grape clusters using the LabelImg tool [28], comprising a total of 300 images with 4,432 annotated grape clusters.

In summary, experiments conducted on the Embrapa WGISD dataset will provide a comprehensive evaluation of the universality and effectiveness of the proposed method. However, the orchard environment is challenging, as depicted in Fig. 1, which categorizes the dataset's characteristics and key detection challenges into four types, including densely packed arrangements of grapes, occlusions by leaves or trunks, complex backgrounds, and varying lighting conditions.

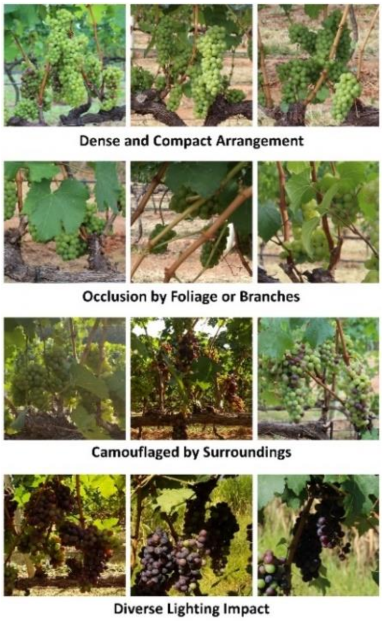

Dense and Compact Arrangement
Occlusion by Foliage or Branches
Camouflaged by Surroundings
Diverse Lighting Impact

Fig. 1. Four typical challenges in dataset images.

B. Model Construction

When applying neural networks for grape detection in orchard environments, numerous factors must be considered. To address these, this study introduces the Grapes Enhanced Feature Detection Network (GEFDNet), a novel high-precision, low-complexity grape detection model for orchard environments. The model adopts a Backbone-Neck structure and integrates the proposed Enhanced Feature Fusion Module (EFFM). The Backbone is responsible for extracting key features from the input images, employing a deep convolutional neural network to ensure the capture of rich spatial hierarchical information while reducing computational complexity. The Neck features a dual-path detection structure [4], including the Main Branch and the Auxiliary Branch, which further process target features, providing additional feature fusion and contextual information through parallel processing paths, enhancing the model's robustness against complex environmental variations. The architectural framework of GEFDNet is illustrated in Fig. 2, with detailed module structures, including EFFM, presented in Fig. 3. The auxiliary detection components are denoted with dashed lines in the diagrams. The following sections will detail their configuration specifics.

C. Enhanced Feature Fusion Module (EFFM)

Feature fusion is crucial for enhancing the model's generalization capability and detection accuracy in grape target detection tasks. By integrating features from different levels and scales, the model can more comprehensively understand image content, leading to more accurate identification and localization of grapes.

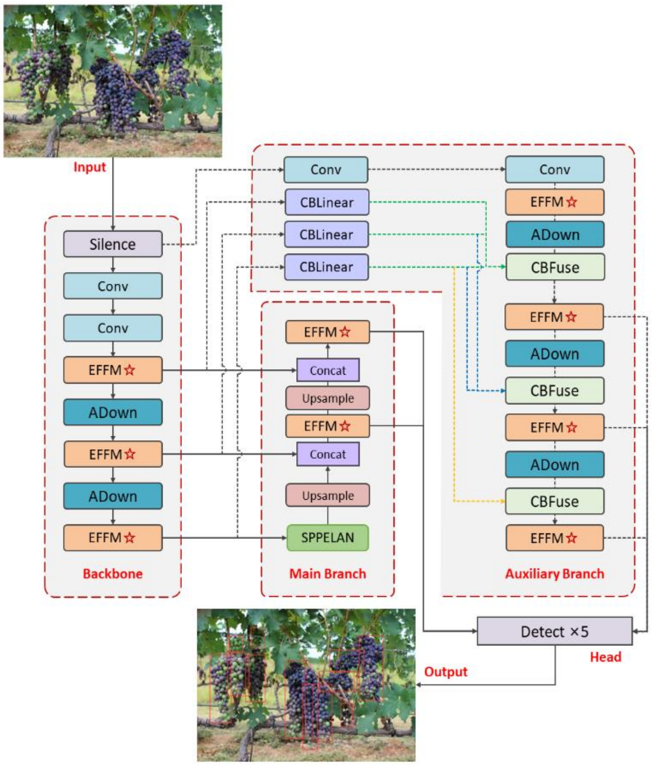


Fig. 2. Architecture of GEFDNet.

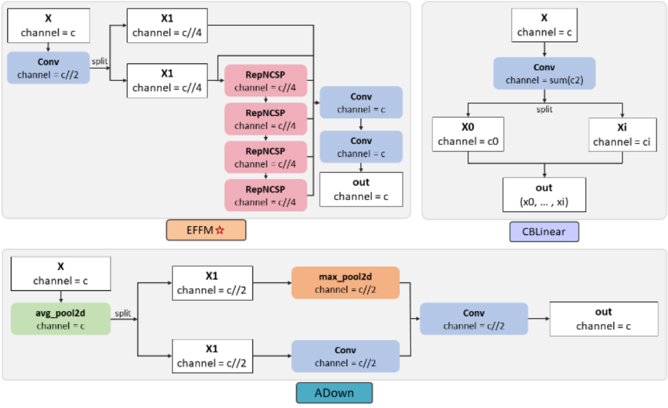


Fig. 3. Detailed module structures.

This paper introduces an innovative and efficient scale fusion module referred to as the Enhanced Feature Fusion Module (EFFM), as depicted in Fig. 3. After the input feature maps undergo a 1×1 convolution, the channel count is halved. The feature maps are evenly divided into two subsets of the same spatial size, denoted as X1, each with a quarter of the channel count of the input feature maps. One X1 subset is retained as is, while the other is processed through the RepNCSP module, further divided into four feature map subsets. These subsets undergo channel adjustments as they pass through each RepNCSP module, achieving efficient feature processing and fusion. For instance, a feature map may transition from channel count c to c//4, processed through a Conv layer, and further refined by the RepNCSP module, ultimately restoring to the original channel count c at the output. It is notable that each convolution can receive feature information from the preceding features, and for each feature branch after the RepNCSP module, the output has a larger receptive field and richer features compared to the unprocessed branch.

D. GEFDNet

1) Backbone Component: The Backbone component is tasked with feature extraction from input images. The architecture initiates with a Silence Module that serves as the preliminary processing unit, accepting raw image data and performing necessary preprocessing steps to maintain the initial feature information of the image, ensuring that the Main Branch and Auxiliary Branch can fully utilize this information for precise target localization. Subsequently, the model integrates multiple standard convolutional and pooling layers, each equipped with a 3×3 convolutional kernel, and employs stride-2 downsampling to reduce the dimensionality of the feature maps.

To enhance the model's nonlinear feature expression and integration capabilities, the Backbone network incorporates the innovative EFFM module, which facilitates deep integration of cross-layer features and effectively captures complex spatial hierarchies within the image. Notably, the core network of GEFDNet innovates in its downsampling strategy, adopting a 16x downsampling design that significantly reduces the loss of spatial resolution, enabling the model to excel in detecting small-sized targets and under varying lighting conditions.

Upon processing through the Backbone network, the model achieves 16x downsampling through four downsampling convolutional layers and three EFFM feature extraction layers, with the output feature map size being 1/16th of the original image, transitioning to multi-scale, multi-depth feature representations. This aids in reducing the model's computational load while retaining sufficient feature information to support subsequent detection tasks.

2) Neck Component: The Neck serves as the critical link between the Backbone and the detection Head, comprising both the Main Branch and the Auxiliary Branch. The Main Branch receives feature maps of varying downsampling levels output from the Backbone and initially processes them through an SPPELAN module to expand the receptive field, enhancing feature abstraction and expression. To further enhance the model's robustness, an attention mechanism is integrated into the Main Branch, allowing the model to adaptively focus on key image regions, such as grape edges and textures, thereby maintaining high accuracy despite challenges like occlusions and overlaps. Computational efficiency is also a significant consideration in the design of the Main Branch, where lightweight network components and depthwise separable convolutions are employed, effectively reducing the model's parameter count and computational complexity without compromising detection accuracy.

A series of upsampling and concatenation operations then follow, merging deep and shallow features from the Backbone to construct a multi-scale feature representation. Upsampling employs nearest neighbor interpolation to enlarge the feature map size, increasing resolution for more precise small target localization and facilitating fusion with larger feature maps. The concatenation operation integrates features from different levels, importantly, further processed through the innovatively designed EFFM feature extraction layer. Ultimately, the Main Branch outputs feature maps with high semantic information and spatial resolution, providing the Head with high-quality inputs for detection.

In addition to the Main Branch, the model's innovation lies in the design of the Auxiliary Branch, incorporating a reversible auxiliary branch design, utilizing cross-layer connections to directly extract and fuse features from the Backbone with high-level features from the Main Branch. Modules such as CBLinear and CBFuse are employed, unifying feature map sizes through cross-block connections and feature fusion strategies, followed by an addition operation to achieve multi-level auxiliary information fusion. This design not only enhances the model's detection capabilities for small targets and complex scenes but also reduces computational load through parallel processing, balancing computational efficiency with detection accuracy. Furthermore, the Auxiliary Branch serves as a regularization technique to prevent overfitting during model training.

3) Head Component: Upon the completion of the Neck's operations, the Head detection component receives five feature maps with varying spatial resolutions and semantic depths from the Neck. This enables the detection head to generate bounding boxes and aim frames of corresponding scales based on the feature map scales, overlaying the model's inference results onto the input image.

E. Activation Functions

In the realm of deep learning, activation functions play an indispensable role in dictating the performance and convergence rate of a model, determining the network's capacity to learn nonlinear relationships. Commonly utilized activation functions include Sigmoid Linear Unit (SiLU) [29], Rectified Linear Unit (ReLU) [30], and Leaky ReLU [31]. In constructing the GEFDNet model, this paper specifically selects SiLU as the activation function due to its combination of linear and nonlinear characteristics, which effectively enhances the network's nonlinear expressive power and learning efficiency. The definition of the SiLU activation function is presented in Eq. (1):

$$SiLU(x) = x \cdot \sigma(x) = x \cdot \frac{1}{1+e^{-x}} \tag{1}$$

The primary features of SiLU include its monotonicity, ensuring that as the input x increases, the output also increases, aiding in mitigating the vanishing gradient problem in deep networks. Its linearity for positive input values simplifies the nonlinear complexity in the positive range. SiLU's zero-centering characteristic, which outputs zero when $x = 0$, helps in centering the data, and when combined with batch normalization techniques, further improves the efficiency of model training. Additionally, SiLU boasts high computational efficiency as it involves only basic exponential and division operations, making it suitable for rapid execution on limited computational resources.

The Rectified Linear Unit (ReLU) activation function is one of the most popular nonlinear activation functions in deep learning. Its definition is straightforward and intuitive, expressed in Eq. (2):

$$ReLU(x) = max(0, x) \tag{2}$$

This function introduces nonlinearity by setting all negative values to zero, allowing only positive values to pass through, while maintaining computational efficiency. The main advantage of ReLU is its acceleration of the neural network training process; however, it also has some drawbacks, the most notable being the "dead ReLU" problem, where neurons corresponding to negative inputs may never activate, causing their weights to no longer update during training. In addition, ReLU's output is not zero-centered, which may affect the stability and convergence speed of the model during training.

Leaky ReLU is an improved version of ReLU, aiming to address the dead ReLU problem. Its formula is given in Eq. (3):

$$Leaky\ ReLU(x) = max(\alpha \cdot x, x) \tag{3}$$

Where α is a small positive number, typically taken as 0.01. Leaky ReLU introduces a small linear term when the input value is negative, ensuring that neurons with negative inputs still have a non-zero gradient, thus alleviating the problem of neuron death. This slight linear operation allows neurons with negative input values to still update their weights during the training process. However, Leaky ReLU introduces an additional hyperparameter α, which, if not chosen properly, may affect the network's convergence speed or lead to suboptimal model performance.

In deep learning, the choice of the appropriate activation function is crucial for model performance. Compared to ReLU and Leaky ReLU, SiLU offers several significant advantages, making it an ideal choice for grape detection models. SiLU's self-normalizing characteristic makes its output a linear transformation of the input in the positive range and approaches zero in the negative range, which helps stabilize network output and enhance generalization ability. Moreover, SiLU does not require additional parameters like Leaky ReLU, simplifying model training and hyperparameter adjustment. At the same time, the biological plausibility of SiLU further ensures the naturalness and efficiency of the activation pattern.

IV. Experience

This section provides an overview of the evaluation criteria and experimental design, followed by a presentation of the GEFDNet model's performance and a comparison with existing technologies. In addition to a comprehensive performance assessment using metrics such as Precision, Recall, and F1-score (F1), visual attention contrast experiments are introduced to further analyze the model's detection mechanisms. Utilizing Grad-CAM technology, the areas of focus when processing different grape samples are visualized, revealing GEFDNet's advantages in target recognition. Finally, the model's lightweight effect is evaluated, emphasizing its potential for efficient deployment in resource-constrained environments.

A. Experimental Conditions and Details

The study was conducted on a PC equipped with an AMD Ryzen 7 5800H 8-core processor (3.20 GHz) CPU and an NVIDIA GeForce GTX 3090 GPU. The software tools included the PyTorch 2.0.0 deep learning framework [32], CUDA version 11.8 parallel computing framework, and CUDNN version 8.9.5 deep neural network acceleration library. Standard data preprocessing methods were employed to fully leverage the dataset's information, including image scaling, cropping, and normalization, along with data augmentation techniques such as random flipping and rotation to enhance the model's generalization capability. Stochastic Gradient Descent (SGD) was used as the optimizer, with network training parameters set to an input size of 640×640 pixels, a batch size of 16, an initial learning rate of 0.01, a decay rate of 0.001, and a momentum parameter of 0.937. Considering convergence, 300 epochs of training were deemed sufficient for the model to reach a state of convergence.

B. Assessment of Model Performance

For the assessment of the proposed model within this study, the metrics of mAP@0.5, mAP@0.5:0.95, and F1-score were selected. The performance was compared against seven benchmark models, namely CenterNet [18], Faster R-CNN [13], SSD [33], FCOS [34], EfficientDet [35], YOLOv7-tiny [36], and YOLOv9 [4], utilizing the same dataset. The GEFDNet model underwent training and testing under identical conditions as the benchmark models, with evaluation based on Precision (P), Recall (R), F1-score (F1), and mean Average Precision (mAP).

Understanding the significance of these metrics requires clarity on the concepts of true positives (TP), false positives (FP), and false negatives (FN). TP represents the count of correctly identified samples, while FP denotes the instances of incorrect identifications. FN corresponds to the number of missed detections. The sum of "TP + FP" indicates the total inferred grape fruits by the model, and "TP + FN" accounts for the actual total count of fruits in the image.

Precision (P), which measures the accuracy of the model's positive predictions, is calculated as the ratio of true positive predictions to the total predicted positives, as illustrated in Eq. (4). This metric reflects the model's proficiency in accurately predicting positive outcomes.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Recall (R), depicted in Eq. (5), is the ratio of true positive predictions to the total actual positives, quantifying the model's effectiveness in capturing all actual positive instances.

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

The F1-score, represented by Eq. (6), is the harmonic mean of Precision and Recall, with a higher F1-score indicating a better balance between Precision and Recall.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \tag{6}$$

Average Precision (AP) signifies the area under the Precision-Recall (P-R) curve, calculated using an integral as shown in Eq. (7), a comprehensive indicator that takes into account both Precision and Recall.

$$mAP = \frac{1}{n}\sum_{1}^{n}\int P(dR) \tag{7}$$

The mAP@0.5 variant computes the average AP value at an Intersection over Union (IoU) threshold of 0.5 for all object categories. Furthermore, mAP@0.5:0.95 is determined to evaluate the model's performance across a spectrum of IoU thresholds, offering a stringent assessment of performance by representing the average mAP at various IoU thresholds ranging from 0.5 to 0.95 with increments of 0.05. The F1-score evaluates the methodology's performance by balancing the importance of accuracy and recall.

The results of the GEFDNet experiments are detailed in Table I, showcasing the performance of each model within the test dataset.

TABLE I. Quantitative Results on the Test Dataset

Model	P	R	F1	mAP@0.5	mAP@0.5:0.95
CenterNet	0.79	**0.85**	0.82	0.751	0.330
Faster R-CNN	0.79	0.82	0.81	0.815	0.398
SSD	0.26	0.59	0.36	0.239	0.095
FCOS	0.82	**0.85**	**0.84**	0.843	0.508
EfficientDet	0.07	0.57	0.12	0.095	0.018
YOLOv7-tiny	0.44	0.45	0.44	0.423	0.111
YOLOv9	**0.88**[a]	0.78	0.83	0.864	**0.601**
GEFDNet	**0.88**	0.81	**0.84**	**0.894**	0.596

[a]. The best performance is indicated in bold.

The experimental results demonstrate GEFDNet's significant advantage in target detection performance compared to other advanced methods. GEFDNet achieved an F1-score of 0.84, tying with YOLOv9 for the highest score, indicating an excellent balance between precision and recall. Particularly, in the key metric of mAP@0.5, GEFDNet surpassed all other models with a value of 0.894, including YOLOv9's 0.864, highlighting its superior detection accuracy at medium IoU thresholds. Furthermore, GEFDNet's comprehensive evaluation of model performance across different IoU thresholds from 0.5 to 0.95, mAP@0.5:0.95, achieved a value of 0.596, slightly trailing YOLOv9's 0.601, but this performance still ranks second among all compared models, showing its consistency and robustness across different IoU threshold ranges.

C. Visual Attention Contrast Experiment

To further examine the differences in detection effects between the proposed GEFDNet model and the existing YOLOv9 model, the Grad-CAM algorithm [37] was employed to visualize and compare the activation heat maps of the two models at different layers. Grad-CAM generates visual heat maps by combining the model's gradients and feature maps, revealing the visual areas the model focuses on when making specific predictions. This intuitive approach allows for a deeper understanding of the model's performance advantages and potential limitations, as shown in Fig. 4, which provides two sets of diagrams.

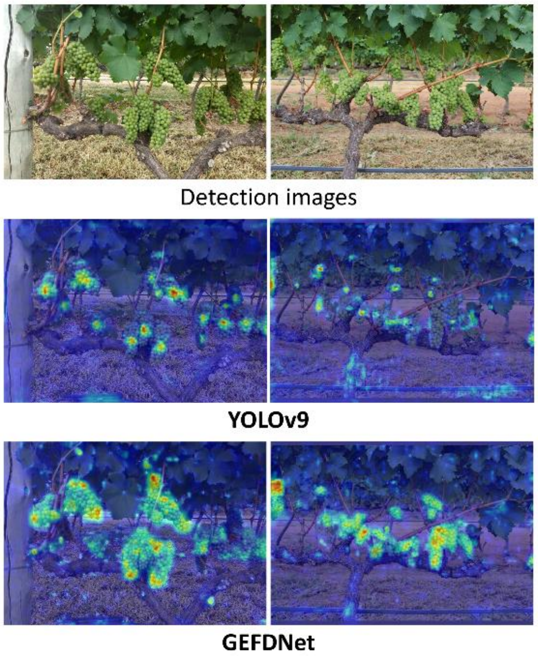


Fig. 4. Grad-CAM Visualization results.

When evaluating heat maps, focus on the following three key features to measure model performance:

- Clear boundary identification: The heat map should clearly depict the outline of the target object, demonstrating the model's high precision in spatial positioning.
- Noise suppression ability: The ideal heat map should not show excessive activation on image noise or irrelevant details, indicating that the model can effectively filter out unimportant information.
- Coverage of important features: The heat map should cover the key features of the target object, which are crucial for the object's recognition and classification.

Through the visualization results of Grad-CAM, it can be observed that GEFDNet has advantages in the above three aspects. Firstly, when localizing fruit targets, GEFDNet shows clearer boundaries and more focused attention, while YOLOv9, although able to recognize targets, also pays attention to the background, leading to scattered attention. Secondly, the heat map of GEFDNet performs better in suppressing image noise, indicating that it has stronger robustness when dealing with complex backgrounds and occlusions. Thirdly, the heat map of GEFDNet better covers the key features of grapes, aiding the model in more accurate recognition and classification of target objects.

D. Validation of Comprehensive Detection Capability

To verify the comprehensive detection capability of the improved GEFDNet model in different environments, we selected samples with varying lighting and density and conducted comparative experiments focusing on the model with similar performance to YOLOv9. We also conducted a detailed analysis of the visualization results of both models. This process revealed potential errors when detecting specific types of targets, thereby providing targeted guidance for subsequent model optimizations. Building on this, we further investigated specific conditions where the model might encounter difficulties. Fig. 5 illustrates the three most severe errors in the test dataset. The blue frames indicate the magnified portions of the images, while the yellow areas denote the regions of the grape clusters that were missed by the detection model.

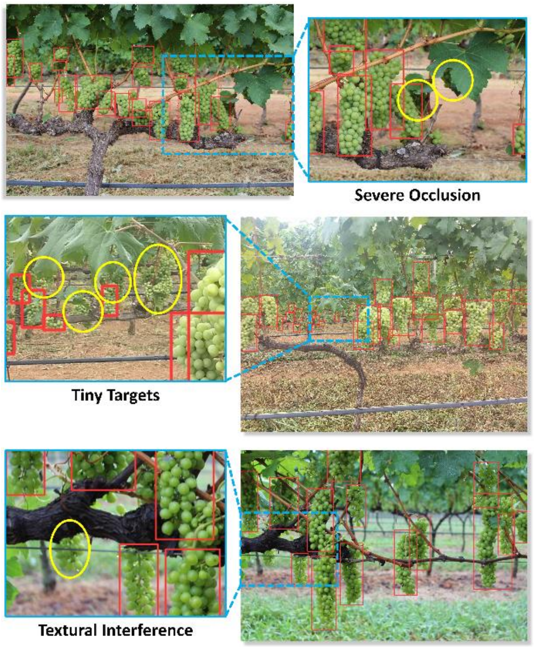


Fig. 5. The three most severe errors.

Firstly, there is the situation of extreme occlusion, where grapes are almost completely obscured by a large amount of foliage, with very little exposed. Under such extreme conditions, although GEFDNet performs better than YOLOv9 overall, there is a decline in detection accuracy. This is mainly because in the extreme occlusion environment, the information available for extracting effective features is greatly reduced. Despite the model's dual-path detection structure and EFFM module trying their best to capture features, it is still difficult to overcome the severe lack of information.

The second scenario is when facing extremely small grapes, the detection accuracy of GEFDNet decreases. This is because during the downsampling process of the model's Backbone network, the detailed features of very small grapes may be lost, making it difficult for the model to identify them accurately. The third scenario is in highly complex backgrounds, which include a large number of distractors similar in color and texture to grapes, as well as scenes with complex lighting and shadow variations. In such cases, although GEFDNet can filter out some irrelevant information, it is still interfered with by similar objects, resulting in a certain degree of false positives and false negatives. This indicates that the model's ability to resist interference needs to be further improved when dealing with highly complex backgrounds.

During the evaluation process, particular attention was given to the confidence threshold setting that yields the optimal mean Average Precision (mAP@0.5) across the entire test dataset. This strategy ensures the objectivity of the assessment while filtering for detections that the model is more confident in, effectively avoiding the impact of low-confidence predictions on the fairness of the evaluation. After the detection process, representative cases of false positives and false negatives were selected and visually presented, as shown in Fig. 6, where the blue areas indicate targets that were either missed or misidentified by the model.

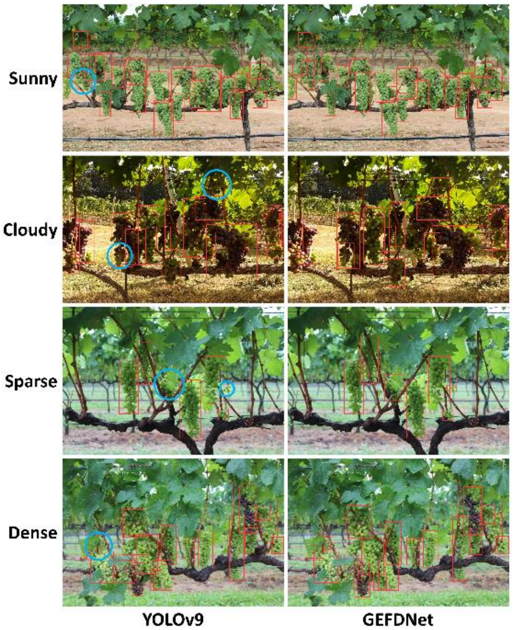


Fig. 6. Comparison of detection effects between YOLOv9 and GEFDNet models.

By carefully examining these results, the following typical errors and their causes can be identified: Under sunny conditions, both YOLOv9 and GEFDNet demonstrated good detection performance. However, YOLOv9 exhibited missed detections for small targets, likely due to insufficient feature extraction capabilities. Under overcast and uneven lighting conditions, YOLOv9 missed detections for grapes obscured by leaves and for small grapes beneath larger grapes. Furthermore, under varying densities, YOLOv9 consistently missed detections for grapes obscured by leaves, whether in sparse or densely clustered distributions. In contrast, GEFDNet effectively addressed these issues, particularly in detecting occluded and densely clustered grape clusters. These analyses not only reveal the limitations of YOLOv9 but also point the way for further optimization and development of GEFDNet.

E. Model Lightweighting

In the development of deep learning models, lightweighting is a critical optimization direction, particularly for application scenarios with constrained resources, such as edge device deployment in the agricultural sector. Lightweight models maintain sufficient detection accuracy while reducing computational load and storage requirements, thereby enhancing model operational efficiency and lowering deployment costs. In the experiments focused on model lightweighting, the recognition performance parameters of the YOLOv9 and GEFDNet models on the test dataset were compared, with the results presented in Table II. Key performance indicators such as Frames Per Second (FPS) were utilized, supplemented by the count of parameters and the size of the weight files to evaluate the models.

TABLE II. LIGHTWEIGHTING COMPARISON BETWEEN YOLOV9 AND GEFDNET MODELS

Model	F1	mAP@0.5	FPS	Parameters	Weights
YOLOv9	0.83	0.864	42.01	48.60 M	98.00 M
GEFDNet	**0.84**	**0.894**	**52.36**	**44.11 M**	**88.90 M**

The comparative experimental data clearly demonstrate the advantages of GEFDNet across multiple key indicators. Specifically, in terms of mAP@0.5, GEFDNet outperformed YOLOv9 with a score of 0.894 versus 0.864, marking a 3.5% improvement. This enhancement indicates that GEFDNet has achieved higher detection accuracy. Moreover, alongside the increase in precision, GEFDNet has also realized optimizations in lightweighting. The model's parameter volume has been reduced from 48.60M in YOLOv9 to 44.11M, and the weight file size has also been minimized from 98.00M to 88.90M. Furthermore, the detection frame rate (FPS) has been increased from 42.01 FPS of YOLOv9 to 52.36 FPS. These enhancements not only alleviate the storage burden but also imply that in resource-constrained environments, such as edge devices in the agricultural sector, GEFDNet can be deployed at a reduced cost. For application scenarios demanding high real-time performance, such as harvesting robots, these improvements are crucial for ensuring the system's response speed and processing capabilities.

V. Discussion and Future Work

The GEFDNet model introduced in this study offers a range of significant advantages in the field of grape detection in orchards. Firstly, the model integrates an innovative and efficient feature fusion module, the Enhanced Feature Fusion Module (EFFM), with a 16x downsampling Backbone network. This integration effectively balances detection accuracy and computational efficiency, reducing the model's parameter volume while increasing the frame rate, which is crucial for applications with high real-time requirements. Secondly, the introduction of the EFFM module enhances the model's ability to detect grapes against complex backgrounds and dense targets. Moreover, the high mean Average Precision (mAP) values demonstrated on the Embrapa WGISD dataset substantiate the model's excellent generalization and robustness.

Despite the positive outcomes of this study, there are certain limitations. It should be noted that the dataset used in this study is derived from a single crop species, and therefore, future testing and validation on more diverse datasets are required. Particularly, testing under poor lighting conditions and for extremely dense or very small-sized grapes should be conducted. Additionally, future research plans should expand and diversify the training datasets. Although the Embrapa WGISD dataset provides valuable resources for grape detection research, it has limitations, such as insufficient images of certain grape varieties, ripeness levels, and environmental conditions [38]. Moreover, to fully assess the potential of GEFDNet in real-world applications, future work will include real-time deployment assessments on actual hardware platforms like edge devices, drones, and agricultural robots. This aligns with the current trend in the field of agricultural automation towards evaluating practical application of models [39, 40]. This will help reveal the model's performance in resource-constrained environments and provide key insights for practical applications.

VI. Summary

Efficient and accurate detection of grapes in orchards has always been a challenging task. In this study, a high-precision, low-complexity deep learning model for grape detection in orchard environments, GEFDNet, was proposed, along with the innovative EFFM module integrated into the 16x downsampling Backbone network and optimized Neck structure. GEFDNet achieves model lightweighting while maintaining high accuracy, significantly enhancing the model's operational efficiency and practicality. The main achievements include a minimum 3.5% increase in mean Average Precision (mAP@0.5) on the test dataset, a reduction of about 9.24% in model parameter volume, and a 10.35 FPS increase in frame rate, validating the effectiveness of model lightweighting. Through Grad-CAM visualization analysis, GEFDNet's superior detection capabilities and precision in target recognition in complex scenarios have been demonstrated.

In summary, the development of the GEFDNet model not only promotes the advancement of agricultural automation technology but also provides a new perspective for the application of deep learning in complex scenarios. With the continuous deepening of future work, it is anticipated that GEFDNet will unleash greater potential in practical applications and make a substantial contribution to agricultural modernization.

Acknowledgement

This research was supported by grants from the “Guangzhou Institute of Software Engineering - Guangzhou Xinwei Internship Base Project" (No. SXJD20210101), the "Guangzhou Institute of Software Engineering - Guangdong Teddy Intelligent Technology Co., Ltd. Big Data Industry-Education Integration Practice Teaching Base Project" (No. SJJD202301), and the "Guangzhou Institute of Software Engineering - Guangdong Teddy Practice Teaching Base" for the 2024 Off-Campus Practice Teaching initiative.

References

[1] Alston, Julian M. and Olena Sambucci. "Grapes in the World Economy." The Grape Genome, Springer International Publishing, 2019, pp. 1-24.

[2] Rakhmatovich, Kholmuminov. "Fundamentals of Targeted Integrative Program Development for Rural Labor Market Growth in Surplus Regions." International Journal of Economics and Financial Issues, vol. 14, 2024, pp. 239-244.

[3] Jiqing, Chen et al. "Efficient and Lightweight Grape and Picking Point Synchronous Detection Model Based on Key Point Detection." Computers and Electronics in Agriculture, vol. 217, 2024, p. 108612.

[4] Wang, Chien-Yao et al. "Yolov9: Learning What You Want to Learn Using Programmable Gradient Information." ArXiv, vol. abs/2402.13616, 2024.

[5] LeCun, Yann et al. "Deep Learning." nature, vol. 521, no. 7553, 2015, pp. 436-444.

[6] Xu, Bo et al. "Apple Grading Method Design and Implementation for Automatic Grader Based on Improved Yolov5." Agriculture, vol. 13, no. 1, 2023, p. 124.

[7] Muhammad, Nur et al. "Evaluation of Cnn, Alexnet and Googlenet for Fruit Recognition." Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, 2018, pp. 468-475.

[8] Raskar, Soham. "Enhancing Agricultural Sustainability: Automated Crop Disease Detection through Image Processing Techniques." International Journal for Research in Applied Science and Engineering Technology, vol. 12, 2024, pp. 3925-3929.

[9] Lapates, Jovelin M. "Corn Crop Disease Detection Using Convolutional Neural Network (CNN) to Support Smart Agricultural Farming," International Journal of Engineering Trends and Technology, vol. 72, no. 6, 2024, pp. 195-203.

[10] S, Deepika et al. "Advancements in Agricultural Technology: A Comprehensive Review of Machine Learning and Deep Learning Approaches for Crop Management and Disease Detection." International Journal of Advanced Research in Science, Communication and Technology, 2024, pp. 111-120.

[11] Mimenbayeva, Aigul et al. "Applying Machine Learning for Analysis and Forecasting of Agricultural Crop Yields." Scientific Journal of Astana IT University, 2024, pp. 28-42.

[12] Virani, VB et al. "Machine Learning-Based Comparative Analysis of Weather-Driven Rice and Sugarcane Yield Forecasting Models." ORYZA-An International Journal of Rice, vol. 61, no. 2, 2024, pp. 150-159.

[13] Ren, S. et al. "Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks." IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, 2017, pp. 1137-1149.

[14] Purkait, Pulak et al. "Spp-Net: Deep Absolute Pose Regression with Synthetic Views." ArXiv, vol. abs/1712.03452, 2017.

[15] Liu, Wei et al. "Ssd: Single Shot Multibox Detector." Computer Vision – ECCV 2016, Translated by Bastian Leibe et al., Springer International Publishing, 2016, pp. 21-37.

[16] Terven, Juan R. et al. "A Comprehensive Review of Yolo Architectures in Computer Vision: From Yolov1 to Yolov8 and Yolo-Nas." Mach. Learn. Knowl. Extr., vol. 5, 2023, pp. 1680-1716.

[17] Redmon, Joseph et al. "You Only Look Once: Unified, Real-Time Object Detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 779-788.

[18] Zhou, Xingyi et al. "Objects as Points." ArXiv, vol. abs/1904.07850, 2019.

[19] Wu, Xinyu et al. "A Lightweight Grape Detection Model in Natural Environments Based on an Enhanced Yolov8 Framework." Frontiers in Plant Science, vol. 15, 2024.

[20] Behera, Santi Kumari et al. "Fruits Yield Estimation Using Faster R-Cnn with Miou." Multimedia Tools and Applications, vol. 80, no. 12, 2021, pp. 19043-19056.

[21] Aguiar, André Silva et al. "Grape Bunch Detection at Different Growth Stages Using Deep Learning Quantized Models." Agronomy, vol. 11, no. 9, 2021, p. 1890.

[22] Pereira, Carlos S. et al. "Deep Learning Techniques for Grape Plant Species Identification in Natural Images." Sensors, vol. 19, no. 22, 2019, p. 4850.

[23] Shuai Rong, Xinghai Kong Ruibo Gao Zhiwei Hu and Yang Hua. "Grape Cluster Detection Based on Spatial-to-Depth Convolution and Attention Mechanism." Systems Science \& Control Engineering, vol. 12, no. 1, 2024, p. 2295949.

[24] Marani, Roberto et al. "Deep Neural Networks for Grape Bunch Segmentation in Natural Images from a Consumer-Grade Camera." Precision Agriculture, vol. 22, 2020, pp. 387-413.

[25] Sozzi, Marco et al. "Grape Yield Spatial Variability Assessment Using Yolov4 Object Detection Algorithm." Precision Agriculture’21, Wageningen Academic Publishers, 2021, pp. 193-198.

[26] Huipeng, Li et al. "A Real-Time Table Grape Detection Method Based on Improved Yolov4-Tiny Network in Complex Background." Biosystems Engineering, vol. 212, 2021, pp. 347-359.

[27] Gebru, Timnit et al. "Datasheets for Datasets." Communications of the ACM, vol. 64, 2018, pp. 86-92.

[28] Tzutalin, D. (2022). LabelImg is a graphical image annotation tool and label object bounding boxes in images. URL https://github.com/tzutalin/labelImg.

[29] Elfwing, Stefan et al. "Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning." Neural networks, vol. 107, 2018, pp. 3-11.

[30] Glorot, Xavier et al. "Deep Sparse Rectifier Neural Networks." Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Translated by Gordon Geoffrey et al., vol. 15, PMLR, 2011, pp. 315-323.

[31] Xu, Bing et al. "Empirical Evaluation of Rectified Activations in Convolutional Network." ArXiv, vol. abs/1505.00853, 2015.

[32] Paszke, Adam et al. "Pytorch: An Imperative Style, High-Performance Deep Learning Library." Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., 2019, pp. 8026-8037.

[33] Bastian Leibe et al. "Ssd: Single Shot Multibox Detector." Springer International Publishing, Computer Vision – ECCV 2016, 2016, pp. 21-37.

[34] Tian, Z. et al. "Fcos: Fully Convolutional One-Stage Object Detection." 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9626-9635.

[35] Tan, Mingxing et al. "Efficientdet: Scalable and Efficient Object Detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781-10790.

[36] Wang, Chien-Yao et al. "Yolov7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." 2023, pp. 7464-7475.

[37] Selvaraju, Ramprasaath R. et al. "Grad-Cam: Visual Explanations from Deep Networks Via Gradient-Based Localization." 2017, pp. 618-626.

[38] Sivasubramanian, Arrun et al. "Object Detection under Low-Lighting Conditions Using Deep Learning Architectures: A Comparative Study." Advances in Data Science and Computing Technologies, 2023, pp. 269-276.

[39] Hert, Daniel et al. "Mrs Drone: A Modular Platform for Real-World Deployment of Aerial Multi-Robot Systems." Journal of Intelligent & Robotic Systems, vol. 108, no. 4, 2023, p. 64.

[40] Alibabaei, Khadijeh et al. "Real-Time Detection of Vine Trunk for Robot Localization Using Deep Learning Models Developed for Edge Tpu Devices." Future Internet, vol. 14, 2022, p. 199.

Vehicular Traffic Congestion Detection System and Improved Energy-Aware Cost Effective Task Scheduling Approach for Multi-Objective Optimization on Cloud Fog Network

Praveen Kumar Mishra, Amit Kumar Chaturvedi
Computer Application Deptt., Govt. Engg. College, Ajmer, BTU, Bikaner, India

***Abstract*—A current research area called fog computing aims to extend the advantages of cloud computing to network edges. Task scheduling is a crucial problem for fog device data processing since a lot of data from the sensor or Internet of Things layer is generated at the fog layer. This research suggested a vehicular traffic congestion detection model and an energy-aware cost effective task scheduling (ECTS) method in a cloud fog scenario. This research proposes an ECTS approach to allocate jobs to the fog nodes effectively. The recommended scheduling approach minimizes energy consumption and decreases expenses for time-sensitive real-time applications. The ECTS algorithm is implemented, and results are analysed using the iFogSim simulator. The proposed method minimizes energy consumption and cost. The suggested ECTS method is tested with five sets of inputs in this paper. The experiment's results show that an ECTS minimizes energy consumption in comparison to alternative algorithms. It also reduces the execution cost. The suggested approach outperforms both the Round-Robin (RR) and Genetic Algorithm techniques. According to the simulation results, the suggested algorithm reduced overall costs by 13.38% and energy usage by 6.59% compared to the Genetic Algorithm (GA). Compared to RR, the proposed method minimizes energy use by 13.76% and total costs by 18.46%.**

***Keywords*—IoT; fog computing; task scheduling; multi objective Model; iFogSim tool**

I. Introduction

One of the most crucial improvements in the realm of technology in the last several years is the Internet of Things (IoT) devices for the computation and exchange of information. IoT devices allow a wide range of items and equipment like sensors, cameras, automobiles, and connectivity with the Internet with smart devices like cell phones and laptops. Numerous applications and service offerings, such as latency minimization, control of traffic, and response time improvement, can be carried out as a result. Large volumes of data are produced through end devices as a result, which require supervision, processing, and analysis to generate appropriate data that will meet the user's objectives and aims. Moreover, the volume of data and a variety of required services and apps are expanding very quickly, demanding more computing power than even the most advanced smart devices can no longer match. The well-known cloud environment is a vast repository of resources that permits the universal ability to share and dynamically provide users with resources through virtualization procedures, which is one potential platform to aid in IoT improvements. By shifting resource and service-intensive jobs to a trustworthy computer environment, like the cloud, allowing smart devices to undertake basic tasks, limitations of current smart devices, such as enhancements, might be made to processing speed, capacity of storage, and resources required at the network. However, combining the use of clouds with the IoT creates further problems. It is anticipated that 50 billion IoT devices has been deployed in 2023. This figure will be increased to 35.8 billion in 2030, with the exponential rise in connected devices and cloud architectures that rely on traditional centralized processing features where storage and computational resources combined won't be enough to handle the demands of the Internet of Things devices burden. The main reason is that IoT gadgets and the cloud's infrastructure are quite far apart. The enormous amount of data of IoT devices send via the Internet to the cloud will strain the network's capacity and bandwidth, causing congestion, particularly near bottlenecks [4]. IoT applications are latency sensitive; therefore, a transmission delay reduces the Quality of Services (QoS), negatively impacting the user experience.

Fog computing [1], an innovative strategy of cloud computing first introduced by Cisco, has the potential to transform connecting the network's boundary to a distributed processing design that can accommodate the end devices services. By using fogging, clients may access computing and data storage power resources more conveniently by extending cloud computing to data-generating and data-receiving IoT gadgets; instead of moving all the processing to the CDS (Cloud Data Center), the Fog layer goals to process the maximum of the traffic load created by end devices nearby to the user's ranger of the network, called fog computing devices. Anywhere there is network connectivity, such as factories, shopping malls, electricity poles, railroads, inside of cars, etc., may use end devices. A fog node is any device with networking, computing, and storage capabilities. These devices include embedded servers, switches, routers, controllers, and security cameras. Requests are optimized for transmission time by putting resources near the network's edge, where the minimum amount of time needed for information to arrive at a point of processing larger-scale and delay-tolerant activities still be routed to the cloud layer. At the same time, smaller jobs or task requirements with low latency have to be given precedence to be handled by fog computing platforms deposited at fog nodes with limited processing capability. Ultimately, fogging and cloud computing combine to create the cloud fog scenario, a new paradigm for the computing environment. This innovative strategy has several benefits, which include latency minimization, minimize high network traffic, and minimizing power consumption. Balancing of load for jobs ensures that no resource remains idle while others are being used [1]. The security risks businesses face using cloud computing has decreased [2]. For enterprises deploying solution of big data on cloud infrastructure, is a crucial factor to consider.

The paper is presented in the following sequence. Specific literature review for task scheduling within the framework of cloud fog atmosphere is provided in section 2. The proposed Vehicular Traffic Congestion Detection (VTCD) System and ETS scheduling algorithm are presented in Section 3. Section 4 provides a simulation environment. Section 5 includes performance analysis and simulation outcomes. Section 6 present the conclusion and forthcoming scope at the end**.**

II. Literature Review

According to Jayasena et al. [3], scheduling of job necessitates optimizing two goals: minimizing cost and decreasing power utilization. The author designed a meta-heuristic Whale Optimization Algorithm (WOA) mapped procedure to explain the recommended system and calculate the outcomes in iFogSim against heuristic techniques like Particle Swarm Optimization (PSO) and RR and SJF. Xu et al. [4] explain the laxity-mapped precedence approach to build a scheduling of task order with a fair priority. According to the author, based on the ant colony system (ACO) algorithm, this strategy minimizes overall energy use.

Tan, et al. [5] proposed an energy-efficient approach and looked at a task scheduling issue along with time limit restrictions in instances when could exist distributed throughout heterogeneous assets, such as fog computing, and an energy-conscious algorithm capable of finding the best solution in a polynomial amount of time. Nikoui et al. [6] developed a genetic-based (CAGB) planning method that improves efficiency at a lower cost for real-world applications with tight deadlines. Its effectiveness is evaluated regarding system overload, expenses, and delay. Fellir Z. et al. [7] presented multi-agent-based planning method, the most important tasks are handled first to ensure that when a packet with the highest importance goes into the waiting line, the job is dealt with, without interfering with the least significant task's implementation if it is presently being run. Madej et al. [8] presented four scheduling schemes named NFCFS (Naive First Come First Serve) technique. The other three schemes are client fair, prioritized fair, and hybridization.

Abdel Basset, et al. [9] suggested method improved the effectiveness of the best outcome and justified the workload across the accessible simulated engines by using a meta-heuristic method and a shift modification technique. Yang et al. [10] presented the superior value efficiency and scale outcome set to tackle the two-parameter collaboration to minimize fog computing scheduling task difficulties. The results demonstrate how the suggested strategy performs better than conventional strategies regarding resource cost, overall job execution time, etc. Hoseiny et al. [11] suggested cost aware scheduling method reduces latency, computation costs, and communication costs for IoT inquiries while increasing the proportion of tasks that are finished earlier than the deadline. This algorithm is compared with genetic algorithm.

The Task Priority Resource Allocation (TPRA) algorithm was presented by Dang et al. [12]. This algorithm's primary goal is to minimize the average latency in the fog network's diverse environment. Abdel-Basset et al. [13] presented the multiple objective task scheduling technique. This algorithm aims to minimize the rate of carbon emissions, make-span, and energy consumption. The resource-aware-cost-efficient (RACE) scheduler was introduced by Arshed et al. [14]. This method distributes the incoming jobs to fog nodes function. This strategy pursues minimizing bandwidth use, maximize Fog Node (FN) utilization at the fog layer, and shorten application make span.

In a fog context, Singh et al.'s [15] hybrid swarm optimization using genetic algorithms (GA) reduces execution time and cost. Compared to GA and PSO, the workflow scheduling experimental result that is being provided is superior. The MGWO multi-objective optimization approach was introduced by Saif, et al. [16] multiple goals, including make-span, throughput, energy, and delay. This approach aims to ascertain the optimal strategy for work scheduling at the fog layer. The MGWO algorithm's experimental result outperforms the equivalent methods regarding power minimization and delay reduction. Zhang et al. [17] introduced the Enhanced Whale Optimization Algorithm (EWOA). It's a technique for scheduling tasks with multiple objectives in a cloud computing environment—a search strategy known as Levy's struggle in EWOA. The results of the various heuristic and meta-heuristic algorithms match the results of the EWOA experiment. EWOA performs better in cost reduction and energy use minimization than these existing algorithms. Alwabel et al. [18] offered a deadline and power-efficient job scheduling method in a fog computing network. This method aims to determine which jobs are crucial and prioritize them so that they may be finished at the fog layer. The simulator iFogSim is used for outcome analysis. The recommended approach outperforms earlier algorithms regarding deadline and energy consumption reduction.

The PEWO (Parallel Enhanced Whale Optimization) approach was created by Khan et al. [19] for task scheduling at the cloud computing layer. This meta-heuristic method aims to minimize make-span and execution time. In a heterogeneous cloud environment, tasks are assigned using the PEWO approach. The experimental result of PEWO is better represented by the random matrix particle swarm optimization (RMPSO). Ali et al. [20] presented DNSG, a task scheduling system, by dynamically assigning the task to a fog node; the proposed approach aims to reduce make-span and cost compared to modified GA. Balancing tasks is also one of the issue in cloud fog environment. Within the Cloud-Fog system [24-25], job scheduling aims to maximize benefits for either service End users are concerned about minimizing make-span, power consumption and cost.

A thorough analysis reveals that there is a trade-off between minimizing costs and minimizing energy use. Consequently, this paper proposes ECTS to assign jobs to FN at the fog layer with the least amount of energy consumption and the best possible cost. WOA is inherited by the proposed ECTS method. This method determines the fitness function by adding up the expenses of RAM and CPU (central processing unit) for every fog node, together with the power used for task execution and FN's power when idle. The primary contribution of this study is the development of the ECTS algorithm. This suggested method for vehicular traffic congestion detection applications is simulated using the iFogSim simulator as the secondary contribution.

Table I presents a survey of the latest published study on cloud and fog computing task scheduling strategies, as well as main idea, improvement parameters and algorithms of previous studies.

TABLE I. REVIEW OF THE EXISTING CLOUD FOG ENVIRONMENT'S SCHEDULING PROCEDURES

Year	Author	Improvement Parameter	Algorithm	Main ideas
2019	Jayasena, et al. [3] IEEE	Minimize energy consumption, Reduce cost	Whale Optimization task scheduling algorithm	A fog processing system job-planning strategy that optimizes two objectives: reducing power consumption and cutting expenses.
2019	Xu, et al. [4] IEEE	Energy consumption, Execution time	Laxity based Ant Colony algorithm [LBACA]	To effectively control the adaptability of work latency and energy consumption, implemented the ant colony systems algorithm and flexibility while accounting for the relevance of each task and when it will be finished.
2020	Tan, et al. [5] Elsevier	Energy, Deadline	Energy Efficient scheduling method	An energy-efficient task scheduling method finding the best solution in a polynomial time.
2020	Nikoui, et al. [6] IEEE	Deadline, Cost	Genetic algorithm	A genetic-based (CAGB) planning method that improves efficiency at a lower cost for real-world applications with tight deadlines. Its effectiveness is evaluated regarding system overload, expenses, and delay.
2020	Fellir Z, et al. [7] IEEE	Priority, Execution Time	Priority based task scheduling algorithm	Multi-agent-based planning method, the most important tasks are handled first to ensure that when a packet with the highest importance goes into the waiting line.
2020	Madej, et al. [8] IEEE	Priority, Job Execution	Priority based task scheduling algorithm	Four scheduling schemes named NFCFS (Naive First Come First Serve) technique. The other three schemes are client fair, prioritized fair, and hybridization are presented
2020	Abdel Basset, et al. [9] IEEE	Energy, Makespan	Energy aware task scheduling algorithm	Improved the effectiveness of the best outcome and justified the workload across the accessible simulated engines by using a meta-heuristic method and a shift modification technique.
2020	Yang, et al. [10] IEEE	Total task execution time, resource cost	Meta heuristic scheduling algorithm	Demonstrate how the suggested strategy performs better than conventional strategies regarding resource cost and execution time.
2021	Hoseiny, et al. [11] IEEE	Cost, deadline	Combined (QoS) quality of service and cost effective scheduling method	In contrast to a genetic algorithm, the suggested technique reduces latency, compute costs, and communication costs all at once for IoT inquiries while increasing the proportion of tasks that are finished earlier than the deadline.
2021	Dang, et al. [12] IEEE	Task priority	An algorithm for allocating resources based on task priorities	Resource allocation algorithm based on task priority reduces the average delay in the heterogeneous environment in the fog environment.
2021	Abdel-Basset, et al. [13] IEEE	Energy, makespan	Multi objective scheduling algorithm	The purpose of this algorithm is minimizing energy, make-span and carbon emission rate.
2021	Arshed, et al. [14] IEEE	Execution time, cost	RACE scheduler	Resource aware cost efficient scheduling algorithm at fog layer.
2023	Singh, et al. [15] IEEE	Makespan, cost	Hybrid particle swarm optimization with genetic algorithm (GA)	The purpose of this algorithm is to reduce execution time and cost in cloud fog environment.
2024	Saif, et al. [16] IEEE	Throughput, makespan	Multi-Objectives Grey Wolf Optimizer (MGWO) algorithm	MGWO is multi objective optimization technique for optimal solution of task scheduling.
2024	Zhang, et al.[17] Springer	Execution Cost, power consumption	Improved Whale Optimization Algorithm (EWOA)	EWOA is advanced task scheduling algorithm at cloud computing environment.
2024	Alwabel, et al. [18] IEEE	Deadline , energy consumption	Power-Aware Placement Mechanism (POAPM)	Deadline and energy consumption minimization scheduling at fog computing network
2024	Khan et al. [19] IEEE	Make-span, throughput	Parallel Improved Whale Optimization (PIWO) algorithm	PIWO algorithm is used for allocation of tasks at cloud computing layer. The purpose of this algorithm is minimizing make-span and execution time
2022	Ali et al. [20] IEEE	Execution time, cost	Non-dominated Sorting Genetic (NSG) algorithm II	DNSG is scheduling algorithm at cloud fog environment. Purpose of proposed algorithm is minimizing cost and execution time

III. PROPOSED SYSTEM ARCHITECTURE

This section describes the proposed three tire architecture of a vehicular congestion detection system, and the ECTS algorithm in cloud fog network.

A. Proposed Vehicular Traffic Congestion Detection Architecture

Vehicular Traffic Congestion Detection (VTCD) architecture in a cloud fog network is shown in Fig. 1. As shown in the diagram this is a three-layer model for detecting vehicular traffic congestion. Layer 1 represents the end devices layer, and at this layer, sensors detect vehicular traffic congestion, and forward requests at layer two, i.e., fog computing layer through IoT enabled devices. At the layer two, clusters of FN are available. Each cluster of FN is connected to a Master Fog Server (MFS) called a fog server. MFS is responsible for checking resource availability, scheduling tasks to FN, and assigning tasks (jobs) to the appropriate FN. These fog nodes process jobs and respond to MFS. Using an actuator, MFS responds to end devices and displays traffic congestion detection-related information. MFS also forwards task results to layer three, the Cloud Data Center (CDC), through a proxy server to store the results for future reference.

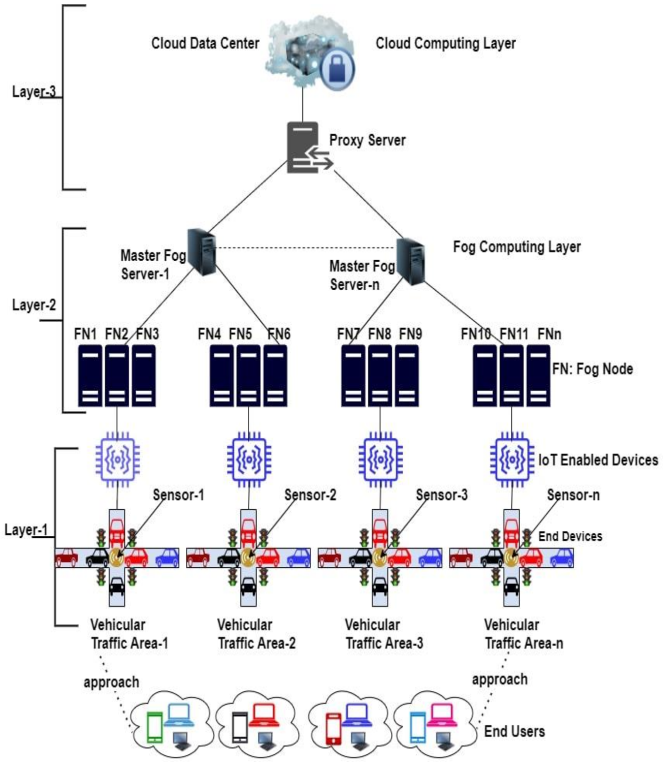


Fig. 1. Proposed three-tier model of vehicular traffic congestion detection.

1) Vehicular traffic area and end devices: As shown in Fig. 1, end users approach end devices using gadgets like tablets, smartphones, desktops, notebook computers, wearable devices, etc. In this paper, end devices are vehicular traffic on roads in different city areas. As we can see in this diagram, vehicular traffic area 1 to vehicular traffic area n are shown where sensor 1 to sensor n detect traffic on the road and forward this information to layer 2 using IoT-enabled devices.

2) The proxy server and fog computing layer: Fog computing is the middle layer in the cloud fog scenario with three-tier architecture called fogging or fog networking. FNs are near-end devices for computing, storage, and communication with end users locally with reduced latency, low bandwidth and lower cost compared to cloud computing environments. An enhanced version of cloud computing, i.e., fogging, reduces stress in the layer of clouds [22-23]. A proxy server is router that communicates with and prevents cyber-attacks, reduces latency between the CDS and layer 2. Data can be retrieved by fog nodes from cloud storage whenever further processing is required.

B. Proposed Task Scheduling Model

The suggested task scheduling technique for assigning jobs to the FN in the cloud fog network is covered in this part. A recommended job scheduling plan is based on the WOA [21] and the multiple-objective model [26]. Fig. 2 depicts the system of the suggested ECTS method of this paper for assigning tasks to the FN. The memory, CPU, and energy consumption functions are all computed using the multiple-objective computation. The cost function and energy consumption are added to get the fitness value. By the fitness rating, the tasks are allocated to the Fog nodes. ECTS first considers the current solution is the best solution. This process is repeated until the optimum solution is identified. In this paper task scheduling aims to minimize energy and total cost while assigning task to the fog node as effectively as feasible.

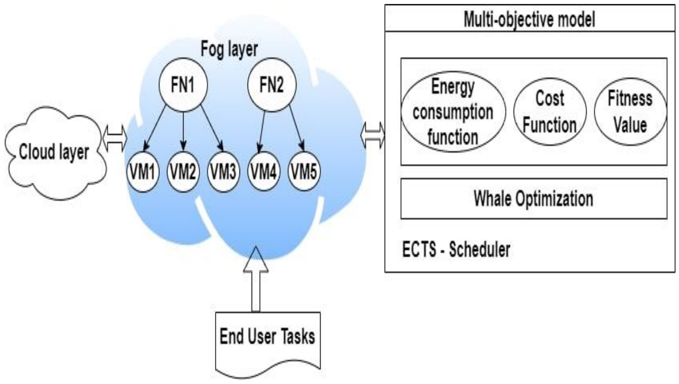


Fig. 2. Model of the proposed ECTS scheduler.

The fog layer which contains of n numbers of fog nodes. Where FL represents fog layer and $\{FN1, FN2, FN3, FN4 \dots\dots\dots.FN_n\}$ represent fog nodes presented at fog layer. This can be represented as,

$$Fog,\ FL = \{FN1,\ FN2,\ \dots FN_n\} \qquad (1)$$

The fog node FN1 can be represented by the equation that follows. Where, $\{FN1,\ FN2,\ \dots FN_n\}$ represents the fog nodes. Master Fog Node (MFS) connected with cluster of fog nodes at layer 2.

$$MFS = \{FN1,\ FN2,\ \dots FN_n\} \qquad (2)$$

Each fog node has the CPU and the memory. Tk represents task and MFS represent master fog server.

$$Tk = \{ T_1,\ T_2 \ldots\ldots T_{500} \} \qquad (3)$$

Where, T_1 is the task first and T_{450} shows 500th task. Assume that one IoT device forward 10 Tk to MFS.

In this paper proposed algorithm: ECTS, where tasks and fog nodes are inputs. The suggested task scheduler aims to best distribute the job across the fog nodes. The WOA is the basis for this scheduler. Initially, the search agent population is initialized. The model of proposed ECTS scheduler computes the fitness value. Equation number 15 updates the search agent's position if the e is greater than or equal to 0.5. The 13th and 14th equation are used to update the search agents' positions if the e is less than 0.5. Until the ideal solution is found, this procedure is carried out repeatedly.

```
Proposed Algorithm : ECTS
Input: Task T, Fog Node FN
Result: Fog nodes are assigned the jobs.
Control Parameters $S^*$, α, N, W, t, p
Begin
    Set up the population initially. $S_i$ (i = 1, 2, …, n)
    Fitness value obtained from the sub function
    Initialize the current best agent $S^*$
    Update α, N, W, t, p
     if ( e < 0.5)
        if ( |N| < 1)
Change the search agent's location with equation 13.
        Else if ( |N| ≥ 1)
   Change the search agent's location with equation 14.
        End if
     End if
     if ( e ≥ 0.5)
         Change the search agent's location with equation 15.
    End if
In the event that (any search agent leaves the search space)
        Update $S^*$
Assign y to y + 1.
    End if
End
Sub function:
Input: Tasks T, Fog Nodes FN
Output: Fitness value
 For (all the Fog nodes)
 Find active energy consumption function by equation 4
 Find idle energy consumption function by equation 5
 Find total energy consumption function by equation 6
 Find the fitness value by equation 12
 End for
End
```

1) Energy Consumption: Energy consumption formula is similar like [15].

$$ER_{usage}(y) = \sum_{i=1}^{n} \mu\, F_i\, V_i^2\ \left(BT_{T_i} \text{ - } ET_{T_i}\right) \qquad (4)$$

$$ER_{idle}(y) = \sum_{j=1}^{m} \sum_{idle_{jk}\ \varepsilon\ IDLE_{jk}} \mu\ RF_{\min i}\ VL_{\min i}^2 \qquad (5)$$

The power consumption during job i, when the resource is operating at peak efficiency and is about to enter sleep mode, and added together to get the total energy consumption.

ER_{usage} is active energy usage and ER_{idle} is idle energy consumed by system at sleep mode. F_i is frequency, V_i is voltage supply fog node where task executes. BT_{T_i}, represent beginning time and ET_{T_i} is end time for task T_i and μ is constant.

$$TER_{con}\ (\mathrm{y}) = ER_{usage} + ER_{idle} \qquad (6)$$

2) Total cost: The fog node total cost is calculated as,

$$\mathrm{C\ (y)} = \sum_{k=1}^{|FN|} C_{cost}\ (k) \qquad (7)$$

$$C_{cost}\ (k) = C_{basic} * C_k * t_{ik} * C_{tr} \qquad (8)$$

$$\mathrm{M\ (y)} = \sum_{k=1}^{|FN|} M_{cost}\ (k) \qquad (9)$$

$$M_{cost}\ (k) = M_{basic} * M_k * t_{ik} * M_{tr} \qquad (10)$$

Where $C_{cost}\ (k)$ is the cost of CPU of fog node FN_k and t_{ik} is the amount of time in which task T_i is executed at node S_k. C_{tr} is the communication cost of the CPU of fog node. Here , C_{basic} and C_{tr} are constant where C_{basic} is 0.16 per hours and C_{tr} is 0.004, much like in [26]. |FN| is the total no of fog nodes. M_{basic} is 0.04 GB per hour and M_{tr} is 0.4, C(y) is the total no of cost of CPU of FN and M(y) is the memory cost of FN . $TC\ (y)$ denote the total cost, it can be calculated as,

$$TC(y) = \mathrm{C(y) + M(y)} \qquad (11)$$

3) Fitness value determining: To determine the best solutions, the fitness value is computed; the solution must have the lowest possible energy consumption and lowest possible cost function. The following formula is used to calculate fitness.

$$FV(y) = TER_{con}\ (\mathrm{y}) + TC(y) \qquad (12)$$

4) Whale optimization algorithm: For distributing the jobs to the fog nodes as efficiently as possible, the whale optimization method [21] is explained. The collection of random solutions is where the whale optimization process starts. It moves forward with the process under the presumption that the present answer is optimal. Repeating this procedure keeps on until the best solution is found.

$$S\ (y+1) = \overrightarrow{S*}(y) - \vec{N} * \vec{D} \qquad (13)$$

$$\vec{S}(y+1) = \overrightarrow{S_{rand}} - \vec{N} * \vec{D} \qquad (14)$$

$$\vec{S}(y+1) = D' * b^{vt} * \cos(2\,\Pi\,t) + S^*(y) \qquad (15)$$

Where, y denotes current iteration, $\vec{S}$ is position vector and $\overrightarrow{S*}$ represents optimal solution. $\overrightarrow{S_{rand}}$ is random location vector, $\vec{N}$ represents the coefficient vector. $\vec{W}$ represents the coefficient vector. In equation 11, v is constant and t shows the value in [-1, 1] interval. | | denotes the absolute value, while * denotes multiplication of elements by elements. D' is calculated as follows,

$$D' = | \overrightarrow{S*}(y) - \vec{S}(y) | \tag{16}$$

$$\vec{D} = | \vec{W} * \overrightarrow{S*}(y) - \vec{S}(y) | \tag{17}$$

Where $\vec{N}$ and $\vec{W}$ can be computed by following formula,

$$\vec{N} = 2 \vec{\propto} * \vec{\beta} - \vec{\propto} \tag{18}$$

$$\vec{W} = 2 * \vec{\beta} \tag{19}$$

The value of $\vec{\propto}$ is move between 2 to 0 and $\vec{\beta}$ denote the arbitrary vector in [0, 1].

IV. Simulation Environment

The simulation environment used to perform calculations is explained in this section. Simulation scenario of Vehicular Traffic Congestion Detection (VTCD) system in fog computing environment has shown in Fig. 3. where S1, S2, S3, S4 are sensors to collect cross-road vehicular traffic information and A1, A2, A3, A4…A8 are actuators to display the results and total four terminals for VTCD system IoT Devices (T_IoT_D1, T_IoT_D2, T_IoT_D3, T_IoT_D4) are used to collect VTCD information from sensors and forward to Fog Node (FN) where four fog nodes (FN1, FN2, FN3, FN4) have used in this simulation at fog computing layer. Controller Fog Server (MFS) collects information from FN and stores it at CDC (Cloud Data Center) via a proxy server. This topology was created and simulated by the iFogSim simulator.

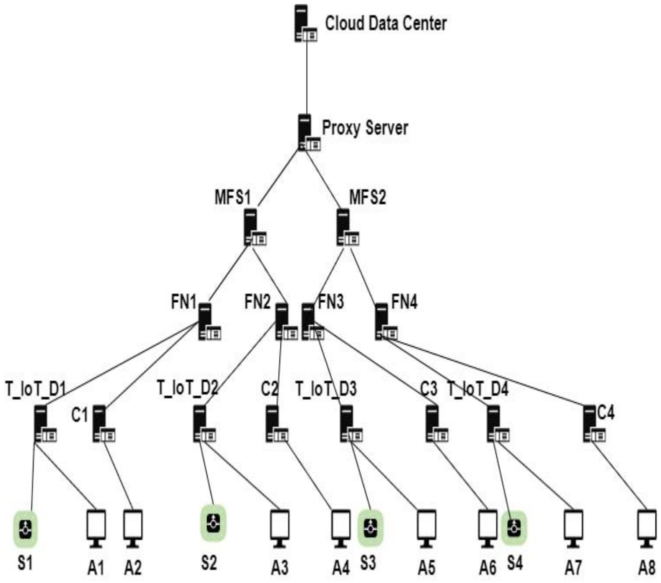


Fig. 3. Scenario of ECTS with VTCD for simulation in cloud fog environment.

Table II shows configuration parameters and values at the Cloud Server, Proxy Server, and Fog Computing layer, which are used to simulate the cloud fog environment. Table III represents the description, various notations used, values assumed in this paper, and system configuration.

TABLE II. Configuration Parameter

Requirement	at Cloud	Proxy server	at Fog
Processing unit (MIPS)	44700	2900	2900
Main memory in MB	9900	3900	3900
Up_bps in MB	100	9900	9900
Dn_bps in MB	9900	9900	9900
Layer	3	2	1
Rate in MIPS	0.01	0	0
power_b in WATT	17*104	107.349	107.349
power_ID in WATT	17*83.24	85.5333	85.5333

TABLE III. Notation, Values and Description

Description	Notation and values
Max no of IoT device	50
p	[-1 , 1]
i	1, 2, 3…
$\text{Max}_{\textbf{itr}}$	100
FN	fog node
EDN	edge device node
System	Intel ® Core(TM) i3 CPU
Tool for simulation	iFogSim
OS(operation system)	Window 7 Ultimate, 64 bit

V. Experimental Results and Performance Analysis for Proposed ECTS Algorithm

Performance evaluation of the proposed algorithm with VTCD application in a cloud fog network is shown in this section. The measurements for energy consumption performance and the simulation result of overall cost are shown in Tables IV and V, respectively. The corresponding bar chart of the parameters shows that energy consumption is minimized when number of IoT devices have increased as shown in Fig. 4, and cost is also minimized when no of IoT devices have increased as shown in Fig. 5. Assume that 10 IoT devices are equal to 100 tasks.

TABLE IV. Simulation Results for the Energy Consumption

No. of IoT Devices	Consumption of energy (in WATTS)
10	188040.91
20	185103.87
30	185103.39
40	182487.18
50	176916.18

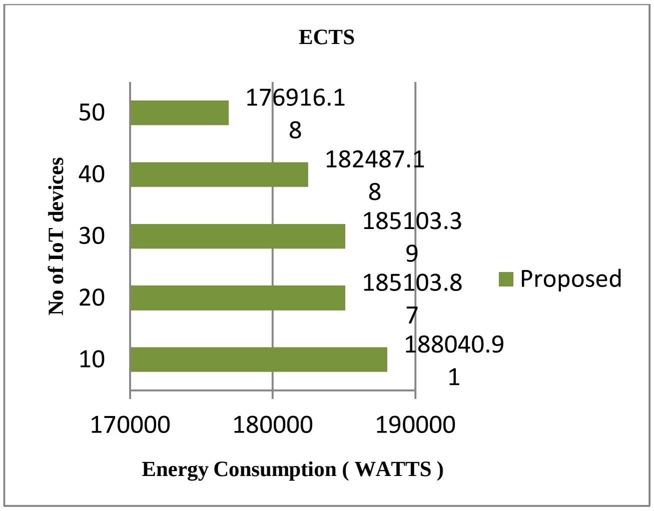


Fig. 4. Energy consumption.

TABLE V. SIMULATION RESULTS FOR THE COST

No. of IoT Devices	Cost ($)
10	395905.15
20	392444.41
30	614328.21
40	761434.06
50	810188.88

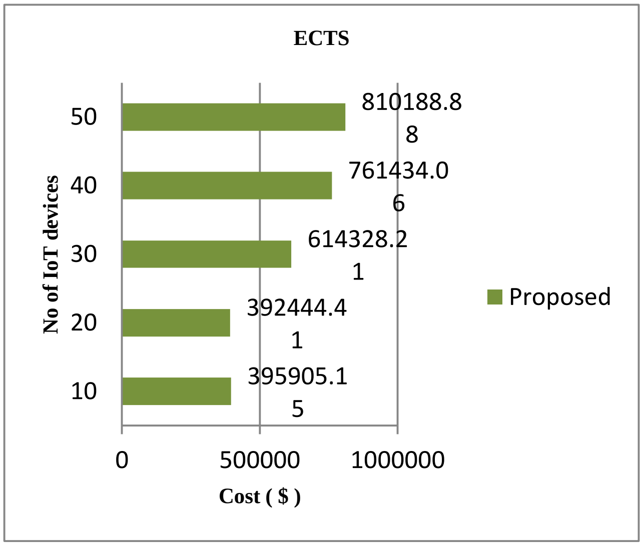


Fig. 5. Cost.

Table VI shows the comparative study of the proposed ECTS method with the existing methods, such as (Round Robin) RR [3] and (Cost aware genetic algorithm) GA [6]. The proposed ECTS method has the minimum energy consumption 176916.18. The energy consumption of RR is 201259.8464 and the energy consumption of GA is 188574.9563. The total cost of RR and GA are 959749.7472 and 918592.1521 respectively while the total cost of proposed ECTS method is 810188.88 which is smaller than the other existing methods.

TABLE VI. COMPARATIVE STUDY OF THE ECTS METHOD WITH THE RR AND GA

	ECTS	RR	GA
Energy consumption	176916.18	201259.8464	188574.9563
Cost	810188.88	959749.7472	918592.1521

This is the analysis of the proposed algorithm in this paper. A range of IoT devices as input of 10-50 have been used for the simulation at the iFogSim simulator. Assume that 10 IoT devices are equal to 100 tasks and 50 IoT devices are equal to 500 tasks. As shown in Fig. 6 energy consumptions are minimized in the proposed algorithm when no of IoT devices are increased as compared to RR and GA. As shown in Fig. 7. The proposed algorithm minimizes the overall cost when the number of IoT devices increases compared to RR and GA.

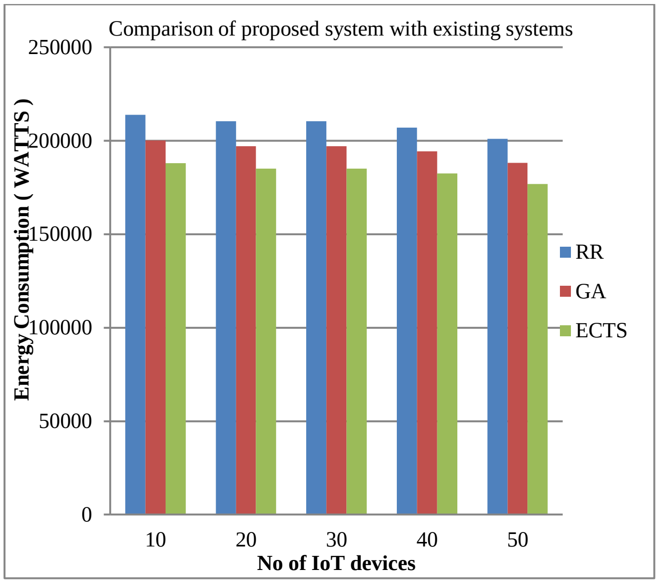


Fig. 6. Energy consumption comparison.

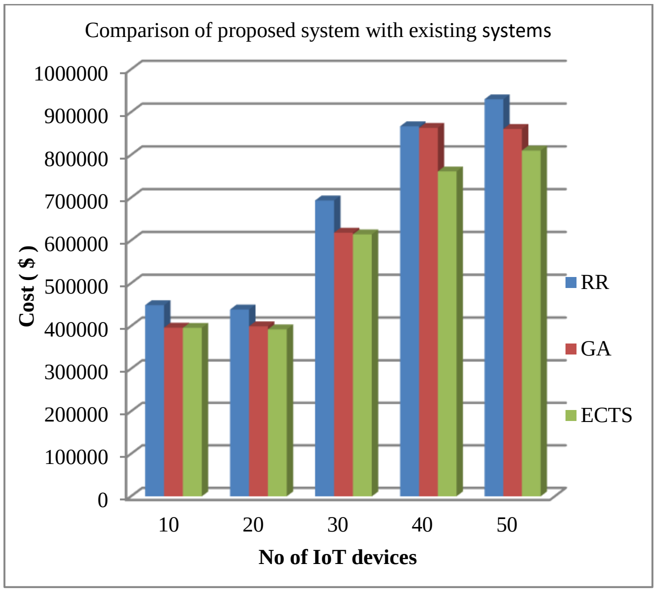


Fig. 7. Cost comparison.

Column chart for comparison of energy consumption, overall cost with GA, RR, and proposed algorithm show that the proposed result is better compared to others, especially in energy and cost parameters.

VI. Conclusion

The design of the task scheduling technique is the primary purpose of this study. The secondary objective is designing a Vehicular Traffic Congestion Detection (VTCD) system. The proposed Energy-aware Cost effective Task Scheduling (ECTS) scheduling algorithm performance has been analyzed using various inputs. Two other approaches, particularly the Genetic Algorithm (GA) and Round-Robin (RR) in a cloud-fog network, were compared with the proposed algorithm using the iFogSim simulator. Especially for energy usage and cost parameters, our proposed algorithm ECTS performed better than the others at five different sets of inputs. The simulation result shows that energy consumption is minimized by 6.59%, and the overall cost is minimized by 13.38% compared to GA. In comparison, energy consumption is minimized by 13.75%, and the overall cost is minimized by 18.46% compared to RR. Here, multi-objective means task’s cost, energy consumption and deadline for scheduling the user's request at the fog computing layer. Furthermore, the suggested algorithm may adapt to the end user's requirement for higher processing performance for other applications.

In the future, improvements may be made in ECTS algorithm to address other issues like reducing make-span, response time, security issues, etc. improvement for other real time applications.

Acknowledgment

I sincerely thank Dr. A. K. Chaturvedi, Head, Computer Application Dept., at Govt. Engineering College, Ajmer, Rajasthan, for truthful academic backing throughout the work.

References

[1] A. I. Abueid, “Big Data and Cloud Computing Opportunities and Application Areas”, Eng. Technol. Appl. Sci. Res., vol. 14, no. 3, pp. 14509–14516, Jun. 2024

[2] M. Ramzan, M. S. Farooq, A. Zamir, W. Akhtar, M. Ilyas, and H. U. Khan, “An Analysis of Issues for Adoption of Cloud Computing in Telecom Industries”, Eng. Technol. Appl. Sci. Res., vol. 8, no. 4, pp. 3157–3161, Aug. 2018.

[3] N. Jayasena, K. P., & Thisarasinghe, B. S. (2019). Optimized task scheduling on fog computing environment using meta heuristic algorithms. 2019 IEEE International Conference on Smart Cloud. doi:10.1109/smartcloud.2019.00019.

[4] Xu, J., Hao, Z., Zhang, R., & Sun, X. (2019). A Method Based on the Combination of Laxity and Ant Colony System for Cloud-Fog Task Scheduling. IEEE Access, 7, 116218–116226. doi:10.1109/access.2019.2936116.

[5] Tan, H., Chen, W., Qin, L., Zhu, J., & Huang, H., Energy-aware and Deadline-constrained Task Scheduling in Fog Computing Systems. 2020, 15th International Conference on Computer Science & Education (ICCSE). doi:10.1109/iccse49874.2020.92017.

[6] Nikoui, T. S., Balador, A., Rahmani, A. M., & Bakhshi, Z. (2020). Cost-Aware Task Scheduling in Fog-Cloud Environment. 2020 CSI/CPSSI International Symposium on Real-Time and Embedded Systems and Technologies (RTEST). doi:10.1109/rtest49666.2020.9140118.

[7] Fellir, F., El Attar, A., Nafil, K., & Chung, L. (2020). A multi-Agent based model for task scheduling in cloud-fog computing platform. 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). doi:10.1109/iciot48696.2020.9089652.

[8] Madej, A., Wang, N., Athanasopoulos, N., Ranjan, R., & Varghese, B. (2020). Priority-based Fair Scheduling in Edge Computing. 2020 IEEE 4th International Conference on Fog and Edge Computing ICFEC). doi:10.1109/icfec50348.2020.00012.

[9] Abdel-Basset, M., El-shahat, D., Elhoseny, M., & Song, H. (2020). Energy-Aware Metaheuristic algorithm for Industrial Internet of Things task scheduling problems in fog computing applications. IEEE Internet of Things Journal, 1–1. doi:10.1109/jiot.2020.3012617.

[10] Yang, M., Ma, H., Wei, S., Zeng, Y., Chen, Y., & Hu, Y. (2020). A Multi-Objective Task Scheduling Method for Fog Computing in Cyber-Physical-Social Services. IEEE Access, 8, 65085–65095. doi:10.1109/access.2020.2983742.

[11] Hoseiny, F., Azizi, S., Shojafar, M., & Tafazolli, R. (2021). Joint QoS-aware and Cost-efficient Task Scheduling for Fog-cloud Resources in a Volunteer Computing System. ACM Transactions on Internet Technology, 21(4), 1–21. doi:10.1145/3418501.

[12] Tran-Dang, H., & Kim, D.-S. (2021). Task Priority-based Resource Allocation Algorithm for Task Offloading in Fog-enabled IoT Systems. 2021 International Conference on Information Networking. (ICOIN). doi:10.1109/icoin50884.2021.9333992.

[13] M. Abdel-Basset, N. Moustafa, R. Mohamed, O. M. Elkomy and M. Abouhawwash, "Multi-Objective Task Scheduling Approach for Fog Computing," in IEEE Access, vol. 9, pp. 126988-127009, 2021, doi: 10.1109/ACCESS.2021.3111130.

[14] J. U. Arshed and M. Ahmed, "RACE: Resource Aware Cost-Efficient Scheduler for Cloud Fog Environment," in IEEE Access, vol. 9, pp. 65688-65701, 2021, doi: 10.1109/ACCESS.2021.3068817.

[15] Singh, G., Chaturvedi, A.K. Hybrid modified particle swarm optimization with genetic algorithm (GA) based workflow scheduling in cloud-fog environment for multi-objective optimization. Cluster Comput 27, 1947–1964 (2024). https://doi.org/10.1007/s10586-023-04071-1.

[16] F. A. Saif, R. Latip, Z. M. Hanapi and K. Shafinah, "Multi-Objective Grey Wolf Optimizer Algorithm for Task Scheduling in Cloud-Fog Computing," in IEEE Access, vol. 11, pp. 20635-20646, 2023, doi: 10.1109/ACCESS.2023.3241240.

[17] Zhang, Y., Wang, J. Enhanced Whale Optimization Algorithm for task scheduling in cloud computing environments. J. Eng. Appl. Sci. 71, 121 (2024). https://doi.org/10.1186/s44147-024-00445-3.

[18] A. Alwabel and C. K. Swain, "Deadline and Energy-Aware Application Module Placement in Fog-Cloud Systems," in IEEE Access, vol. 12, pp. 5284-5294, 2024, doi: 10.1109/ACCESS.2024.3350171.

[19] Z. A. Khan, I. A. Aziz, N. A. B. Osman and S. Nabi, "Parallel Enhanced Whale Optimization Algorithm for Independent Tasks Scheduling on Cloud Computing," in IEEE Access, vol. 12, pp. 23529-23548, 2024, doi: 10.1109/ACCESS.2024.3364700.

[20] I. M. Ali, K. M. Sallam, N. Moustafa, R. Chakraborty, M. Ryan and K. -K. R. Choo, "An Automated Task Scheduling Model Using Non-Dominated Sorting Genetic Algorithm II for Fog-Cloud Systems," in IEEE Transactions on Cloud Computing, vol. 10, no. 4, pp. 2294-2308, 1 Oct.-Dec. 2022, doi: 10.1109/TCC.2020.3032386.

[21] Mirjalili S. and Lewis A., “The whale optimization algorithm,” Advances in engineering software, vol. 95, pp. 51–67, Elsevier, 2016.

[22] Celso A. R. L. Brennand, Daniel Ludovico Guidoni (2021). Fog Computing-based Traffic Management Support forIntelligent Transportation Systems. 17165-217-13756-1-10-20210911.

[23] Ning, Z., Huang, J., & Wang, X. (2019). Vehicular Fog Computing: Enabling Real-Time Traffic Management for Smart Cities. IEEE Wireless Communications, 26(1), 87–93. doi:10.1109/mwc.2019.1700441.

[24] Mishra P. K., A. K. Chaturvedi, "State-Of- The-Art and Research Challenges in Task Scheduling and Resource Allocation Methods for Cloud-Fog Environment," 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, IEEE, doi: 10.1109/ICCT56969.2023.

[25] Mishra P. K., Chaturvedi A. K., "Research Challenges in Job Scheduling and Resource Distribution Methodology for Cloud Fog Network: An

Organized Analysis. " International Conference on Computational Intelligence, Communication Technology and Networking, 2023, IEEE, doi: 10.1109/CICTN57981.2023.

[26] Sreenu, K., Sreelatha, M. W-Scheduler: whale optimization for task scheduling in cloud computing. Cluster Comput 22 (Suppl 1), 1087–1098 (2019). https://doi.org/10.1007/s10586-017-1055-5

Deep Learning with IoT-Based Solar Energy System for Future Smart Agriculture System

Vidya M S, Ravi Kumar B. N, Anil G. N, Ambika G. N
Department of Computer Science and Engineering, BMS Institute of Technology, Bangalore, India

***Abstract*—Agriculture has a considerable contribution to the economy. Agriculture automation is a serious issue that is becoming more prevalent around the world. Farmers' traditional practices were insufficient to achieve these objectives. Artificial Intelligence (A1) and the Internet of Things (IoTs) are being used in agriculture to improve crop yield and quality. Distributed solar energy resources can now be remotely operated, monitored, and controlled through the IoT and deep learning technology. The development of an IoT-based solar energy system for intelligent irrigation is critical for water- and energy-stressed areas around the world. The qualitative design focuses on secondary data collection techniques. The deep learning model Radial Basis Function Networks (RBFN) is used in conjunction with the Elephant Search Algorithm (ESA) in this IoT-based solar energy system for future smart agriculture. Sensor systems help farmers understand their crops better, reduce their environmental impact and conserve resources. These advanced systems enable effective soil and weather monitoring, as well as water management. To provide the required operating power, the proposed system, RBFN-ESA, employs an IoT-based solar cell forecasting process. The proposed model RBFN-ESA will collect these data to predict the required parameter values for solar energy systems in future smart agriculture systems. The results of the RBFN-ESA model are effective and efficient. According to the findings, RBFN-ESA outperforms CNN, ANN, SVM, RF, and LSTM in terms of energy consumption (56.764J for 100 data points from the dataset), accuracy achieved (97.467% for 600 nodes), and soil moisture level (94.41% for 600 data).**

Keywords*—Precision agriculture; smart monitoring; Internet of Things; Radial Basis Function Networks; Elephant Search Algorithm (ESA)

I. Introduction

Food manufacturing in the 20th century is a pressing issue as long as population growth continues to increase. Between 9.4 and 10.1 billion people will rely on biodiversity for their livelihood by 2050, which would raise the demand for locations set aside for agricultural production, especially for farming and the rearing of farm animals [1]. Human-induced changes to the environment can result in conditions that make it difficult for new crops to flourish. Similarly, rising urbanization raises food prices while decreasing food.

Production and employment in food-producing regions. In an effort to address the challenges of fulfilling the demands of food production and the working population decline, smart agriculture aims to lower farm management expenses [2–3]. It employs techniques and technology at various agricultural production scales and levels. Precision farming, for example, can employ a range of sensing devices to collect data (heat, moisture, light, stress, presence, etc.), connectivity networks to receive and send that data, information management systems to keep and process that data, and analysis tools to do so [4]. "IoT" is a term used frequently to describe this network of connected devices [5]. The right actions can be taken thanks to the knowledge that intelligent farming generates.

Recent developments in wireless technology have completely changed how farmers can interact with their crops and track their growth [6]. Advanced management concepts can be used to monitor crops using new technologies and respond to their needs appropriately. Precision agriculture (PA) is one method that combines technology and conventional farming methods [7, 8]. Using PA in farming can increase control and accuracy when raising animals and crops. Farmers are becoming increasingly productive and cost-effective by utilizing new technology to enable agriculture because they can use more precise solutions rather than simply attempting to manage the many elements of their farming systems.

Instead of using modern technology, traditional farming practices are used to manage fields. More experience is required to maintain proper efficiency. With traditional farming methods, the best course of action for a successful harvest must be determined by considering both the current weather and historical data when making decisions about planting, harvesting, and irrigation. Contrarily, PA helps farmers use less labour while giving their crops more attention as needed by using tools like sensors, actuators, the Global Positioning System (GPS), robots, and data analysis software. Monitoring livestock and vegetation with Internet of Things (IoT) devices is one effective method for achieving PA [9]. IoT devices are minuscule, energy-efficient embedded electronics with network data transmission capabilities. An IoTs network is frequently used to describe a group of connected devices that work together to accomplish a common objective. Sensors, for example, can be installed in an IoT-based agricultural system to collect environmental information about soil moisture. An automated irrigation system can use the measured data to water plants appropriately, avoiding over- and under-watering. Farmers might be able to instantly and remotely monitor field conditions thanks to an IoT system. It is just as crucial to keep an eye on the vegetation in a field as it is to keep an eye on livestock to ensure that they are fed and cared for properly. The use of IoT devices can lower labour costs significantly and enhance animal welfare. IoT devices can be used to find the livestock's location and assess its health.

II. LITERATURE SURVEY

Systems remain a type of feedforward neural system that activates using radial basis functions and universal approximators. Classification, regression, pattern recognition, and time series forecasting problems are frequently solved using RBFN [10–11]. In addition to their strong ability to approximate any continuous network, RBFNs also possess strong characteristics like their compact structure, noise tolerance, and ability to approximate any global approximation. The new elephant algorithm is among the most recent meta-heuristic methodologies to be suggested. The search areas of elephant males are widened as they travel farther and farther. The female elephants focus on seeking out the best response locally. A lifespan mechanism that regulates birth and death gives all agents a gradually increasing chance of dying as they age. The heuristic knowledge of these elephants' forebears will be passed down to them, and this mechanism is designed to keep whole agents from entering the local optimum. The solar energy-driven polygene ration system configurations and classified them based on design, benefits, technical potentials, challenges, and market prospects. A solar-driven multigeneration system enhances the system's efficiency and reduces the capital and operation costs as well as carbon dioxide emissions to improve the environment [12].

Soil temperature modelling to assess the viability of using soil air exchangers for agricultural structures. In this context, the ability of soil to cool or heat agricultural structures such as greenhouses was determined by modifying temperature behaviour at various depths [13]. A solar thermal system produces inexpensive, environmentally friendly heat using the sun's energy. Temperature pushes, electronic warmers, and rotation forces are all controlled in accordance with the need for hot water in a building. First, take into account clear, cloudy, rainy, and dark weather[14]. The network's overall node count could be decreased while the sampling frequency was raised. Although reducing the number of sensor nodes has been shown to result in a similar network lifetime, it is unknown how much data is lost from specific locations within a field. Even though there are more samples, most agricultural systems don't need quick responses because the environment doesn't change quickly over short periods of time [15]. The temperature readings taken by the drone while it was flying over the crop were incorrect, according to experiments. The drone was able to get more precise readings when it was nearer the area of interest. The devices had to be in constant time sync for the data collected among the drone and nodes to be accurate [16]. In order to implement multiple networks, the system was built with nodes that could switch between two operating frequencies. Nodes were organized into clusters, and the cluster leader forwarded data from each cluster to the target node. The outcomes showed that the design used very little energy and could work for a whole season on just one charge of the battery [17].

Wen-tai Li et al. presented by [18]. Building managers can achieve their energy management objectives with the aid of the Solar Water Heating (SWH) control mechanisms. The methods are based on the price of electricity, the weather, and the demand for hot water. An important source of solar energy for buildings, solar thermal systems are the subject of this study. A solar thermal system produces inexpensive, environmentally friendly heat using the sun's energy. Temperature pushes, electronic warmers, then rotation forces remain all controlled in accordance with a building's need for hot water. First, take into account clear, cloudy, rainy, and dark weather. To run the simulations, three different days were picked: a cloudy day, a sunny day, and a semi-synthetic day with no solar. The ideal control mechanism for heat pumps, electric heaters, and circulator pumps has been researched to enhance the SWH system's performance.

Mohammadi et al. [19] reviewed various solar and hybrid solar energy driven polygene ration system configurations and classified them based on design, benefits, technical potentials, challenges, and market prospective. Solar-driven multigeneration system enhances the system efficiency and reduces the capital and operation costs as well as carbon di-oxide emissions to build environment.

Faridi et al. [20] Soil temperature modelling was used to assess the viability of using soil-air exchangers for agricultural structures. The ability of soil to heat or cool agro constructions like greenhouses was detected by means of modifying the behavior of temperature at various depths in this context.

The objective of the paper is

1) First this IoT-based solar energy system for a future smart agriculture system, a deep learning [36, 37] model called Radial Basis Function Networks (RBFN) with the Elephant Search Algorithm (ESA) has been used.

2) IoT and automation are linked with agriculture and farming practises in order to recover the efficacy and efficiency of the entire process.

3) Sensory systems promoted resource conservation, decreased detrimental environmental effects, and improved farmers' understanding of crops. These innovative systems allow for effective soil and weather monitoring as well as efficient water management.

4) To supply the necessary operating power, the proposed system, RBFN-ESA, makes use of a forecasting method from an IoT-based solar cell.

5) The IoT controller reads the data from the humidity, field-based temperature and soil moisture sensors and then outputs the required actuation command signals to drive irrigation pumps.

6) The proposed model RBFN-ESA will collect these data to predict the values of the important solar energy system parameters for a future smart agriculture system.

III. PROPOSED SYSTEM

Every aspect of conventional farming practices can be drastically altered by integrating the most recent sensing and IoT technologies. Now that the IoTs and wireless sensors are available, smart agriculture [35] can reach new heights. By implementing smart farming techniques like drought response, yield enhancement, land applicability, irrigation [34], and pest control, the Internet could really help improve options for so many traditional farming problems. The RBFN-ESA method's block diagram is shown in Fig. 1.

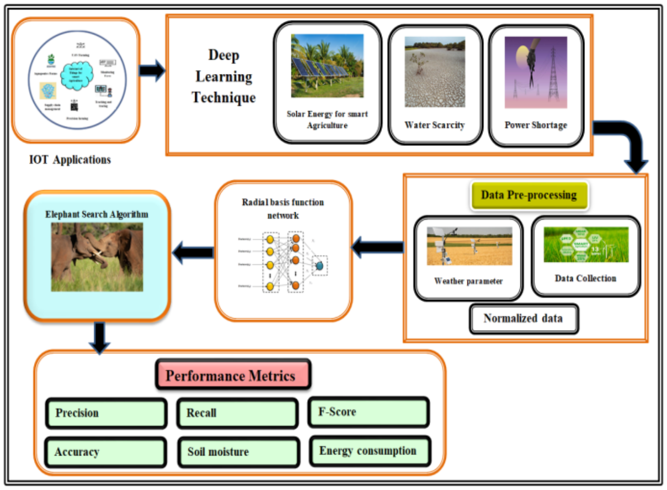


Fig. 1. Block diagram for the approach proposed by RBFN-ESA.

A. Data Pre-processing

The method for considering the weather parameter, data collection, and normalised data during the data pre-processing is described as follows.

1) Weather metric: Accuweather is used to get the daily weather parameters and their measurement units. Based on the temperature (in degrees Celsius), date (dd/mm/yy), season, and daily rainfall, this parameter is used to calculate the amount of rain that will fall on a specific day. The probability of rainfall is taken into account when choosing the aforementioned parameters. The chosen parameter is only equipped to predict the weather.

2) Data collection: For this experiment, we made use of actual data, particularly weather information from Kolkata, West Bengal. The data has been standardized. This data was gathered from the online weather resource Accuweather.com. Data from the first year is used for training, and data from the next 50 days is used for testing.

3) Normalized data: The next stage of data processing, known as "normalization," has arrived after the choice of weather parameters and completion of data collection. Random data in GA must be in normalized form for training and testing. It might be challenging to combine when the GA is trained using real data. Every bit of data is fixed and changed to a value of 0 or 1.

B. Web-Based Water Motor Control Service

A web server built atop the HTTP protocol has been developed to stop and start the water motor. The programming language in R-Pi has accessed this web service to start or stop this same water motor. The Pic Microcontroller's programming language sends signals to the Arduino-Uno, which controls the spread circuit to start and stop the fluid motor.

C. Digital Water Pump

In this subsystem, an aquatic force is attached to a convey button that is managed by a base station with Bluetooth capabilities. For real time monitoring, the web service stimulates base station control from the flexible web-based interface. The water pump can be controlled remotely, both automatically and manually, using this web-based interface.

D. Internet of Things

The IoT is a station of smart, interrelated substances that can transmit information and generate useful data about the market environment. As a result, almost any object that can connect to the Internet can be referred to as a "thing" in the context of the IoTs, including furniture, electronics, appliances, agricultural or industrial machinery, and level public [21-23].

The IoT concept is not new-fangled, but acceptance has recently risen. Some of the technologies that have developed to support it include big data, cloud computing, artificial intelligence, and hardware advancements that have reduced the scope and control of feasting and improved connectivity via the Internet and among plans via wireless connections [24]. Together, these technical parts make a net of nodes that can send and receive information and data and react to interference from the network.

Even though the structure of an IoT network is similar to that of other computer system architectures, [25] says that the identification, sensing, and control of remote devices, as well as the limited computing power of the equipment, are some of the unique aspects of this framework that must be taken into account.

E. Classification using RBFN

The most basic RBFN configuration is a three-layer feed-forward neural network. The network's inputs are represented by the first layer, and its final output is represented by the second layer, which is a hidden layer made up of numerous RBF non-linear activation units. Gaussian functions are frequently used in RBFNs to implement activation functions [26]. An illustration of the RBFN framework can be found in Fig. 2. Let's say we have a dataset D that contains N structures of (xp, yp), where xp is the data set's input.

Eq. (1) can be used to calculate the production of the ith initiation function φi in the net's unseen coating based on the separation among the input pattern x and the centre i.

$$\phi_i(\| x-d_i \|) = exp\left(-\frac{\| x-d_i \|^2}{2\sigma^2{}_j} \right) \qquad (1)$$

The ||.|| hidden neuron j's centre and width are represented by d_i and σ_j, the Euclidean norm.

Then, Eq. (2) can be used to determine the output of node k of the network's output layer:

$$yk = \sum_{j=1}^{n} \omega jk \phi_j(x) \qquad (2)$$

The majority of conventional training methods for RBFNs described in the literature consist of two step [27]. For example, in the first stage, an unsupervised clustering algorithm is used to compute the widths and centers. In order to reduce an error criterion, such as the common mean squared error (MSE) over the whole dataset, the hidden layer and output layer's connection weights must be determined in the second step.

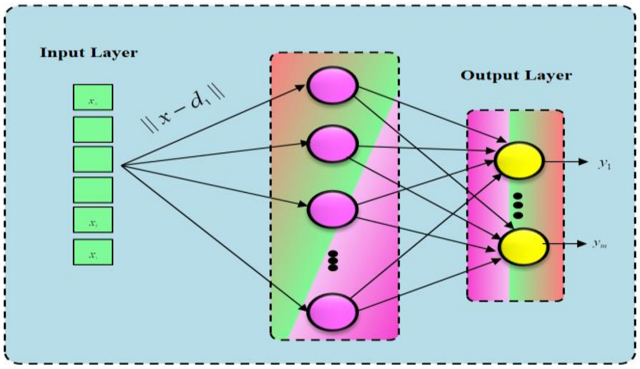


Fig. 2. The RBFN's structure.

F. Elephant Search Algorithm (ESA)

The most recent generation of meta-heuristic search optimization algorithms includes ESA. A dual search mechanism, or the ability to divide the search agents into two groups, is the foundation of this algorithm's approach, which mimics the characteristics and behaviours of an elephant [28]. Elephants live in herds, and each herd is made up of several smaller clans or groupings, each led by the eldest elephant in the herd. The ESA mimics elephant herds' major characteristics and qualities. Elephants have different social systems, with males preferring solitary living and females preferring family units. Female elephants are more concerned with improving their surroundings, whilst male elephants are in charge of discovering new locations to explore.

In this case, ESA is a good search optimization algorithm that has the following three main traits:

1) The search process enhances the present response iteratively in order to identify the ideal one. Chief female elephants also conduct extensive local searches in regions where they believe there is a better chance of finding the greatest solution.

2) Male elephants are in charge of foraging outside the neighborhood's ideal range.

3) Elephants possess a variety of traits, making it crucial to draw inspiration from their biological behaviour. Here is a description of the ESA.

Algorithm 1: Elephant Search Algorithm (ESA)

```
Input: SearchSpace, HerdSize, MaxIterations
Output: BestSolution
Initialize the herd
Herd ← InitializeHerd(HerdSize, SearchSpace)
BestSolution ← None
Main search loop
for iteration = 1 to MaxIterations do
Evaluate the herd's position
  for each elephant in Herd do
elephant.fitness ← EvaluateFitness(elephant.position)
if BestSolution is None or elephant.fitness is better than BestSolution.fitness then
   BestSolution ← elephant
end if
  end for
Communication among elephants (sharing the best known solution)
BestElephant ← FindBestElephant(Herd)
for each elephant in Herd do
  if elephant ≠ BestElephant then
elephant.position ← MoveTowards(BestElephant.position, elephant.position)
  end if
end for
Random exploration to avoid local optima
for each elephant in Herd do
  if rand() < ExplorationProbability then
elephant.position ← RandomMove(SearchSpace)
  end if
end for
Memory retention (remembering good positions)
for each elephant in Herd do
   if rand() < MemoryRetentionProbability then
elephant.position ← elephant.bestKnownPosition
  else
elephant.bestKnownPosition ← elephant.position
  end if
end for
end for
Return the best found solution
return BestSolution
End Algorithm
```

Each elephant must be a member of a clan since they all live together in a herd under the leadership of the oldest elephant [29]. The equation below can be used to represent the animal j in the cli clan.

$$Y_{new,cli,j} = Y_{cli,j} + c(Y_{Best,cli} - Y_{cli,j}).r \tag{3}$$

where $Y_{Best,cli}$ denotes the clan cli and $r \in [0,1]$, and $Y_{new,cli,j}$ and $Y_{cli,j}$ are the elephant j's newly updated and old places in clan cli, separately. $c \in [0,1]$ determines how clan cli influences $Y_{cli,j}$. Eq. (3) cannot be applied when $Y_{cli,j} = Y_{Best,cli}$, but the fittest elephant can be determined using the formula shown below.

$$Y_{new,cli,j} = \alpha.Y_{center,cli} \tag{4}$$

where $\alpha \in [0,1]$ stands for the $Y_{center,cli}$ s impact on the $Y_{new,cli,j}$. The d th dimension of the new individual $Y_{new,cli,j}$ is then updated using the formula below.

$$Y_{center,cli,d} = \frac{1}{n_{cli}}.\sum_{j=1}^{n_{cli}} X_{cli,j,d} \qquad (5)$$

There are that many elephants in the cli clan, $1 \le d \le Dindicate\ the\ dth\ \dim ension$ e, D is its overall dimension, and $Y_{cli,j,d}$ is the d th of the individual $Y_{cli,j}$ elephant.

As mentioned earlier, adult male elephants continue living alone in a remote area after leaving their families [30]. By using a separating operator to solve challenging optimization problems, this scenario can be simulated. Let's assume that the animal individual people with worst fitness particular instance will use the trying to separate operator in compliance with the appropriate equation to enhance the search functionality of ESA [31].

$$Y_{worst,cli,d} = Y_{Min} + (Y_{Max} - Y_{Min} + 1).Rand \qquad (6)$$

where Y_{Max} and Y_{Min} are the highest and lowest limits of an elephant's position, $Y_{worst,cli}$ is the worst elephant member of clan cli, and $Rand \in [0,1]$ is a random distribution [32]. The description of the clan updating and separating operator has been included in the ESA development.

IV. Result and Discussion

The main goal of this experiment is to use sensors to gather the physical characteristics of a farming area. From there, an algorithm will be developed using the sensor data and weather forecast information to predict soil moisture for the upcoming days [33]. This study compares the proposed MPNN-MCOA algorithm with the convolution neural network, support vector machine, artificial neural network, long short-term memory and random forest as five machine learning algorithms.

A. *Assessment Criteria*

- True Positives (TP) are instances where both the actual yield and our expectations came true.
- True Negatives (TN): Occurrences in which the true yield turned out to be incorrect, as predicted.
- False Positives (FP): We expected real results, but the yield was incorrect.
- False Negatives (FN): When an outcome that we anticipated to be untrue proved to be accurate.

Precision: It is also known as the ratio of results that were correctly predicted as positive to results that were actually positive.

$$P\,recision = \frac{TP}{TP + FP} \qquad (7)$$

Recall: It is determined by separating the total amount of successful results by the total amount of conjugate samples.

$$Recall = \frac{TP}{TP + FN} \qquad (8)$$

F1-score: It also goes by the name "harmonic mean" and aims to balance precision and recall. The computation works well on an unbalanced dataset and allows for both false negatives and false positives.

$$F1 - score = \frac{2TP}{2TP + FP + FN} \qquad (9)$$

Accuracy: The percentage of precise predictions to all input models is referred to by this expression.

$$Ac\chi\upsilon\rho\alpha\chi\psi = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

B. *Precision Analysis*

Fig. 3 and Table I provide a comparison of the RBFN-ESA method's precision with that of other methods now in use. The precision with which the deep learning with IOT method has enhanced performance is illustrated by the graph. For example, the precision of the RBFN-ESA method for data 100 is 86.743%, whereas the CNN, ANN, SVM, RF, and LSTM methods have precision values of 83.487%, 78.256%, 75.187%, 69.664%, and 72.387%, respectively. However, the RBFN-ESA method has shown optimal performance over a range of data set sizes. Under 600 data points, the RBFN-ESA methods precision value is 92.864%; in contrast, the CNN, ANN, SVM, RF, and LSTM methods have precision values of 84.754%, 81.242%, 77.854%, 71.643%, and 74.532%, respectively.

TABLE I. PRECISION ANALYSIS OF THE RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	83.487	78.256	75.187	69.664	72.387	86.743
200	82.954	78.654	75.533	70.532	72.854	87.953
300	82.843	79.054	76.863	70.843	73.454	88.435
400	83.543	79.435	77.095	69.853	72.964	91.653
500	84.864	80.774	76.346	71.254	73.964	93.643
600	84.754	81.242	77.854	71.643	74.532	92.864

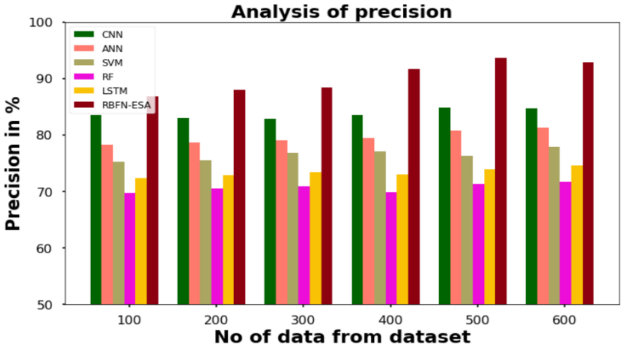


Fig. 3. Precision analysis for RBFN-ESA method.

C. Recall Analysis

Fig. 4 and Table II compare the recall analysis of the RBFN-ESA method with existing methods. The graphic shows how recall performance has increased with the deep learning with IOT method. For example, the recall value for data 100 for the RBFN-ESA method is 90.542%, whereas the corresponding values for the CNN, ANN, SVM, RF, and LSTM methods are 77.76%, 86.543%, 80.187%, 74.875%, and 85.765%. The RBFN-ESA method has performed at its best with various data sizes, though. Similar to this, for 600 data, the recall value of the RBFN-ESA is 94.765%, while for CNN, ANN, SVM, RF, and LSTM methods, it is 79.942%, 88.864%, 84.854%, 78.543%, and 87.912%, respectively.

TABLE II. RECALL ANALYSIS FOR RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	77.765	86.543	80.187	74.875	85.765	90.542
200	76.643	86.954	81.286	75.278	85.265	89.542
300	77.923	87.254	84.543	73.478	87.397	91.467
400	78.743	87.854	83.567	76.187	86.093	93.965
500	79.654	88.145	82.864	77.098	87.743	92.376
600	79.942	88.864	84.854	78.543	87.912	94.765

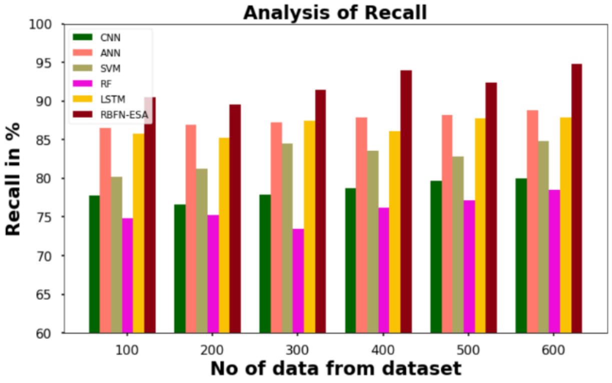


Fig. 4. Recall analysis for RBFN-ESA method.

D. F-Score Analysis

Fig. 5 and Table III provide comparative f-score analyses of the RBFN-ESA method with other existing methods. The graph shows that the f-score performance has improved with the deep learning with IOT method. For example, the f-score value of the RBFN-ESA method for data 100 is 94.095%, whereas the corresponding values for the CNN, ANN, SVM, RF, and LSTM methods are 88.643%, 85.865%, 78.543%, 91.754%, and 81.765%. However, the RBFN-ESA method has shown optimal performance over a range of data sizes. In comparison to the CNN, ANN, SVM, RF, and LSTM methods, which have respective f-score values of 90.345%, 86.324%, 80.864%, 94.865%, and 83.265%, the RBFN-ESA method has an f-score value of 97.565% under 600 data points.

TABLE III. F-SCORE ANALYSIS FOR RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	88.643	85.865	78.543	91.754	81.765	94.095
200	87.045	84.345	79.465	92.865	82.644	94.345
300	87.345	85.234	78.843	91.245	81.438	95.346
400	88.934	84.846	80.245	92.533	82.835	95.755
500	87.834	86.987	81.258	93.546	83.095	96.346
600	90.345	86.324	80.864	94.865	83.265	97.565

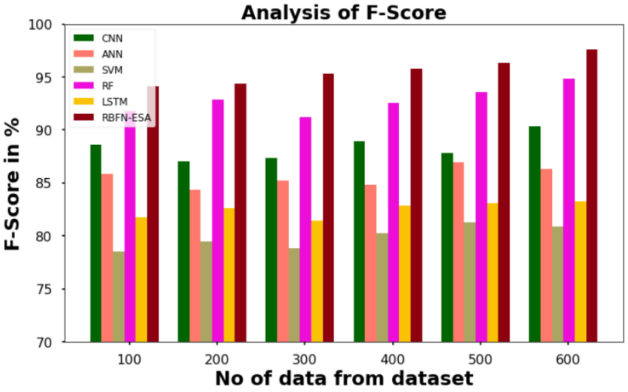


Fig. 5. F-Score analysis for RBFN-ESA method.

E. Accuracy Analysis

Fig. 6 and Table IV compare the accuracy of the RBFN-ESA method to other methods. The graph shows how applying the deep learning with IOT method has improved performance with accuracy. For example, the RBFN-ESA method accuracy value for data 100 is 94.509%, while the accuracy values for CNN, ANN, SVM, RF, and LSTM methods are 79.346%, 89.453%, 84.578%, 90.353%, and 81.756%, respectively. However, the RBFN-ESA method has shown optimal performance over a range of data sizes. Comparing the accuracy values of CNN, ANN, SVM, RF, and LSTM method, which are 83.653%, 88.245%, 86.953%, 92.465%, and 83.543%, respectively, to the RBFN-ESA, which has an accuracy value of 97.467 % is 600 data.

TABLE IV. ACCURACY ANALYSIS FOR RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	79.346	89.453	84.578	90.353	81.756	94.509
200	79.754	89.775	84.965	91.654	82.467	94.356
300	80.645	87.464	85.196	93.464	82.776	95.864
400	80.356	87.865	85.853	92.098	83.854	96.245
500	81.246	88.353	86.257	93.834	81.943	96.865
600	83.653	88.245	86.953	92.465	83.543	97.467

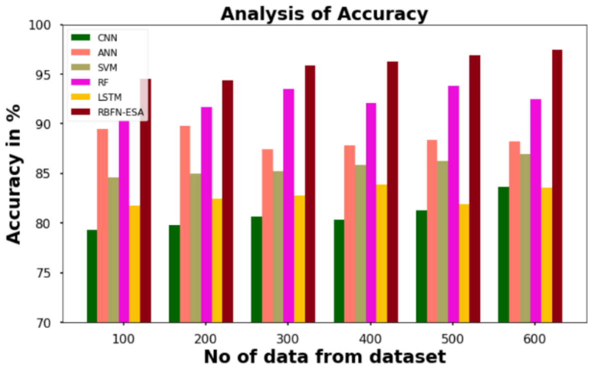


Fig. 6. Accuracy analysis for RBFN-ESA method.

F. Training Validation and Training Loss

Fig. 7 shows training validation and training loss analysis for RBFN-ESA method.

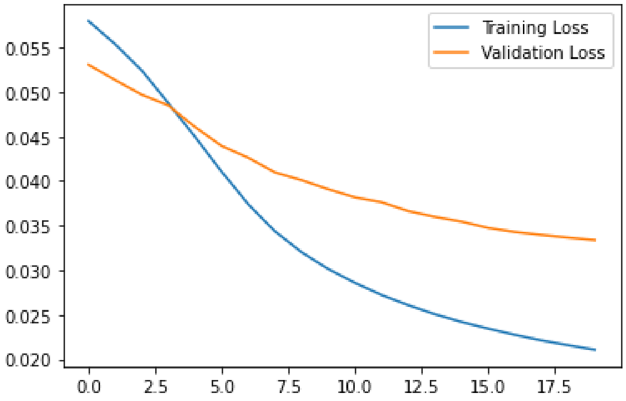


Fig. 7. Training validation and training loss analysis for RBFN-ESA method.

G. Soil Moisture

Fig. 8 and Table V compare the RBFN-ESA method with existing methods for examining soil moisture. The graph shows how soil moisture performance has increased with the deep learning with IOT method. For example, with 100 data, the RBFN-ESA method's soil moisture is 90.16%, whereas the CNN, ANN, SVM, RF, and LSTM methods' soil moisture values are 71.87%, 77.76%, 76.17%, 79.96%, and 84.67%, respectively. However, the RBFN-ESA method has shown optimal performance over a range of data sizes. Similarly, the RBFN-ESA has soil moisture of 94.41% under 600 data, while CNN, ANN, SVM, RF, and LSTM methods have 73.18%, 79.17%, 76.62%, 82.78%, and 88.76%, respectively.

TABLE V. SOIL MOISTURE ANALYSIS FOR RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	71.87	77.76	76.17	79.96	84.67	90.16
200	73.43	76.17	74.36	77.12	83.87	92.65
300	73.12	78.19	75.42	79.65	85.15	91.76
400	70.98	76.54	74.17	81.65	88.44	93.43
500	72.98	78.66	75.77	80.32	88.91	95.17
600	73.18	79.17	76.62	82.78	88.76	94.41

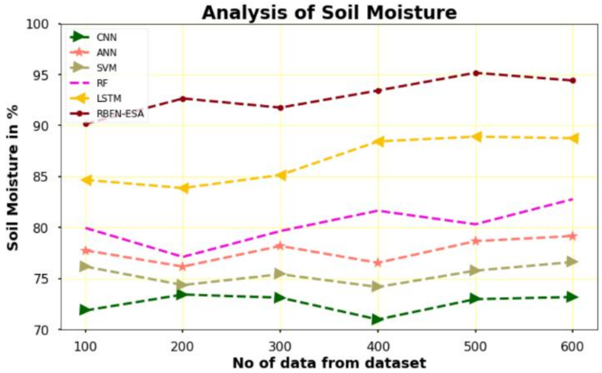


Fig. 8. Soil moisture Analysis for RBFN-ESA method.

H. Energy Consumption Analysis

Table VI and Fig. 9 provide a comparison of the energy consumption of the RBFN-ESA method with existing methods. With 100 data, the CNN, ANN, SVM, RF, and LSTM methods consume 59.324J, 57.276J, 64.865J, 67.897J, and 69.256J of energy, respectively, whereas the proposed RBFN-ESA method uses 54.632 J. In a similar vein, the proposed RBFN-ESA method uses just 56.764 J with 600 data, compared to 62.543 J, 58.721 J, 65.443 J, 68.432 J, and 73.876 J for CNN, ANN, SVM, RF, and LSTM. The recommended method shows enhanced performance with lower energy usage.

TABLE VI. ENERGY CONSUMPTION ANALYSIS FOR RBFN-ESA METHOD

No of data from dataset	CNN	ANN	SVM	RF	LSTM	RBFN-ESA
100	59.324	57.276	64.865	67.897	69.256	54.632
200	60.633	57.642	63.269	66.265	70.765	55.853
300	61.287	58.973	64.249	66.875	70.236	53.842
400	62.843	58.423	63.865	65.854	69.246	55.062
500	61.865	59.053	65.089	68.532	71.663	55.187
600	62.543	58.721	65.443	68.432	73.876	56.764

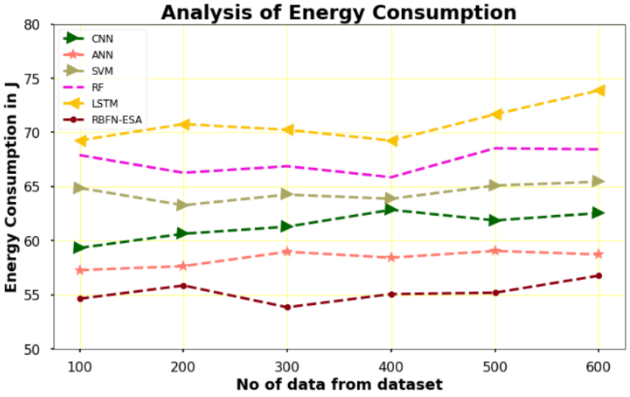


Fig. 9. Energy consumption analysis for RBFN-ESA method.

V. CONCLUSION

Environmental variables, such as relative humidity, temperature, soil temperature, UV rays, etc., impact soil moisture. Technology advancements have greatly enhanced the precision of weather forecasts, and the information can now be used to predict variations in soil moisture. The intelligent irrigation system described in this week's IoT-based planet comprehensive liveliness classification is essential for areas of creation where water and energy are uncommon. Using a qualitative methodology and focusing on secondary data collection, a deep learning model called Radial Basis Function Networks (RBFN) with the Elephant Search Algorithm (ESA) was used for this IoT-based solar energy system for a Future intelligent agriculture system. To supply the necessary operating power, the proposed RBFN-ESA uses a forecasting process from an IoT-based solar cell. The IoT controller reads the data from the humidity, field-based temperature and soil moisture sensors and then outputs the required actuation command signals to drive irrigation pumps. In forecast the value systems of the crucial solar power system variables for a future intelligent agriculture system, the suggested framework RBFN-ESA will gather these data. In terms of defining whether a user will belong to a specific group, the proposed model performed better than other models like Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM). This approach makes use of existing models, such as SVM, RF, LSTM, and convolution neural. We want to conduct further assessments of water savings based on the proposed algorithm with numerous nodes and system cost reduction.

REFERENCES

[1] Kasai, Takeshi, “Preparing for population ageing in the Western Pacific Region,” The Lancet Regional Health–Western Pacific, vol. 6, 2021, https://doi.org/10.1016/j.lanwpc.2020.100069.

[2] A. Walter, R. Finger, R. Huber, and N. Buchmann, “Opinion: Smart farming is key to developing sustainable agriculture,” Proc. Natl. Acad. Sci, vol. 114, pp. 6148–6150, 2017.

[3] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, “Big Data in Smart Farming—A review,” Agric. Syst, vol. 153, pp. 69–80, 2017.

[4] D. Pivoto, P. D. Waquil, E. Talamini, and C. P. S. Finocchio, “Corte, V.F.D.; de Mores, G.V. Scientific development of smart farming technologies and their application in Brazil,” Inf. Process. Agric, vol. 5, pp. 21–32, 2018.

[5] S. Madakam, R. Ramaswamy, and S. Tripathi, “Internet of Things (IoT): A Literature Review,” J. Comput. Commun, vol. 3, pp. 164–173, 2015.

[6] E. C. Leonard, “Precision Agriculture. In Encyclopedia of Food Grains; Elsevier: Amsterdam,” The Netherlands, vol. 4, pp. 162–167. ISBN 9780123947864, 2016.

[7] M. Ponti, A. A. Chaves, F. R. Jorge, G. B. P. Costa, A. Colturato, and K. R. L. J. C. Branco, “Precision agriculture: Using low-cost systems to acquire low-altitude images,” IEEE Computer Graphics and Applications, vol. 36, no. 4, pp. 14–20, July 2016

[8] P. Abouzar, D. G. Michelson, and M. Hamdi, “Rssi-based distributed self-localization for wireless sensor networks used in precision agriculture,” IEEE Transactions on Wireless Communications, vol. 15, no. 10, pp. 6638–6650, Oct. 2016.

[9] C. Brewster, I. Roussaki, N. Kalatzis, K. Doolin, and K. Ellis, “Iot in agriculture: Designing a europe-wide large-scale pilot,” IEEE Communications Magazine, vol. 55, no. 9, pp. 26–33, 2017.

[10] E. Kovac-Andri ˇ c, A. Sheta, H. Faris, and M. Š. Gajdošik, “Forecasting ozone concentrations in the East of Croatia using nonpara- ´ metric neural network models,” J. Earth Syst. Sci, vol. 125, no. 5, pp. 997–1006, 2016

[11] W. Jia, D. Zhao, and L. Ding, “An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample,” Appl. Soft Comput, vol. 48, pp. 373–384, 2016.

[12] K. Mohammadi, S. Khanmohammadi, H. Khorasanizadeh, and K. Powell “A comprehensive review of solar only and hybrid solar driven multigeneration systems: Classifications, benefits, design and prospective,” Appl Energy, pp. 268:114940, 2020.

[13] H. Faridi, A. Arabhosseini, G. Zarei, and M. Okos “Utilization of Soil Temperature Modeling to Check the Possibility of Earth-Air Heat Exchanger for Agricultural Building,” Iran J Energy Environ, pp. 10:260–8, 2019.

[14] Wen-Tai Li, Kannan Thirugnanam, Wayes Tushar, Chau Yuen, Kwee Tiang Chew, and Stewart Tai, “Improving the Operation of Solar Water Heating Systems in Green Buildings via Optimized Control Strategies,” IEEE Transactions on Industrial Informatics, vol. 14, no. 4, pp. 1646-1655, 2018.

[15] C. T. Kone, A. Hafid, and M. Boushaba, “Performance management of ieee 802.15.4 wireless sensor network for precision agriculture,” IEEE Sensors Journal, vol. 15, no. 10, pp. 5734–5747, Oct 2015.

[16] T. Moribe, H. Okada, K. Kobayashl, and M. Katayama, “Combination of a wireless sensor network and drone using infrared thermometers for smart agriculture*,” in 2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pp. 1–2, Jan 2018.

[17] A. R. Concepcin, R. Stefanelli, and D. Trinchero, “Ad-hoc multilevel wireless sensor networks for distributed microclimatic diffused monitoring in precision agriculture,” *in 2015 IEEE Topical Conference on Wireless Sensors and Sensor Networks*, pp. 14–16, Jan 2015.

[18] Wen-Tai Li, Kannan Thirugnanam, Wayes Tushar, Chau Yuen, Kwee Tiang Chew & Stewart Tai, “Improving the Operation of Solar Water Heating Systems in Green Buildings via Optimized Control Strategies,” IEEE Transactions on Industrial Informatics, vol. 14, no. 4, pp. 1646-1655, 2018.

[19] K. Mohammadi, S. Khanmohammadi, H. Khorasanizadeh, K. Powell, “A comprehensive review of solar only and hybrid solar driven multigeneration systems: Classifications, benefits, design and prospective,” Appl Energy, pp. 268:114940, 2020.

[20] H. Faridi, A. Arabhosseini, G. Zarei, and M. Okos “Utilization of Soil Temperature Modeling to Check the Possibility of Earth-Air Heat Exchanger for Agricultural Building,” Iran J Energy Environ, pp. 10:260–8, 2019.

[21] M. S. Mekala, and P. Viswanathan, “CLAY-MIST: IoT-cloud enabled CMM index for smart agriculture monitoring system,” Measurement, vol. 134, pp. 236–244, 2019.

[22] R. Trogo, J. B. Ebardaloza, D. J. Sabido, G. Bagtasa, E. Tongson, and O. Balderama, “SMS-based smarter agriculture decision support system for yellow corn farmers in Isabela,” in Proceedings of the 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), Ottawa, Canada, June 2015.

[23] S. K. Roy, S. Misra, N. S. Raghuwanshi, and S. K. Das, “AgriSens: IoT-based dynamic irrigation scheduling system for water management of irrigated crops,” IEEE Internet Things J., vol. 8, no. 6, pp. 5023–5030, 2020.

[24] C. M. Angelopoulos, G. Filios, S. Nikoletseas, and T. P. Raptis, “Keeping data at the edge of smart irrigation networks: a case study in strawberry greenhouses,” Comput. Netw, vol. 167 pp. 107039, 2020.

[25] B. Keswani, et al., “Adapting weather conditions based IoT enabled smart irrigation technique in precision agriculture mechanisms,” Neural Comput. Appl, vol. 31, no. 1, pp. 277–292, 2019.

[26] R. Prabha, E. Sinitambirivoutin, F. Passelaigue, and M. V. Ramesh, “Design and development of an IoT based smart irrigation and fertilization system for chilli farming, in: 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),” IEEE, pp. 1–7, 2018.

[27] Ambika, G.N., Singh, B.P., Sah, B., Tiwari, D. “Air quality index prediction using linear regression”, International Journal of Recent Technology and Engineering, 2019, 8(2), pp. 4247–4252.

[28] Y. A. Rivas-Sánchez, M. F. Moreno-Pérez, and J. Roldán-“Cañas, Environment control with low-cost microcontrollers and microprocessors: application for green walls,” Sustainability, vol. 11, no. 3, pp. 782, 2019.

[29] S. Ali, H. Saif, H. Rashed, H. AlSharqi, and A. Natsheh, “Photovoltaic

energy conversion smart irrigation system-Dubai case study (goodbye overwatering & waste energy, hello water & energy saving), in: 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC)(A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC),” IEEE, pp. 2395–2398, 2018.

[30] J. R. dela Cruz, R. G. Baldovino, A. A. Bandala, and E. P. Dadios, “Water usage optimization of Smart Farm Automated Irrigation System using artificial neural network, in: 2017 5th International Conference on Information and Communication Technology (ICoIC7),” IEEE, pp. 1–5, 2017.

[31] S. S. Mathurkar, N. R. Patel, R. B. Laanjewar, and R. S. Somkuwar, “Smart sensors based monitoring system for agriculture using field programmable gate array,” in Proceedings of the International Conference on Circuits, Power and Computing Technologies (ICCPCT-2014), Nagercoil, India, March 2014.

[32] Ambika, G.N., Suresh, Y. “Optimal Deep Convolutional Neural Network Based Face Detection and Emotion Recognition Model”, International Journal of Intelligent Systems and Applications in 2023, 11(3), pp. 841–` 849.

[33] Z. Unal, ‘‘Smart farming becomes even smarter with deep learning— A bibliographical analysis,’’ IEEE Access, vol. 8, pp. 105587–105609, 2020, doi: 10.1109/ACCESS.2020.3000175.

[34] S. B. Saraf and D. H. Gawali, “IoT based smart irrigation monitoring and controlling system,’’ in Proc. 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. (RTEICT), Bengaluru, India, May 2017, pp. 815–819, doi: 10.1109/RTEICT.2017.8256711.

[35] R. Laurett, A. Paço, E. W. Mainardes “Antecedents and consequences of sustainable development in agriculture and the moderator role of the barriers: Proposal and test of a structural model,” J Rural Stud, pp. 86:270–81, 2021.

[36] Ambika, G.N., Suresh Y. “An Efficient Deep Learning with Optimizatio Algorithm for Emotion Recognition in Social Networks”, International Journal of Advanced Computer Science and Applications, 2023, 14(8), pp. 206–215.

[37] Ambika, G.N., Suresh, Y. “Mathematics for 2D face recognition from real time image data set using deep learning techniques”, Bulletin of Electrical Engineering and Informatics., 2024, 13(2), pp. 1228–1237.

Subjectivity Analysis of an Enhanced Feature Set for Code-Switching Text

Emaliana Kasmuri, Halizah Basiron*
Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—The phenomenon of code-switching has posed a new challenge to the linguistic computing area. Conventionally, the computer will process monolingual text or multilingual text. However, code-switching is different from this kind of text. Two or more languages are used to construct a piece of code-switching text, particularly a code-switching sentence. It is challenging for the computer to process a piece of code-switching text with languages that exist simultaneously. The challenge is more intense for the computer in subjectivity analysis, where the computer should distinguish subjective from objective code-switching text. This paper proposed three feature sets for subjectivity analysis on Malay-English code-switching text: Embedded Code-Switching Feature Sets, Unified Code-Switching Feature Sets, and Stylistic Feature Sets. These feature sets were enhanced from the monolingual feature set of subjectivity analysis. Experiments were conducted using the data harvested from Malay-English blogs. These data were labelled as either subjective or objective. Two machine learning classifiers – the Support Vector Machine (SVM) and Naive-Bayes, were used to evaluate the classification performance of the proposed feature sets. The experiments were carried out on individual feature sets and the combination of them. The results show the classification performance from combining the unified and stylistic feature sets surpassed other proposed feature sets at 59% accuracy. Therefore, it is concluded that the combination of unified and stylistic feature sets is necessary for the subjectivity analysis of Malay-English code-switching text.**

Keywords—Subjectivity analysis; code-switching; enhanced feature sets; Malay-English text

I. Introduction

It is common for a person to master multiple languages. Mastering multiple languages benefitted an individual in various ways, including vocabulary enrichment, communication improvement, and knowledge improvement through information exchange and sharing. It is also a common situation to see a multi-lingual person communicate, either in spoken or written, using a mix of languages. The use of mixed languages is known as code-switching [1].

A code-switching text is a piece of text that is constructed using words from at least two languages. For example, 'I really like matte satin silk nak buat collection tapi I am still a student tak ada income if dapat giveaway ni mesti best'. In the example, the sentence is constructed using English (underlined) and Malay (italic) words. Code-switching may occur within a text where the words from the second language interleave in between the words from the first language, or the first part of the text was constructed using the first language while the remaining part was constructed with the second language. For example, 'Aspirasi saya adalah banyak peluang dari segi pendidikan, ekonomi, sukan dan sebagainya untuk masyarakat khas for them to excel in those field'.

The code-switching phenomenon is not new in the linguistic research domain. It was acknowledged in the 1980s [2]. However, with the advancement of computer technology, the use of code-switching in open platforms such as blogs and social media has become more apparent than before. This situation has posed a new challenge to language computational areas in text analysis and machine comprehension of this new kind of text. In general, most of the research effort has been channelled towards computational and knowledge augmentation on monolingual and multilingual text. The presence of two languages simultaneously in a code-switching text was not considered in the current subjectivity analysis study.

A set of research questions are drawn to address the code-switching issues in subjectivity analysis. The research questions are as follows:

1) What are the most effective feature sets for subjectivity analysis in Malay-English code-switching text?

2) How do different machine learning classifiers perform in subjectivity classification when using the proposed feature sets?

3) What is the impact of combining various enhanced feature sets on the accuracy of subjectivity analysis in code-switching text?

Three objectives were drawn to align with the research questions. The first objective is to develop enhanced feature sets that effectively analyse subjectivity in Malay-English code-switching text. The second objective is, to compare the performance of machine learning classifiers using the proposed feature sets for subjectivity analysis. Finally, the last objective is to assess the accuracy improvement achieved by combining different feature sets for subjectivity classification in code-switched text.

Therefore, this article proposes a method that enhances feature sets to analyse subjectivity in Malay-English code-switching sentences to achieve the objectives. The enhanced feature sets used the subjectivity feature from both languages to represent the subjectivity of the code-switching feature sets. The enhanced feature sets consist of embedded, unified, and stylistic feature sets. Two machine learning classifiers, Naïve-Bayes (NB) and Support Vector Machine (SVM), were used to evaluate the feasibility of the feature sets classifying the Malay-English code-switching sentences into subjective and objective classes.

*Corresponding Author, halizah@utem.edu.my

This research contributes to the growing field of computational linguistics by addressing the challenge of subjectivity analysis in code-switching text. The findings from this study can be applied to improve text processing systems used in social media, blogs, and other platforms where code-switching is prevalent. The proposed enhanced features set that contains embedded, unified, and stylistic feature sets provide a foundation for future research and development in computational methods for analysing subjectivity in code-switching text.

The article's structure is as follows: In Section II, we review related works and discuss research motivation. Section III presents a description of our proposed solution and the features used. Section IV describes the dataset used in the experiment. Section V gives the details of our experiments and discusses the results. Sections VI and VII present the discussion and limitations, respectively. Lastly, Section VIII concludes and summarizes our work in this article.

II. Related Work

Subjectivity analysis is a linguistic computational task that determines the existence of a private state such as opinion, emotion and stance in a piece of textual document. Subjectivity analysis is a precursor to tasks such as sentiment analysis, and emotion classification.

Ting et al. investigated subjectivity classification using a window-based self-attention approach [2]. The approach improves the sentence encoding by leveraging context within variable-sized windows, instead of the entire sentence. Different sizes of windows (the size of surrounding words) were used to determine the importance of each word using trigrams or five-grams. The max-pooling method was used to extract the most relevant feature. As a result, multiple windows were captured from various phases where meaningful combinations of words were evident. This method simplifies attention computation, working directly on word embeddings and relying on n-gram features to provide context-aware sentence representations. The author used monolingual Cornell's movie data set to classify the review into subjective and objective classes. The proposed method performed at 94.60% accuracy.

Belisario et al. studied subjectivity analysis for Portuguese book reviews [3]. The authors used 350 reviews that were equally divided into subjective and objective sentences. The features of these sentences were extracted and represented using several methods including Sentilex-PT and WordnetAffectBR for the lexicon-based method, global centrality graph-based method using Eigenvector Centrality, Katz Index and PageRank and Naïve-Bayes, SVM and Neural Network for machine learning based method. The result reveals the Neural Network outperformed other methods at 83.20% accuracy.

Kasnesis et al. compare three transformer-based architectures to classify the the 10,000 Cornell movie review sentences into subjective and objective classes [3]. The authors used Bidirectional Encoder Representation (BERT), Robust Optimised BERT Pretraining Approach (RoBERTa) and Efficient Learning Encoder Classifies Token Replacements Accurately (ELECTRA) to evaluate the effectiveness of the subjectivity classification. ELECTRA achieved the highest accuracy score, 98.30% among the three.

Al Hamoud et al. analyse subjectivity expressed from an online English political and ideological debate forum [4]. Various controversial topics were debated in the forum including abortion, creationism, gay rights, the existence of God, gun rights and health care. The authors have created a dataset of 53, 453 sentences from the forum. These sentences were labelled as either subjective or objective. The features of the dataset were represented using one-hot encoding and GloVe pretraining embedding vectors. These features were experimented with using six deep learning models to find the most effective model that could classify the dataset into subjective/objective classes. The models are Long Short Term Memory Networks (LSTM), Gated Recurring Units (GRU), bidirectional GRU, bidirectional LSTM, LSTM with attention and bidirectional LSTM with attention. The results of the experiment show that LSTM with attention performed the best out of all deep learning models – 97.39% accuracy.

The studies that were presented above were working on monolingual textual documents – that is English and Portuguese. The excellent performance result in studies has shown the proposed classification solutions have generated effective classification models that has outperformed the other models in the respective experiment. Even though the studies can produce effective subjectivity classification models, their feasibility on code-switching text is still unknown. Therefore, this article attempts to fill the gap using Malay-English code-switching text for subjectivity analysis.

The issue of code-switching in linguistic computation has been addressed as early as the 1980s [5]. However, the issue received proper attention from the linguistic computation area due to the availability of the data sets and the maturity of computer language processing tools. With the advance of social media, the research of linguistic computation using code-switching text has started to receive attention. Among the research on code-switching computational are text generation [6], [7], [8] and sentiment analysis [9], [10], [11].

One of the hurdles in code-switching linguistic computation is data scarcity. The advancement of social media has accelerated the phenomenon of code-switching. Therefore, data become more available in this research area. However, the amount of data is insufficient for various tasks in linguistic computation. Code-switching data is artificially generated to support these tasks. Hu et al. generate code-switching training text to improve code-switching automatic speech recognition (ASR) [8]. The text injection method known as PaLM 2 and prompt tuning were used to generate the training data. The experiment to generate the data was carried out using the Large Language Model (LLM). The result shows that the proposed method achieves a 3.60% Word Error Rate (WER) for Mandarin-English. The research concludes leveraging LLMs for text generation and text injection benefits code-switching ASR tasks.

Three methods were compared to select the best method that generates code-switching data for Egyptian Arabic-English Hamed has been compared [7]. The first method is lexical replacement, which replaces a number of random Arabic words with English words using code-switching point assignment and point prediction. Code-Mixing Index (CMI) was used to measure the performance of this method. CMI indicate the level of mixing from both languages. Both methods achieve 28.00% and 25.00% CMI. The second method used linguistic theories - Equivalence Constraint (EC) Theory and Matrix Language Frame (MLF) Theory. Both achieved 30.00% and 25.00% CMI. The final method used back translation. The method performed at 18.00% CMI.

Research by Chi et al. also addressed code-switching data scarcity [6]. Monolingual BERT models (Mandarin and English) were used to train a Transformer encoder-decoder model to translate between any pair of languages. The translated sentences are forced to code-switch to any degree. Disjoint union of vocabulary from two languages were used as parallel text. Grid beam search is the method to force the degree of the code-switching text. The proposed method performed at 30.58% WER.

A deep convolutional neural network (CNN) was used to determine the sentiment expressed in a comment that contains a mixture of English and Hindi languages [9]. A dataset that contains 17,155 comments was created, and the comments were extracted from a governmental website. Each comment was manually labelled as either positive or negative. The study used Word2Vec to represent the English words. The Hindi words embedding were developed using Word2Vec and trained using Hindi Wikipedia and other sources. The solution used on the dataset has yielded 67.00% of accuracy.

A study on a mixture of English-Spanish that leverages existing English and Spanish text analytical tools has been carried out using a supervised learning algorithm [10]. In this study, 3,062 tweets that were labelled as positive and negative were used. Four atomic features, which are word, lemma, psychometric properties and part-of-speech (POS) tags were used as basic features. The best accuracy obtained is 59.34% using combination lemmas and psychometric properties as features.

A recent study of sentiment analysis using a mixture of English-Hindi and English-Bengali tweets has demonstrated using deep learning [11]. The study classifies the dataset into positive, negative and neutral. Three deep learning algorithms, which are BiDirectional Long Short-Term Memory (BiLSTM) Convolutional Neural Network (CNN), a Double BiDirectional Long Short-Term Memory (D-BiLSTM) and an Attention-based model, were used in the experiment. Pre-trained word embeddings, which are Global Vector (GloVe) and Bidirectional Encoder Representations from Transformers (BERT), were used in the study. The study has revealed that accuracy performance between 42.00% and 68.00% was achieved using a 10-fold cross-validation classification. The result from the experiment shows the best performance for the English-Hindi dataset was achieved by the deep learning attention model using GloVe with 68.00% accuracy, whereas, for the English-Bengali dataset, the best performance was achieved by the deep learning attention model with 67.00% accuracy.

All the research efforts described in this section focus on code-switching sentiment classification and text generation. The attempt for subjectivity classification on code-switching text, especially for subjectivity analysis on Malay-English code-switching text, was found to be limited at the time this research was conducted. Therefore, this article is attempts to find out a feasibile solution for subjectivity analysis on Malay-English code-switching text.

III. Enhanced Feature Sets of Subjectivity Analysis for Code-Switching Text

The syntactical and semantical features were used to analyse subjectivity in a monolingual text such as English [12]. These features are described in Table I. These features were enhanced for a Malay-English code-switching text.

TABLE I. INITIAL FEATURE SET FOR SUBJECTIVITY ANALYSIS

Syntactical Features	Semantical Features
1. Sentence that located in beginning of a paragraph. 2. The co-occurrence of words and punctuation.	1. Pronoun 2. Adjective 3. Cardinal number 4. Modal (other than will) 5. Adverb (other than not)

The enhanced feature sets are Embedded Code-Switching Feature Sets, Unified Code-Switching Feature Sets and Stylistic Feature Sets. The Embedded Code-Switching Feature Set include the Malay text analytical components in the code-switching feature set representation. The Unified Code-Switching Feature Set fused two text analytical components into a unified feature set. The stylistic feature sets used non-vocabulary features to represent the subjectivity clues in the code-switching text. Fig. 1 illustrates the proposed feature sets.

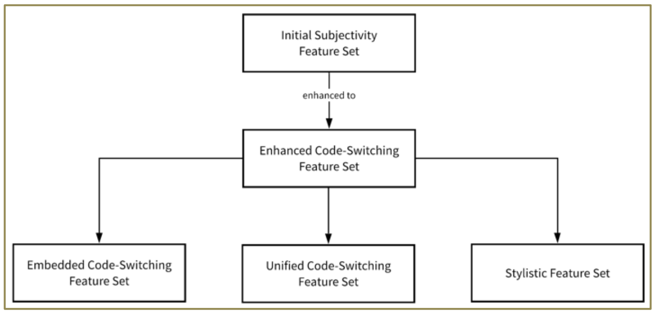

Initial Subjectivity Feature Set
enhanced to
Enhanced Code-Switching Feature Set
Embedded Code-Switching Feature Set
Unified Code-Switching Feature Set
Stylistic Feature Set

Fig. 1. Proposed feature set for code-switching subjectivity analysis.

A. Embedded Code-Switching Feature Set

The subjectivity analysis feature set for a monolingual text is derived from a monolingual part-of-speech (POS) tag. A code-switching sentence is constructed using at least two different languages. Therefore, two POS taggers are used to produce the feature set for the code-switching sentence. The union of the POS tags from two languages produces the embedded code-switching feature set. The union process is shown in Fig. 2.

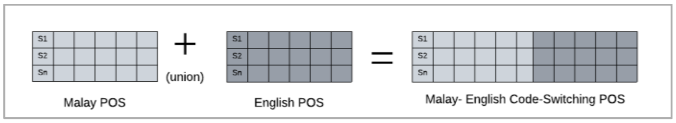


Fig. 2. The union process to construct embedded code-switching POS tags.

The Hawkin-UKM Malay and Penn Treebank POS tags are used to produce the embedded code-switching feature sets. The POS taggers consist of several tags to represent a single POS. For example, the Hawkin Malay POS tagger used ADJ, ADJS and ADJT to represent an adjective word, while the Penn Treebank POS used JJ, JJR and JJS. These POS tags are combined and grouped. Table II shows the grouping of the POS tags that represent the subjectivity features for the Malay-English code-switching text.

TABLE II. MALAY-ENGLISH CODE-SWITCHING POS TAGS

Part of Speech	Malay POS Label	English POS Label
Adjective	ADJ, ADJS, ADJT	JJ, JJR, JJS
Adverb	ADV	RB, RBR, RBS, WRB
Noun	KN, KNG, KNK, KNT, KPB	NN, NNP, NNPS, NNS
Verb	KK, KKIA, KKIK, KKIW, KKT	VB, VBD, VBG, VBN, VBP, VBZ
Pronoun	KGDD, KGDP, KGDT, KGNT, @KG	PRP, PRP$, WP
Conjunction	KH	CC, IN
Cardinal number	KBIL	CD
Modal	MD	MD

Modal POS was absent in the initial Hawkin-UKM Malay POS tagger. Modal is included in the Malay POS for a balanced and uniform feature. The modal POS tag is defined as MD. A list of English modals retrieved from MyEnglishPages.com[1] is translated into Malay using Cambridge Online English-Malay Dictionary[2], and Kamus Dwibahasa Bahasa Inggeris – Bahasa Melayu [13], [14]. Thirteen English modals were translated into Malay. The translation process produces 30 Malay words with equal meaning to the English modal.

The process to generate the Malay-English code-switching feature set is shown in Fig. 3. The process consists of five phases. The process begins with the sentence tokenization process in Phase 1. The code-switching sentences retrieved from the repository are tokenized. Tokenization is a process that breaks the words into the sentence into individual pieces. The individual piece is called a token. The space between the words is used as a boundary to mark the beginning and end of a word. This is known as a delimiter.

The process continues with POS tagging in Phases 2 and 3. The tokens are tagged using the Hawkin Malay and Penn Treebank POS taggers. Phases 2 and 3 produced two sets of tagged tokens. In Phase 4, the English POS-tagged tokens are embedded into the Malay POS-tagged tokens. The phase standardized the POS tags using the group described in Table III. In Phase 5, the presence of the feature is coded as 1, while others are coded as zero, indicating the absence of the feature. The features are extracted and converted into a matrix. This matrix is known as an embedded code-switching feature set.

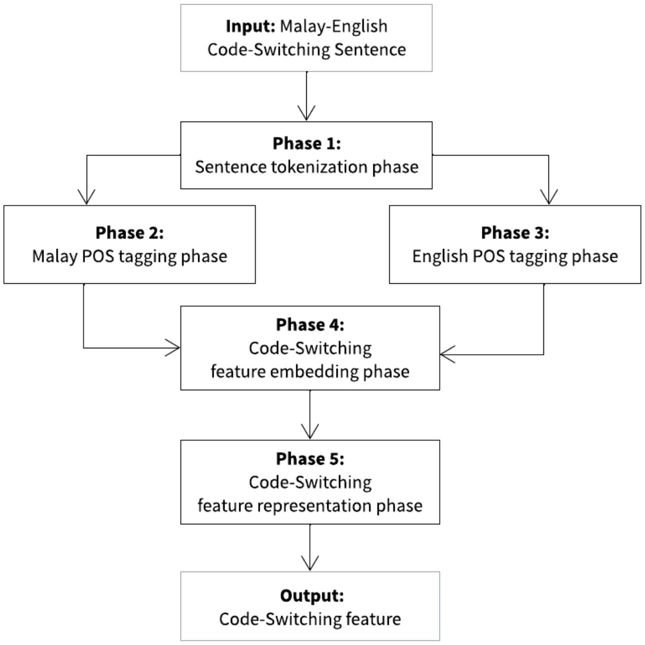


Fig. 3. The process to generate embedded code-switching feature set.

TABLE III. EMBEDDED CODE-SWITCHING SUBJECTIVITY FEATURE SET

Feature	Description
ADJ_MS	Presence of adjective for Malay words
ADV_MS	Presence of adverb for Malay words
NN_MS	Presence of noun for Malay words
VB_MS	Presence of verb for Malay words
PR_MS	Presence of pronoun for Malay words
CC_MS	Presence of conjunction for Malay words
CD_MS	Presence of cardinal number for Malay words
MD_MS	Presence of modal for Malay words

B. *Unified Malay-English POS Feature Set*

The second feature set is the unified Malay-English POS feature set. The process to produce the unified code-switching feature set is shown in Fig. 4. The process consists of six phases. The process starts with Phase 1, which tokenized the sentences. The tokens are POS-tagged using the Hawkin-Malay and Penn Treebank POS tags in Phase 2 and Phase 3. The phases produce two sets of tokens. However, there were discrepancies in token lengths identified in Phase 4. The length of the token is standardized by removing the additional token.

In Phase 5, the Malay and English POS Tags are unified using a rule-based algorithm. The algorithm was adopted from [15], as shown in Algorithm 1. The algorithm requires three (3) sentences: the sentence with language label tags, the sentence with English POS tags and the sentence with Malay POS tags. Each word in the sentence with language label tags will be mapped to either a sentence with English POS tags or a sentence with Malay POS tags. The mapping of each word will be according to the language label.

[1] https://www.myenglishpages.com/site_php_files/grammar-lesson-modals.php

[2] https://dictionary.cambridge.org/dictionary/english-malaysian/

Algorithm 1: Unified_cs_pos_tag

```
Input:
sentence_cs: Malay-English code-switching sentence with language tags
sentence_en_pos: Malay-English code-switching sentence with English POS tags
sentence_my_pos: Malay-English code-switching sentence with Malay POS tags
sentence_unified_POS = " "
For each word in sentence_cs
    Get the language of the word
    If the language labelled for the word is either as English or Shared
            Get the matching word from sentence_en_pos
            Concatenate word and POS Tag into sentence_unified_POS
    Else if the word is labelled as Malay
            Get the matching word from sentence_my_pos
            Concatenate word and POS Tag into sentence_unified_POS
    Else
            Discard the word
    End
End
Return sentence_unified_POS
```

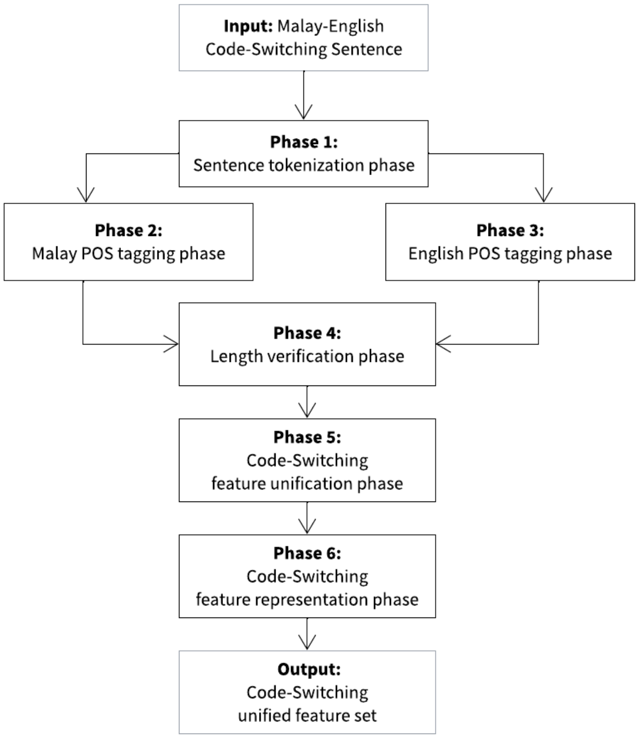

Input: Malay-English Code-Switching Sentence
Phase 1: Sentence tokenization phase
Phase 2: Malay POS tagging phase
Phase 3: English POS tagging phase
Phase 4: Length verification phase
Phase 5: Code-Switching feature unification phase
Phase 6: Code-Switching feature representation phase
Output: Code-Switching unified feature set

Fig. 4. The process to generate a unified POS feature set.

C. *Stylistic Feature*

The stylistic features are non-vocabulary words representing the author's emotional state in a text. The stylistic features that influence the code-switching text are exclamation marks, emoticons, exaggeration, intensifiers and interjections. An emoticon is a representation of the facial expression, mood and emotion of an author that is created by combining numerous keyboard strokes. An emoticon strengthens the message conveyed in the text [16]. It is observed that the widespread usage of emoticons in electronic documents has indicated the development of emotion [17]. An emoticon is also used to clarify the intent of the electronic message, such as sarcasm and criticism [16]. The findings of these studies have shown that emoticons are used in a subjective textual document. Table IV shows examples of emoticons and their underlying meaning.

Creative spelling with exaggeration of characters is another stylistic feature considered in the code-switching text. The exaggeration of characters is defined as the repetition of characters in a word that occurs more than two times. The repetition of characters in the text indicates the expression of information intensifying.

TABLE IV. REPRESENTATION OF EMOTICON AND ITS UNDERLYING MEANING

Emoticon	Meaning
:) or :-) or (^_^)	An emoticon that represents a smiling face that indicates the author is happy or pleased.
:(or :-(	An emoticon that represents a frown face that indicated the author is sad or unhappy.
;) or ;-)	An emoticon that represents a winking, indicating being flirtatious.

An interjection is used as a form of human expression of feeling in an electronic document. Interjection represents an immediate feeling by not literally describing it [18]. Some interjections are language-independent, such as "wow", "haha", and "hmm". Whilst some interjections are language-dependent, such as "jeng", "gap", and "seh", which are commonly used by Malay speakers, were found in the code-switching sentences. Table V shows some examples of interjections and their meaning.

TABLE V. EXAMPLE OF INTERJECTIONS AND ITS MEANING

Interjection	Meaning
wow	Represents astonishment
haha	Represent regular laughter
hmm	Represent thinking or hesitation
jeng	Represent surprise. Usually appear multiple times such as jeng…jeng..jeng
ngap	Represent act of eating

IV. BUILDING MALAY-ENGLISH SUBJECTIVE DATA SET

The data used in this research was harvested from 45,964 Malay-English blog posts. The blog is chosen as the source of data for numerous reasons: 1) The content is rich with objective and subjective information 2) The content was written using a mixture of Malay and English 3) The blog is still relevant 4) The blog is publicly available 5) The casual writing style.

Preparing the data set begins with extracting the content of the blog post. For that purpose, a Python program using the BeautifulSoup module was developed. The blog content was separated into individual sentences, and the distribution of Malay and English words was computed using the procedure shown in Algorithm 2.

Algorithm 2: Compute_MS_EN_distribution

```
For every sentence in the database , sn
    $t_{malay} = 0$, $t_{english} = 0$, $t_{shared} = 0$, $t_{ood} = 0$
    Separate the sentences into individual word, $s=\{w1, w2, w3, \ldots ,wn\}$
    For every word in the sentence, wn
            If $((wn \in mswn) \cap (wn \notin enwn))$
            $t_{malay}$ += 1
            else if $((wn \notin mswn) \cap (wn \in enwn))$
            $t_{english}$ += 1
            else if $((wn \in mswn) \cap (wn \in enwn))$
            $t_{shared}$ += 1
            else
            $t_{ood}$ += 1
            End
    Record the language distribution for this sentence, sdn = { $t_{malay}$, $t_{english}$, $t_{shared}$, $t_{ood}$}
    End
End
Return record of language distribution
```

After that, the Malay-English code-switching sentences were extracted using the procedure shown in Algorithm 3. A Malay-English code-switching sentence should contain at least one Malay and one English functional word. This research defined a functional word as a word that is not categorized as either a stop word or the name of an entity.

Algorithm 3: selecting_MS_EN_sentences

```
Get language distribution for every sentence from the database
For each language distribution of a sentence
    Get Malay words distribution
    Get English words distribution
    If ((Total Malay words distribution >= 1) and (Total English words distribution >= 1))
            Label the sentence as MS-EN-CS
     Else if ((Malay words distribution >= 1) and (English words distribution == 0))
            Label the sentence as MS
     Else if ((Malay words distribution == 0) and (English words distribution >= 1))
            Label the sentence as EN
End
End
```

The procedure has extracted 93,796 Malay-English sentences. Sentences containing 3 to 25 words are selected to be labelled as subjective or objective. Mohammad (2016) and Belisário et al. (2020) used a similar number of words per sentence [19], [20]. Sentences with less than three words were considered uninformative. Sentences containing more than 25 words contain overwhelming information that increases the complication of the annotation task. 56,207 sentences were selected after discarding the unwanted sentences. These sentences were labelled by two annotators using the annotation scheme described in Table VI.

The annotation process has produced 35,067 Malay-English code-switching sentences - 25,164 were subjective, while 9,903 were objective sentences. These sentences will be used as a dataset in the experiment section. The result is shown in Table VII.

TABLE VI. ANNOTATION SCHEME FOR MALAY-ENGLISH CODE-SWITCHING CORPUS

Label	Language			Description of sentence	
	Malay	*English*	*Malay-English*	*Fact*	*Opinion*
EN-OPI		√			√
EN-FAC		√		√	
MS-OPI	√				√
MS-FAC	√			√	
CS-EN-OPI			√		√
CS-MS-OPI			√		√
CS-FAC			√	√	

TABLE VII. DISTRIBUTION OF ANNOTATED SENTENCES

Language / Labels	Subjective	Objective
Malay-English	25, 164	9,903
Code-Switching	25, 164	9,903
English	3, 957	1, 197

The dataset contains 9,903 objective sentences and 25,164 subjective sentences. The number of subjective sentences is two and a half times greater than the number of objective sentences, making the data imbalanced. However, a machine learning algorithm works best with a balanced dataset.

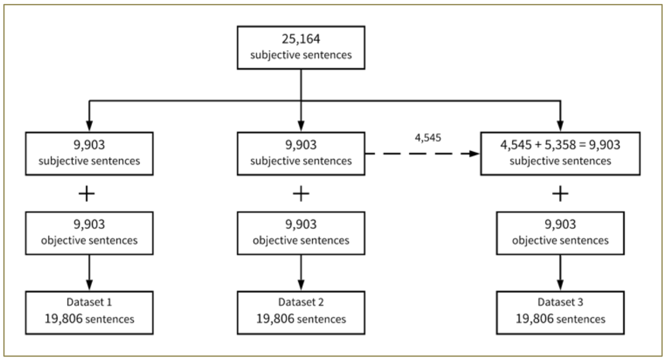


Fig. 5. Distribution of dataset.

This research divided the subjective sentences into three parts to create an equal number of sentences to objective sentences, as shown in Fig. 5. Krawczyk (2016) used a similar approach [21]. This research named these parts as Dataset 1, Dataset 2 and Dataset 3. With this distribution, all datasets have an equal number of subjective and objective sentences.

V. RESULT

Nine experiments were designed and carried out to evaluate the proposed feature sets' ability to analyse the subjectivity in Malay-English code-switching text. The experiments were carried out using two different machine learning classifiers, Naïve-Bayes and Support Vector Machine (SVM). Two experiments were carried out exclusively for English and Malay sentences using the initial feature sets as baseline experiments. The baseline experiments are established as comparisons for the proposed feature sets. Other experiments were conducted to identify the optimal feature sets and machine learning classifiers that function well with subjectivity classification for Malay-English code-switching text at the sentence level.

A. *Baseline Feature Sets Performance Results*

The results from the baseline experiments will be compared to those of the other experiments. The baseline experiments are performed separately, considering only English and Malay initial feature sets. These feature sets are trained and tested using two different classifiers and three different datasets. The accuracy performance of the experiments is shown in Fig. 6. The results were obtained after performing 10-fold cross-validation for both classifiers.

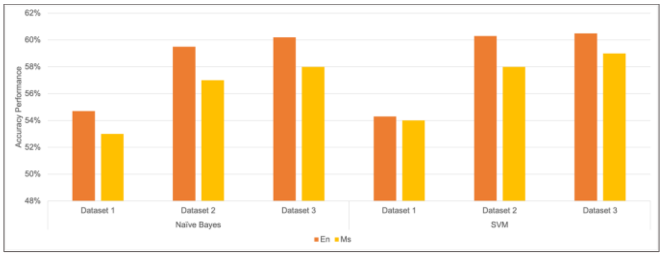


Fig. 6. Accuracy performances of baseline initial features models using different classifiers across multiple datasets.

Fig. 6 shows the baseline feature set performed at 55.00% accuracy for the English (En) feature set and 53.00% accuracy for the Malay (Ms) feature using Dataset 1 using the Naïve-Bayes classifier. The performance of the baseline feature sets using Dataset 2 is 59.00% accuracy for the English (En) feature set and 57.00% accuracy for the Malay feature set, using the same classifier. The baseline feature sets yielded 60.00% accuracy and 58.00% accuracy, respectively, for the English initial feature set and Malay initial feature set, using Dataset 3 and the same classifier. Dataset 3 has the highest accuracy performance among the three datasets using the Naive-Bayes classifier.

The accuracy of the baseline initial feature sets was lower in Dataset 1 using the SVM classifier, which is 54.00% for both, as shown in Fig. 9. The accuracy of the models increased significantly using Dataset 2 using the same classifier, where the English initial feature set achieved 60.00% accuracy, and the Malay initial feature set achieved 58.00% accuracy. The accuracy performance of the English initial feature set showed a slight improvement using Dataset 3. However, the accuracy performance for the Malay feature set increased by 1.00% to 59.00% using Dataset 3 and the SVM classifier.

In general, the bar chart in Fig. 9 showed increments of accuracy performance from Dataset 1 to Dataset 3 in both classifiers. Comparatively, the accuracy performances of the initial feature sets between the datasets and classifiers show a consistent increment pattern. Therefore, the baseline results can be compared with the proposed feature set models.

B. *Embedded Code-Switching Feature Set Performance Results*

The same setting of experiments that were carried out for the baseline experiment is used to evaluate the embedded code-switching feature set. The result of the experiment is shown in Fig. 7. The accuracy results show the embedded feature sets perform at equal accuracy, 54.00% for Dataset 1, using both classifiers, Naive-Bayes and SVM classifier. The result also shows that Dataset 3 outperforms other datasets for both classifiers.

There is an increment pattern of accuracy performance between the three datasets for both classifiers. There is a significant accuracy increment between Dataset 1 and Dataset 2 using both classifiers. Datasets 1 and 2 performed at 54.00% and 57.00% using the embedded feature set and Naive Bayes classifier. The accuracy increased by 3.00%. The accuracy performance increased by 5.00% using the SVM classifier between Datasets 1 and 2. Datasets 1 and 2 performed at 54.00% and 59.00%, respectively, with the SVM classifier. The accuracy increased by 1.00% between Dataset 2 and 3 using both classifiers. It is also noted that there are noteworthy accuracy differences between Dataset 1 and 3 for both classifiers. The differences in accuracy performance between Datasets 1 and 3 are 4.00% differences using the Naive Bayes classifier and 5.00% using the SVM classifier.

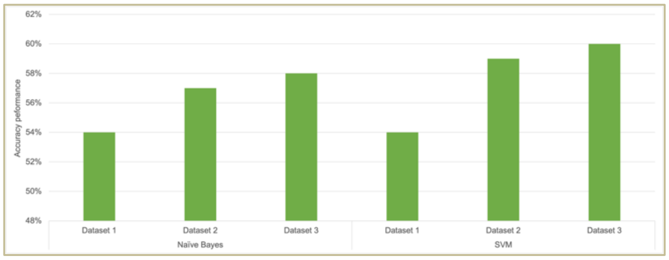


Fig. 7. Result of accuracy performance for subjectivity classification on Malay-English code-switching text using embedded feature set.

The significant differences in accuracy performance between the datasets are due to the presence of the subjectivity features in the dataset. The presence of the Malay and English subjectivity features is more significant in Datasets 2 and 3 compared to Dataset 1. The differences reveal the SVM classifier works better to identify the subjectivity presence in Malay-English code-switching sentences using the embedded feature-set, with 60.00% accuracy, where Datasets 3 outperformed others. Therefore, using both classifiers, the embedded code-switching feature set can be used to distinguish subjective and objective Malay-English code-switching sentences.

The accuracy performance results of the embedded feature set were compared with the initial feature sets, which are the English and Malay initial feature sets. The comparison is shown in Fig. 8. The accuracy performance of the English initial feature set (En) was superior in all datasets and classifiers. The superiority of accuracy performances indicates the English initial feature sets are stable and robust. However, the accuracy performance of the embedded feature set is better in comparison to the Malay initial feature set (Ms) in three experiment settings, which are Dataset 1 with Naive Bayes classifier and Datasets 2 and 3 using the SVM classifier. The accuracy performance of the feature set is on par with the other three experiment settings. This is clearly shown using Datasets 2 and 3 with the Naive-Bayes classifier and Dataset 1 using the SVM classifier.

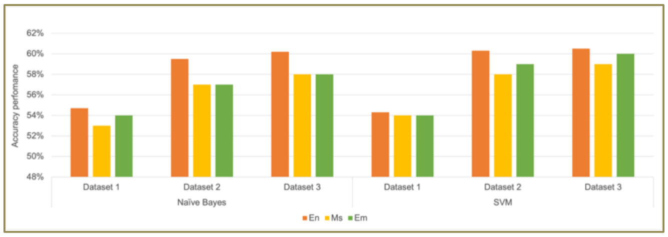


Fig. 8. Comparison of embedded feature sets with baseline initials feature sets.

The accuracy performance of the embedded feature set (Em) is either superior or on par with the Malay initial feature set (Ms). It is concluded that the Em feature set performed as good as the Ms feature set. The Em feature set had given a competitive advantage to the Ms initial feature set, given the appearance of multiple languages in a single sentence. It is also obvious that the Em feature set improved the subjectivity classification performance on code-switching text instead of using the Malay initial feature alone for the classification.

C. *Unified Code-Switching POS Feature Set Performance Results*

The unified code-switching POS feature set fused Malay and English using the algorithm described in Algorithm 1. The unified process determined the POS of a word based on its language. The unified feature set was trained and tested using a similar experimental setting as the embedded code-switching feature sets and the initial feature sets. The accuracy performances from the classification of this feature set are captured and shown in Fig. 9.

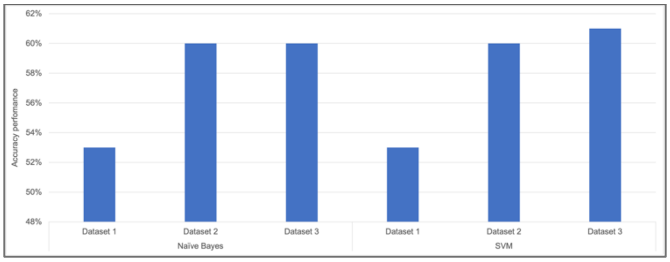


Fig. 9. Results of accuracy performance for the subjectivity of classification on Malay-English code-switching text using the unified feature set.

The bar chart in Fig. 9 shows the unified feature set performed at 52.00% accuracy using Dataset 1 and both classifiers. The accuracy performance of the unified feature set is on the same level at 60.00% for Datasets 2 and 3 using the Naive Bayes classifier and Data Set 2 using the SVM classifier. There is a slight increment in Dataset 3 using the SVM classifier, in which the feature set performed at 61.00% accuracy. The consistent performance of around 60.00 to 61.00% shows the feature set is usable to differentiate the subjective and objective sentences from a Malay-English code-switching dataset.

The bar chart in Fig. 10 compiled the accuracy results from the baseline experiments (En and Ms) and unified feature set (Un) using the Naïve-Bayes and SVM classifiers. In general, the unified feature set outperformed the English initial feature set using Dataset 2 with the Naive Bayes classifier at 60.00% and Dataset 3 with the SVM classifier at 61.00%. The unified feature set performed on par accuracy in comparison to the English initial feature set using Datasets 2 and 3 with the Naive-Bayes and the SVM classifier. The performances are 60.00%. The outperformance and on-par accuracy result shows the unified feature set is as good as the English initial feature set, given the fact that code-switching is a relatively new challenge to text processing and subjective analysis. However, the unified feature set did not share the same excellence of accuracy performance result using Dataset 1 for both classifiers. The unified feature set performed at 53.00% accuracy using Naive Bayes and SVM classifiers on the same dataset, less 2.00% and 1.00% for each classifier. Even though the same level of excellence is not shared by using Dataset 1 and the same classifiers, the unified feature set can still distinguish the subjective from the objective sentences of Malay-English code-switching text.

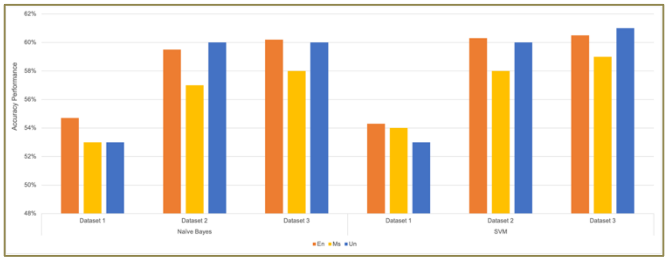


Fig. 10. Comparison of accuracy performance for the unified feature set with baseline initial feature sets.

The unified feature set outperformed the Malay initial feature set in four experiments. The accuracy performance of the unified feature set is 60.00% using Dataset 2 with Naive Bayes classifier, which is 3.00% higher than the Malay initial feature set. The same accuracy performance is seen as consistent for the feature set using Dataset 3 and the same classifier. However, it is 2.00% higher than the Malay initial feature set. The same pattern of differences is also apparent for Datasets 2 and 3 using the SVM classifier compared to the Malay initial feature set. The unified feature set performed at the same level of accuracy performance, 55.00%, using Dataset 1 with the Naive Bayes classifier. However, the feature set did not share the same accuracy performance as the SVM classifier using the same data set. The feature set performed at 53.00%, less than 1.00% from the Malay initial feature set. The significant differences in the accuracy performance between the unified Malay-English code-switching and the Malay initial feature sets across datasets and multiple classifiers show the unification of feature sets for two different languages is necessary to distinguish the code-switching sentences into objective and subjective classes.

D. *Stylistic Feature Sets Performance Results*

Stylistic features significantly influence subjectivity analysis [22][23]. Six experiments were conducted to investigate the ability of stylistic features to distinguish subjective and objective sentences from Malay-English code-switching datasets. The experiments were conducted using the same settings as the embedded and unified feature sets. The results are shown in Fig. 11.

The bar chart in Fig. 11 shows a consistent accuracy performance achieved by the stylistic feature set using Datasets 3 with the Naive Bayes classifier and Datasets 1, 2 and 3 with the SVM classifier. The stylistic feature set performed at 55% accuracy. It is worth mentioning that the differences in the stylistic performance from these datasets and Datasets 1 and 2 using the Naive Bayes classifier are 2.00% and 1.00%, respectively. The insignificant accuracy performance differences across datasets and classifiers show that the stylistic feature set performance is nearly consistent. Therefore, these findings support the influence of stylistic feature sets in code-switching text.

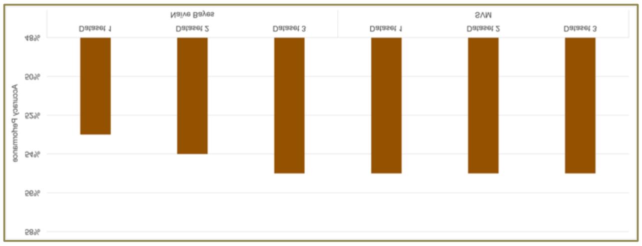


Fig. 11. Results of accuracy performance for the subjectivity of classification on Malay-English code-switching text using the unified feature set.

The accuracy performance of the stylistic feature set (St) is compared to the initial feature sets for English (En) and Malay (Ms). The accuracy performances are shown in Fig. 12. The bar chart in Fig. 11 shows the English and Malay initial feature sets have outperformed the stylistic feature sets using Datasets 2 and 3 for both classifiers. The stylistic feature set gives the same performance as the Malay initial feature set using Dataset 1 and the Naive Bayes classifier, but the English initial feature set still gives superior performance. However, the stylistic feature set can surpass the English and Malay initial feature sets using Dataset 1 and the SVM classifier.

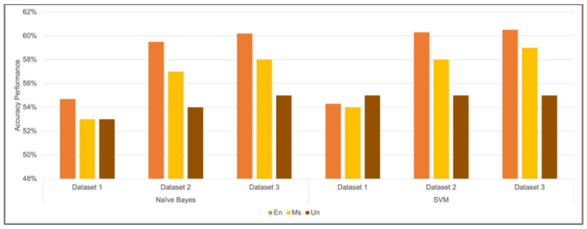


Fig. 12. Comparison of accuracy performance for a stylistic model with baseline initial feature sets.

E. *Combinations of Feature Sets Performance Results*

The experiments in the previous section were executed independently of each other. The proposed feature sets were then combined to determine the possibility of improved accuracy performances. Four more experiments were conducted for this purpose. The combinations are listed as follows:

1) The combination of an English (En) initial feature set and a stylistic (St) feature set is known as En + St.

2) The combination of the Malay (Ms) initial feature set and stylistic (St) feature set is known as Ms + St.

3) The combination of embedded (Em) feature set and stylistic (St) feature set is called Em + St.

4) The combination of a unified (Un) feature set and a stylistic (St) feature set is known as Un + St.

The experiments were carried out using a similar experimental setup as the previous experiments – using three datasets and two classifiers, the Naive Bayes and the SVM. The accuracy performances from each dataset were captured and averaged. The averaged accuracy performances are illustrated using bar charts in Fig. 13 and Fig. 14.

The average accuracy performances for the combination of proposed features using the Naïve Bayes classifier are shown using a bar chart in Fig. 13. The bar chart includes the averaged accuracy performance result from Malay and English feature sets. The result shows the combination of the unified (Un) and stylistic (St) feature sets outperformed other feature sets at 59.00% accuracy. The combination of embedded and stylistic feature sets (Em + St) shows an on-par performance with the English initial feature set. The results show that considering features from languages other than English in code-switching text and stylistic features improves the accuracy performance. Thus, the combination of feature sets is practical to classify the Malay-English code-switching sentences into subjective and objective classes. The bar chart also shows that the stylistic feature set alone performed at the lowest accuracy, 54.00%, in comparison to other feature sets. The performance reveals using only stylistic features to classify the subjectivity of code-switching sentences is insufficient.

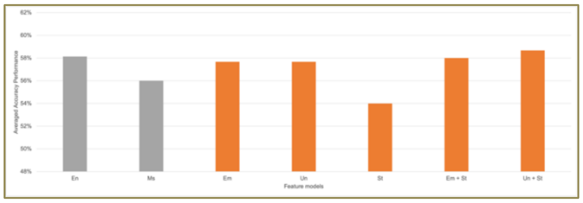


Fig. 13. Comparison of averaged accuracy performances for combined feature sets using the Naïve-Bayes classifier.

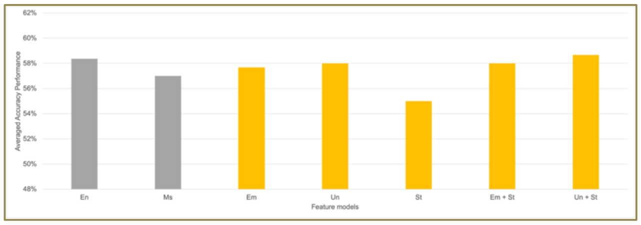


Fig. 14. Comparison of averaged accuracy performances for combined feature sets using the SVM classifier.

The results of averaged initial feature sets (Malay and English) proposed feature sets (embedded, unified and stylistic) combination feature sets that were classified using the SVM classifier are shown in a bar chart of Fig. 14. The bar chart shows the averaged accuracy performance using the combination of unified and stylistic (Un + St) feature set surpassed other feature sets. The Un + St feature set performed at 59.00% accuracy. The combination of embedded and stylistic (Em + St) feature sets performed at 58.00% accuracy, slightly lower than the English (En) initial feature set and on par as a unified feature set. The bar chart shows a slight accuracy performance improvement when the embedded and unified feature sets are combined with the stylistic feature set. The performance of both feature sets increased by 1.00% with the combination of stylistic feature sets. The bar chart also shows the stylistic feature set alone gave the lowest performance, that is, 55.00% accuracy. Therefore, a stylistic feature set should not be used as the only feature set to distinguish subjective and objective sentences for Malay-English code-switching datasets. In addition to that, stylistic feature sets have proven their influence on subjectivity classification for Malay-English code-switching text.

VI. Discussion

The results of this study provide a comprehensive evaluation of the subjectivity classification in Malay-English code-switching text using enhanced feature sets and machine learning classifiers. Several key insights emerged from the result analysis.

A. Performance of Baseline Feature Sets

The baseline feature sets focused exclusively on English and Malay feature sets. It provides a starting point to evaluate the effectiveness of the enhanced feature sets. The English feature set has outperformed the Malay feature sets across all datasets. This finding aligns with previous research that suggests English-based models tend to perform better in text classification tasks due to the availability of more extensive language resources and tools.

B. Embedded Code-Switching Feature Set

A key contribution of this study is how the embedded code-switching feature sets successfully captured subjectivity in code-switched text. These feature sets performed better than the baseline Malay sets, especially when using Dataset 3 with both Naive Bayes and SVM classifiers. This shows how important it is to include both languages in the feature sets to handle code-switching effectively. Additionally, the improvement in accuracy from Dataset 1 to Dataset 3 suggests that having a more diverse dataset could help models better recognise subjectivity in code-switched sentences.

C. Unified Code-Switching POS Feature Set

The unified feature set, which combined Malay and English parts of speech (POS), delivered competitive results, particularly with Dataset 3 when using the SVM classifier. This approach outperformed the baseline English feature set, highlighting the importance of considering the syntactic structures of both languages in a unified way for tasks involving code-switching. Interestingly, the unified feature set exhibited limitations in Dataset 1, indicating that while this method is advantageous, its performance depends on the characteristics of the dataset and the quality of POS tagging.

D. Impact of Stylistic Feature Set

Stylistic features, which include linguistic cues such as punctuation, sentence structure, and word choice, were also explored for their role in subjectivity classification. While the stylistic feature sets performed consistently across all datasets, they were outperformed by the initial English and Malay feature sets. This suggests that stylistic features alone may not be sufficient for classifying subjectivity in code-switched text. However, when combined with other feature sets, such as the unified and embedded feature sets, stylistic features contributed to incremental improvements in accuracy.

E. Combination of Feature Sets

The experiments that combined multiple feature sets revealed that unifying linguistic and stylistic features yield the best performance. Specifically, the combination of the unified and stylistic feature sets resulted in the highest accuracy across datasets, particularly with Dataset 3. This finding suggests that a hybrid approach—leveraging multiple linguistic levels (e.g., syntax, style, language features)—is most effective for handling the complexity of Malay-English code-switching text. These findings reinforce the idea that no single feature set can comprehensively capture the nuances of subjectivity in code-switching. Instead, a combination of feature sets is required to achieve optimal performance.

VII. Limitation and Future Research

There are several limitations to this study. First, the datasets used in the experiments, while diverse, may not fully capture the complexity of real-world code-switching scenarios. Future research could explore larger and more diverse datasets that include different levels of code-switching. Additionally, this study focused on sentence-level classification; however, subjectivity may vary within a single sentence. Thus, future work could explore word-level or phrase-level classification to gain a more fine-grained understanding of subjectivity in code-switching.

Moreover, the study was limited to Malay-English code-switching. As such, future research could apply the proposed methods to other language pairs where code-switching is common, such as Spanish-English or Arabic-French. Investigating whether the same feature sets and classifiers work across different language combinations could provide more generalizable insights into the phenomenon of code-switching in subjectivity classification.

VIII. Conclusion

The results have several implications for future research and practical applications. First, the importance of language-specific features in code-switched text is evident, highlighting the need for models that can process multiple languages simultaneously. This is particularly relevant in multilingual societies like Malaysia, where code-switching is prevalent in everyday communication. The findings also suggest that existing models and datasets can be improved by focusing on the interaction between syntactic and stylistic features in both languages.

Three feature sets were designed to distinguish the code-switching text into subjective and objective classes. The first feature set includes the Malay feature into the English feature – known as the embedded feature set. The second feature set combined and unified the Malay and English feature set, which is known as a unified feature set. The last feature set is non-vocabulary, known as a stylistic feature set. Part-of-speech is used as a feature of the code-switching text. The 1 and 0 are used to represent the presence and absence of the features. The feature sets are more interpretable where the contribution of a specific can be directly seen to the model decisions.

The experiments were carried out using three datasets and two machine learning classifiers – the Naive Bayes and SVM to verify the proposed feature sets. The results from the proposed feature sets were compared with the initial feature sets. The experiment shows the unified feature performed as good as the English initial feature set. Therefore, unifying the Malay and English feature sets is necessary to distinguish the subjective and objective sentences in the Malay-English code-switching dataset. An improvement in accuracy performance is achieved when the unified feature set is combined with the stylistic feature set. The improvement reveals this method is computationally lighter which requires fewer data to perform well.

Acknowledgment

The authors would like to thank Social Analytics and Intelligence Lab (SAIL) and the Applied Intelligent Computing (APIC) research group, the Center of Advanced Computing Technology (C-ACT), and Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for their incredible support for this research.

References

[1] P. Muysken, Bilingual Speech: A Typology of Code-Mixing. Cambridge University Press, 2000.

[2] T. Huang, Z. H. Deng, G. Shen, and X. Chen, 'A Window-Based Self-Attention approach for sentence encoding', Neurocomputing, vol. 375, pp. 25–31, Jan. 2020, doi: 10.1016/j.neucom.2019.09.024.

[3] L. B. Belisário, L. G. Ferreira, and T. A. S. Pardo, 'Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese', Information (Switzerland), vol. 11, no. 9, Sep. 2020, doi: 10.3390/INFO11090437.

[4] A. Al Hamoud, A. Hoenig, and K. Roy, 'Sentence Subjectivity Analysis of Political Debate and Ideological Debate Dataset using LSTM and BiLSTM with Attention and GRU Models', Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 7974–7987, Nov. 2022, doi: 10.1016/j.jksuci.2022.07.014.

[5] A. K. Joshi, 'Processing of Sentences with Intra-Sentential Code-Switching', in Proceedings of the 9th Conference on Computational Linguistic, 1982, pp. 145–150. doi: 10.1017/cbo9780511597855.006.

[6] J. Chi, B. Lu, J. Eisner, P. Bell, P. Jyothi, and A. M. Ali, 'Unsupervised Code-switched Text Generation from Parallel Text', in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech Communication Association, 2023, pp. 1419–1423. doi: 10.21437/Interspeech.2023-1050.

[7] I. Hamed, N. Habash, and N. T. Vu, 'Data Augmentation Techniques for Machine Translation of Code-Switched Texts: A Comparative Study'. [Online]. Available: http://arzen.camel-lab.com/

[8] K. Hu et al., 'Improving Multilingual and Code-Switching ASR Using Large Language Model Generated Text', in 2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASRU57964.2023.10389644.

[9] D. Gupta, A. Lamba, A. Ekbal, and P. Bhattacharyya, 'Opinion Mining in a Code-Mixed Environment: A Case Study with Government Portals', Proc. of the 13th Intl. Conference on Natural Language Processing, pp. 249–258, 2016, [Online]. Available: http://ltrc.iiit.ac.in/icon2016/proceedings/icon2016/pdf/W16-6331.pdf

[10] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, 'Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora', no. 2011, pp. 2–8, 2015, doi: 10.18653/v1/w15-2902.

[11] A. Jamatia, S. D. Swamy, B. Gambäck, A. Das, and S. Debbarma, 'Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus', International Journal on Artificial Intelligence Tools, vol. 29, no. 05, pp. 20–35, Jun. 2020, doi: 10.1142/S0218213020500141.

[12] V. Hatzivassiloglou and J. M. Wiebe, 'Effects of Adjective Orientation and Gradability on Sentence Subjectivity', in COLING '00: Proceedings of the 18th conference on Computational linguistics, 2000, pp. 299–305. doi: 10.3115/990820.990864.

[13] S. Ibrahim, Kamus Dwibahasa Bahasa Inggeris-Bahasa Melayu, Edisi Kedu. Dewan Bahasa dan Pustaka, 2019.

[14] J. M. Hawkins, Kamus Dwibahasa Oxford Fajar Inggeris-Melayu, Melayu-Inggeris. Fajar Bakti, 2006.

[15] T. Solorio and Y. Liu, 'Learning to Predict Code-Switching Points', in EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL, 2008, pp. 973–981. doi: 10.3115/1613715.1613841.

[16] M. A. Ullah, S. M. Marium, S. A. Begum, and N. S. Dipa, 'An algorithm and method for sentiment analysis using the text and emoticon', ICT Express, vol. 6, no. 4, pp. 357–360, 2020, doi: https://doi.org/10.1016/j.icte.2020.07.003.

[17] B. Jung, H. Kim, and S. H. (Shawn) Lee, 'The impact of belongingness and graphic-based emoticon usage motives on emoticon purchase intentions for MIM: an analysis of Korean KakaoTalk users', Online Information Review, vol. 46, no. 2, pp. 391–411, Jan. 2022, doi: 10.1108/OIR-02-2020-0036.

[18] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, 'Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations', IEEE/ACM Trans Audio Speech Lang Process, vol. 29, pp. 1675–1686, 2021, doi: 10.1109/TASLP.2021.3076364.

[19] L. B. Belisário, L. G. Ferreira, and T. A. S. Pardo, 'Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese', Information, vol. 11, no. 9, pp. 1–14, 2020, doi: 10.3390/INFO11090437.

[20] S. M. Mohammad, 'Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text', Emotion Measurement, pp. 201–237, 2016, doi: 10.1016/B978-0-08-100508-8.00009-6.

[21] B. Krawczyk, 'Learning from imbalanced data: open challenges and future directions', 2016. doi: 10.1007/s13748-016-0094-0.

[22] F. Bravo-Marquez, M. Mendoza, and B. Poblete, 'Meta-Level Sentiment Models for Big Social Data Analysis', Knowl Based Syst, vol. 69, no. 1, pp. 86–99, 2014, doi: 10.1016/j.knosys.2014.05.016.

[23] A. S. Altheneyan and M. E. B. Menai, 'Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts', Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pp. 473–484, 2014, doi: 10.1016/j.jksuci.2014.06.006.

A Lightweight Privacy Preservation Protocol for IOT

A Data and Metadata Protection Protocol

Ahmed Mahmoud Al-Badawy[1], Mohammed Belal[2], Hala Abbas[3]
Teaching Assistant, Computer Science Department-Faculty of Computers and Artificial Intelligence,
Helwan University, Cairo, Egypt[1]
Prof., Computer Science Department-Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt[2]
Assist. Prof., Computer Science Department-Faculty of Computers and Artificial Intelligence, Helwan University[3]
Faculty of Computer Studies, Arab Open University, Cairo, Egypt[3]

***Abstract*—Due to rapid evolution of Internet of things (IOT) in terms of hardware, software and communication leads to widespread expansion across many domains and sectors. This expansion consequently results in sensitive data transfer increase for purposes of complex calculations and decision making which in turn leads to increase of data attacks and leakage which results in data privacy violation. Although, a lot of current solutions tried to fulfill data privacy via lightweight mechanisms but neither provided end to end protection nor gave a focus to metadata protection which can reveal valuable information about data it describes. This paper presents a lightweight complete data privacy protocol which manages the lifecycle of data starting from object registration till data transfer to cloud. The proposed protocol is a trusted third party free (TTP-Free) which adopts anonymization techniques, lightweight key agreement protocol, end to end encryption and message authentication code to fulfill identity and data protection which in turn fulfill complete data privacy.**

Keywords*—*IOT; data privacy; lightweight protocols; end to end protection; data and metadata protection

I. Introduction

Internet of Things (IOT) is a physical network of resource constrained objects (ex. sensors, actuators, wearables, IIOT devices) connected together in order to rapidly exchange data to fulfill a specific job. IOT has three main visions [1] to focus on:

- Things vision which tends to focus on generic objects and integration of them into a framework (ex. RFID, NFC).
- Internet vision which tends to present IOT as a network-oriented (ex. IPO, communicating things).
- Semantic vision [2] which tends to view IOT as a worldwide network of interconnected things that can be uniquely identified (ex. reasoning over data, semantic technologies).

IOT in general aims to facilitate people's life and enhance countries' economies via introducing smart solutions capable of serving required needs which in turn leads to a better world. It can connect people, things, objects, and devices regardless of time and location with barrier free manner. With, the advances in IOT hardware and software started from enhancing communications networks, devices and reducing things sizes and cost reduction for constructing IOT networks led to invasion and domination of IOT in many domains due to benefits got from.

Healthcare [3][4] is one of domains that IOT tried to support to enhance people health and saves their life. Medical IOT aims to serve patients via presenting a lot of services started from monitoring services [5][6] where patients' health is tracked to avoid any health disaster till complex healthcare solutions for malignant diseases like cancer [7][8]. Moreover, IOT used to fight against pandemic diseases like Covid 19 [9][10].

Agriculture is another domain that IOT gave attention to enrich and support due to its economic importance to countries. IOT developed a special type of network [11] called agriculture IOT sensor monitoring networks (ASMN) in order to fully monitor farmland in (temperature, humidity, light and soil moisture) and take appropriate actions needed. These networks [12] aim to continuously monitor crops to protect crops' health.

Industry is the third domain that IOT tried to automate and support with industrial IOT (IIOT). IIOT [13][14] is a specialized network which manufacturers adopt to enhance production process started from supplying raw materials till customer services.

The advancement of IOT networks and increasing NO. of objects used led to increased exchange of sensitive data which in turn lead to a lot of security and privacy problems. Data privacy is one of the most important problems to be focused on due to sensitivity of data. Data here can be personal, healthcare, industrial or even militaria which needs to be protected while being transferred.

A lot of attacks aim to leak data to be abused, attacks can be categorized into two types:

- Active attacks which attacker tries to change the whole or part of data while being transferred.
- Passive attacks which attacker tries to read data only without any change.

In this paper, a lightweight trusted third-party free (TTP-Free) data privacy protocol is presented to build a secure communication channel for Device to cloud (D2C) which aims to protect IOT data and metadata as well. Although metadata might seem less important, it can reveal valuable information about the data it describes which in turn leads to privacy compromise. The protocol depends mainly on four parts to fulfill privacy:

- Anonymization: regularly changing objects' identities in order to avoid tracking and impersonation attacks based on objects known IDs.
- Lightweight Key agreement Protocols: used between objects and cloud to construct session key without directly exchanging it which will be used later for data encryption/decryption.
- Lightweight End to End Encryption: designed especially for constrained devices to encrypt/decrypt data using already constructed key between object and cloud to make sure that no party can decrypt data transferred except cloud.
- Message Authentication Code: used to authenticate message via edge to ensure integrity and authenticity of data which in turn resist active attacks.

The paper is structured as follows: Section II gives an overview of the related work and the main differences between proposed work and existing research. Section III provides an overview of proposed protocol, all related algorithms and techniques used. Section IV discusses security analysis of proposed protocol via threat model and its analysis. Section V evaluates the proposed protocol against existing one to clarify strength and weakness of each one at predefined criterion. Section VI provides conclusion and future work.

II. Related Work

A lot of researches tried to provide solutions and mechanisms to IOT data privacy leakage due to its necessity. Many attacks include impersonation, injection, eavesdropping, data theft and reprogram attacks aims to track IOT networks for data leakage and abuse.

J. Andrew [15] proposed an anonymization clustering schema which aims to fulfill data privacy in medical IOT. This schema depends on two parts, client side which is responsible for anonymizing data generated by things using clustering K anonymity which fulfill privacy via clustering methodology, server side which uses cluster combination to reduce communication overhead which achieve privacy. This schema employs a trusted intermediate aggregation node to anonymize data got from client then send it to untrusted server to be sent to data collector. Usage of anonymization techniques with trusted third-party only fulfills data privacy partially due to a lot of attacks that can lead to original data restoring (ex. Re-identification attacks) besides relying on TTP -aggregation node- which can be attacked.

Xuezhen [16] proposed a framework that aims to fulfill security and data privacy via cryptography and behavior pattern analysis. The framework is divided into levels according to IOT main entities:

- Objects: which is defined as sensors and actuators each of them has a security and privacy requirements.
- Communication networks: is responsible for communication between objects that needs to be protected to protect network from abuse.
- Users: who use the IOT, which is the most sensitive part as part of people in this context will be attackers themselves so users' behaviors must be carefully analyzed and stored to detect any malicious behavior.

In order to provide security and privacy to users and data. The framework used secured channel to fulfill required security but did not mention how to accomplish this. Moreover, the framework deals direct with object real identities which makes the system vulnerable to impersonation and tracking attacks.

Uzair Javaid [17] focused on data provenance and integrity by using BlockpPo framework which is a combination between PUFs which produces a unique response so data provenance is established with each IOT device, and blockchain which enforces data integrity to fulfill data privacy. Although, blockchain tried to fulfill data privacy across IOT networks, but still has a lot of challenges [18] which may affect that fulfillment starting from choosing blockchain platform (public, private) which will affect confidentiality and integrity of data. Moreover, the identity will be disclosed due to sharing transactions with their owners.

Othman [19] proposed a privacy preserving schema using homomorphic encryption in order to protect healthcare data privacy. Its main goal is to provide safe and secure aggregation for data with respect to energy consumption. The schema tried to protect data from active, passive, internal and external attacks. Although the proposed schema depends mainly on cryptography using homomorphic encryption which enables coordinator to work on without needs to decryption, it does take into consideration objects' identity protection which in turn can lead to tracking and impersonation attacks. Moreover, the schema does not state the data encryption decryption key mechanism used which is considered a very critical part to be covered due to diversity of attacks occurred on that part.

Prem Prakash [20] proposed a technique for data privacy preserving via introducing privacy preserving IOT architecture based on OpenIOT [21]. This architecture provides end to end privacy by giving ability to control access to sensitive IOT data via distributing and decomposing data into multiple data stores and then aggregated again when needed [22]. The technique composed of four communication parts (IOT device and gateway, gateway and data store, data store and data access finally, data access and user) which assumes that these communication channels are secured by applying cryptography and key sharing mechanisms only. Although this approach tried to fulfill data privacy by focusing on how to hide data that is transferred from between communication parties, this approach is not adequate to fulfill objective needed. Focusing on data only without paying attention to object identities can lead to data leakage which in turn leads to data privacy issues.

Mamun Abu-Tair [21] proposed a new architecture that aims to support IOT applications with a specified level of security and privacy. The architecture is bundled with new algorithm that is responsible for configuration of newly added sensors in terms of cryptographic suite to match target applications. The architecture employs cryptography and anonymization to fulfill complete privacy but relying on trust management schema is considered a weak point. Moreover, key management schema is not stated which can lead to critical attacks.

Shancang Li [23] presents a lightweight privacy preserving protocol which aims to address privacy issues between objects, cloud and users using cryptography – homomorphic encryption-. The protocol depends on a key management schema which employs users' keys beside objects' keys to make sure that the data will be delivered to correct user. Although the protocol tried to fulfill data privacy but it has a major concern to be addressed, the protocol did not state in details key sharing mechanism which can be a weak point to the whole protocol moreover, the protocol deals with objects with their real identity which makes the system vulnerable to impersonation and tracking attacks.

Mohammed Ahmed [24] used remote patient monitoring as a case study in healthcare domain to fulfill security and data privacy via proposing a new system that provides mutual authentication and employed cryptography to protect data while being transferred. Although the proposed system tried to fulfill privacy via applying cryptography, the system does not pay attention to object identities which can be tracked and impersonated. Moreover, the registration phase for objects is not powerful to forbid injection attacks.

Xi lou [25] presented a lightweight security protocol which aims to fulfill data privacy via cryptography and symmetric key mechanism. This protocol tries to maximize symmetric keys generated via key delegation which uses chaotic system and logistic map to ensure unpredictability and unrepeatability of keys generated. The protocol depends mainly on control center as a trusted third party to be responsible for key management between communication parties which is considered a weak point if got controlled by attackers.

A lot of protocols and systems tried to fulfill data privacy by focusing on either protecting objects' identities or data which is considered a partially fulfillment. Some of them is trusted third party and others is trusted third party free. Up to our knowledge, all protocols focus mainly on objects' data as a protection level, but no one pay attention to metadata – like gateway id, manager id -. This gap is critical to be protected since leakage will lead to disclosure of much sensitive information (ex. Objects cluster, network location and sometimes object itself) will lead to data privacy violation. The proposed protocol is a trusted third party free which aims to fulfill full data privacy by protecting objects' identities and their data. Moreover, the proposed protocol has put into consideration the protection of metadata to avoid any privacy violation.

III. Proposed Protocol

The proposed protocol is considered a communication protocol with a set of defined rules that regulate exchange of data between parties in a secure manner to fulfill data privacy. The protocol focuses mainly on both:

- Data mainly reads from objects and needs to be transferred to the cloud.
- Metadata, which is data about data like timestamp, gateway id, edge id which needs to be protected as well.

The proposed protocol is trusted third party free that focuses on fulfilling data privacy via securing communication channels between objects and cloud by using lightweight key agreement protocol and end to end encryption to ensure that only cloudcan decrypt the data issued by objects. In addition, the protocol employs anonymization techniques for objects in order to prevent tracking and impersonation attacks for objects. Therefore, being trusted third party free and providing secure communication channel beside objects' identities anonymization will provide full data privacy.

Notations used are summarized in Table I.

TABLE I. Notations Summary

Notation	Description
O_i	Object_i
GT	Gateway
Mgr	Manager
IDS	Identity Server
Edg	Edge
Cld	Cloud
Fid_i	Fake ID_i
Id_i	ID_i
TS	Timestamp
Y-> Send(X,{Z})	Y Sends parameters Z to X
Y->Construct(X,Z)	Construct Part X With Z and store it in Y.
H(X)	Hashing X
Key	Session key between Object and cloud
O_i ->Enc(P, Key)	Encrypt plain text (P) with Key for O_i
O_i ->Dec(P, Key)	Decrypt cipher text (P) with Key for O_i
Read	Senor Captured Read
CldPubK	Cloud public Key
CldPrK	Cloud private Key
ChkElg	Check Eligibility
CAT(X,Y)	Concatenate X and Y

The proposed protocol consists of six main components as below:

- Object: is denoted by O_i which is responsible for gathering required data and do necessary functionality to it.
- Identity Server generators: is denoted by IDS, responsible for satisfying objects' requests to form fake identities.
- Manager: is denoted by Mgr which is responsible for managing objects lifecycle starting from object registration till data exchange. Each network cluster has its Mgr to do required jobs.

- Edge: is responsible for preparing object requests and adding necessary meta data.
- Gateway: is responsible for verifying eligibility for objects to send data or not and doing necessary functions to send data to cloud.
- Cloud: is responsible for mapping data to correct object identities, storing data and perform required analysis to take needed decisions.

Fig. 1 shows the key components of the proposed protocol.

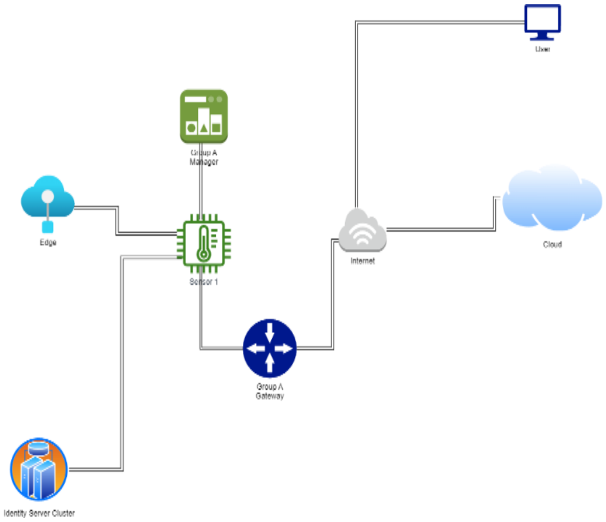


Fig. 1. Proposed Protocol.

The protocol consists of four phases according to below:

1) Registration Phase: Each object (Oi) to be added to IOT network must firstly send to it's Mgr to be registered and approved. Once got approval, Oi starts to request its fake identity to start communicating with.

2) Anonymization Phase: It is responsible for changing real object identity to fake one to resist any tracking attacks or impersonation for any object based on its real identity, as below:

- O_i sends a request with timestamp (T_1) for more than one IDS (n) where n >1 to form its fake identity if it is expired.
- IDS validate timestamp against timestamp threshold (T) via $|T_{IDS} - T_1| < \Delta T$ to determine if the request will be satisfied or rejected.
- Each server receive request generates part of identity uniquely and send it back to O_i with timestamp and expiry date.
- The severs send parts generated to cloud with other information required (server id, object id, timestamp).
- Cld concatenate parts received from servers based on received timestamps ascending, validate timestamp against timestamp threshold (T) via $|T_{cld} – T_i| < \Delta T$, hash the concatenation output to get fake identity, set expiry date for that fake identity based on system configuration and then store mapping for fake identity to real one in mapping tables.
- O_i prepares its fake identity as cloud did, validate timestamp and send it to cluster mgr to update its list.
- Cluster mgr updates its list and broadcast it to GT. Table II shows construction of fake identities.

TABLE II. FAKE IDENTITY GENERATION ALGORITHM

Algorithm 1: Fake Identity Generation by object Oi

```
$O_i$ -> Send( $IDS_1$, { $Id_i$ ,$T_1$ }).
$IDS_1$ -> Validate Timestamp if( $|T_{IDS1} - T_1| > \Delta T$ then rejected
$IDS_1$ -> Send($O_i$ , { $Fid_1 \|$ $T_{s1}$ })
$IDS_1$-> Send( Cld, {$Fid_1 \|$ $IDS_1 \|$ $id_i \|$ $T_{s1}$}) .
$O_i$ -> Send( $IDS_2$, { $Id_i$ ,$T_2$ }).
$IDS_2$ -> Validate Timestamp if ( $|T_{IDS2} – T_2| > \Delta T$) then rejected
$IDS_2$ -> Send($O_i$ , { $Fid_2 \|$ $T_{s2}$ })
$IDS_2$ -> Send( Cld, {$Fid_2 \|$ $IDS_2 \|$ $Id_i \|$ $T_{s2}$}) .
$O_i$ -> Construct(CAT( $Fid_1$, $Fid_2$ , H( $id_1$, $T_{s1}$, $T_{s2}$)), $T_{f1}$) and output Fid to be stored
Cld -> Construct(CAT( $Fid_1$, $Fid_2$ , H( $id_1$, $T_{s1}$, $T_{s2}$), $T_{f1}$))) and output Fid to be stored.
$O_i$ -> Send( $Mgr_i$ , { $id_1$ $\|$ $Fid_1$}).
$Mgr_i$ -> Update (List, {$Fid_1$}).
$Mgr_i$ -> Send (GT, List).
```

3) Session Key Generation Phase: It is the phase responsible for constructing session key between object and cloud via lightweight key agreement protocol [26]. This protocol consists of two stages as below:

Registration stage: The object register itself on the cloud through the following:

- O_i chooses identity ID_i and password PW_i then two parameters a and b.
- O_i calculates $Mpw_i = h(PW_i \| a \| b \| ID_i)$, $HID_i = h(ID_i \| b)$ and $d_i = a \oplus b$.
- O_i sends { HID_i , Mpw_i , d_i , a} to Cld.
- Cld then calculates $v_i = h(HID_i \| Mpw_i)$ then chooses random numbers c_i z_i.
- Cld calculates both $B_i = h(HID_i \| x_s)$, $E_i = (B_i \oplus Mpw_i)$.
- Cld will store (z_i, HID_i) in its database.
- Cld will calculate $A_i = E_{xs} (c_i \| HID_i \| d \| a)$
- Cld will send { A_i , E_i , z_i , v_i , c_i , b, a}.
- O_i will calculate $T_i = A_i \oplus Mpw_i$
- O_i will store [T_i , E_i , z_i , v_i , c_i , b, a]

- Login and authentication stage: The object successfully logged in to authenticate itself and start information exchange as below:
  - O_i submit its ID_i and PW_i

- O_i calculates $Mpw_i = h(PW_i \| a \| b \| ID_i)$, $HID_i = h(ID_i \| b)$, $v_i = h(HID_i \| Mpw_i)$, $B_i = E_i \oplus Mpw_i$.
- After calculating new v_i , it compares it with old one.
- O_i calculates $d_i = a \oplus b$ and $cd_i = c_i \oplus d_i$.
- O_i chooses random number e_i and selects T_1 .
- O_i then calculates $A_i = T_i \oplus Mpw_i$, $M_i = E_{Bi} (T_1 \| e_i` \| A_i)$.
- O_i sends { cd_i , M_i , T_1 , HID_i } to Cld.
- Cld will select T_2 then check whether $| T_2 - T_1 | < \Delta T$.
- Cld will calculate $B_i = H(HID_i \| X_s \| z_i)$.
- Cld will $DEC(M_i)_{Bi} = (T_1, A_i, e_i)$ and $DEC(A_i)_{Xs} = (HID_i , a, c_i, d_i)$.
- Cld will then calculate $cd_i = c_i \oplus d_i$ and then check if new cd_i equals old one or not and the same for T_1 then chooses random number q_i.
- Cld finally will calculate $Q_i = H(A_i \| B_i)$, $s_i = q_i \oplus B_i$, $w_i = h(cd_i \| e_i)$, $N_i = E_{Ai} (s_i \| T_2 \| w_i)$.
- Cld will send N_i , T_2 back to object.
- O_i will select T_3 and check whether $| T_3 – T_2 | < \Delta T$.
- O_i will $DEC(N_i)_{Ai} =(s_i, T_2 , w_i)$.
- O_i will calculate w_i `= $H(cd_i \| e_i)$ and then checks whether $w_i$ ` = w_i and $T_2 = T_2$.
- O_i will calculate $q_i = s_i \oplus B_i$, $Q_i = H(A_i \| B_i)$, $sk= H(e_i \| B_i \| Q_i \| q_i \| z_i \| s_i)$, $M N_i = H(Sk \| q_i \| s_i \| Q_i)$
- O_i will send $M N_i$ and T_3 to cloud to finalize key construction.
- Cld will select T_4 and then checks $| T_3 – T_4 | < \Delta T$ and then calculates $sk= H(e_i \| B_i \| Q_i \| q_i \| z_i \| s_i)$ and $M N_i$ `$= H(Sk \| q_i \| s_i \| Q_i)$ and then checks whether new and old $M N_i$ are equal.

4) Transferring Data to Cloud Phase: It is the phase responsible for transferring data from objects to cloud to be processed and stored as below:

- O_i encrypts current read with lightweight Speck-R algorithm [27] using constructed session key on session key generation phase.
- O_i will send encrypted read with newly generated fake identity to Edg.
- Edg will prepare the request by adding needed meta data (timestamp, edge id, gateway id) to the request
- Edg will encrypt the whole data and metadata with cloud public key and send it to gateway in conjunction with message authentication code [28] to ensure data integrity and resist active attacks.
- GT will verify whether object has right to send data to cloud or not, if yes, the GT will forward data to cloud.
- Cld receives request then decrypts it using its private key and then go through verification in terms of timestamp and message authentication code [28] attached to ensure both data integrity and no reply attack took place.
- Cld will decrypt the data using previously constructed session key with O_i.
- Cld will get mapping for fake identity and check expiry date to validate if it is still used or not to avoid impersonation and identity theft attacks.
- Cld will store data with real identity.

IV. Security Analysis

This section provides a complete security analysis for proposed protocol by formulating threat model and threat model analysis with informal and formal analysis to prove correctness of designed protocol in terms of security and attacks resistance.

A. Threat Model

In IOT environments, integrity and confidentiality are considered critical part to be dealt with as stated in Dolev-Yao adversary [29].

According to proposed protocol, objects, gateways, managers, edges and identity servers are communicated to each other using internal IOT network. Objects before sending any data, must firstly acquire their fake identities to replace their real ones - in order to be protected from impersonation and identity tracking attacks - while sending data. Moreover, all communications to cloud must go through gateway which has ability to forward request or drop it due to any violation. Manager is responsible for managing authorization of objects in terms of sending data even though they got new fake identities. Edge is responsible for adding necessary metadata and providing a second layer of security to data by encrypting whole data by cloud public key to be sent to cloud which in turn fulfill data privacy. All parties (Manager, Gateway, Edge, identity) are assumed to be dishonest which means they are curious about data.

Attackers aim to reveal as much data as possible by trying to gain access to any IOT network party or sniffing communication network itself. Data is not only objects generated reads but its metadata as well due to its importance. Metadata can be used to extract valuable information about data itself (ex. object cluster, Location and object itself in many cases) to attackers which in turn causes data privacy leakage.

Our objective is to minimize the amount of information attackers can gain to protect data privacy through IOT networks via fulfilling confidentiality and integrity of data.

A lot of Assumptions to be considered:

- Attacker has ability to intercept any message between cloud and IOT network.

- Lightweight key agreement protocol [26] for key construction between objects and cloud and Speck-R [27] algorithm are secured.
- Cloud and object themselves are secured.

B. Threat Model Analysis

Recall that from our threat model, our objective is to minimize the amount of information attackers can gain to protect data privacy through IOT networks.

Attacks can be classified in to two main parts:

- Internal attacks which are carried out inside IOT network.
- External attacks which are carried out outside IOT network.

And each part of them can be classified into:

- Active attacks which are considered unauthorized access aim to alter networks data or injecting data other than correct one.
- Passive attacks are considered unauthorized access to gain data without modifying it.

And according to our threat model assumptions, attacks on objects and cloud themselves are out of scope.

C. Informal Analysis:

The proposed protocol aims to fulfill the following:

- Confidentiality: The protocol aims to fulfill confidentiality via end-to-end encryption between object and cloud. Speck-R is used as a first layer encryption to protect data from unauthorized access and eavesdrop in conjunction with light key agreement protocol. Key agreement protocol is used to form encryption key without directly exchanging it which provides more security and prohibit key sniffing attacks. Therefore, any data to be transferred from objects will be in ciphertext form which in turn fulfills Confidentiality.
- Integrity: The protocol aims to fulfill integrity and authenticity of data via message authentication code which in turn resists active attacks. The edge before encrypting metadata will authenticate message by adding message tag to ensure that message is not tampered or altered through communication.
- Anonymization: The protocol aims to protect identity of objects by changing real identities of objects with other ones while sending reads to avoid tracking attacks and impersonation attacks which in turn preserve data privacy.

And the protocol has ability to resist the following:

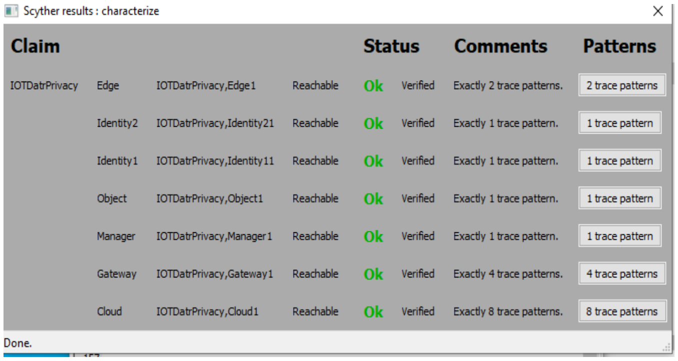
- Man in the middle attack (MITM): The attacker position himself between two communication parties in order to intercept exchanged messages and modify the content. Even though, the attacker can store the message for a while and resend it later. The proposed protocol has ability to deal with this attack via fulfilling both Confidentiality and integrity by applying both E2E encryption and using message authentication code which in turn helps on defending against MITM attack.
- Eavesdropping and Interference: The attacker aims to eavesdrop on any part of the network to extract any valuable information. The proposed protocol has the ability to resist that by applying E2E encryption between object and cloud and adding a second layer of encryption between edge and cloud for metadata which in turn resists any eavesdropping attacks.
- False Data Injection attack: The attacker tries to inject false data instead of correct one which in turn leads to wrong decision on cloud. The proposed protocol has ability to resist that by firstly apply Anonymization for objects to anonymize objects’ identities and register these identities on cloud. Secondly, two level encryption one with the session key using key agreement protocol between cloud and object, other with edge and cloud. Even though the attacker tried to inject a node into the network with the purpose of injecting false data, the gateway will not forward this data to cloud due to being unauthorized from network manager which in turn makes it difficult for any attacker to inject any false data to cloud.
- Advanced Persistent Threat: The attacker tries with many tactics and techniques to infiltrate IOT network and be silent and undetected for a long time with aim to steal and leak valuable information. The protocol has the ability to resist that via achieving Confidentiality and integrity.
- Reply attacks: The attacker tries to intercept network and retransmit a message between communication parties which in turn will lead to wrong timed message received. This wrong message can lead to disasters and wrong decisions due to being received correct at the wrong time. The proposed protocol has ability to resist this type of attack, by taking into consideration timestamp which message came with, if the difference between receivers’ timestamp and message timestamp greater than threshold defined on protocol, the request will be rejected.

D. Formal Analysis

The proposed protocol had been verified by scyther tool [30], used to analyze and verify security protocols in terms of vulnerabilities and flaws in their design. According to our protocol implementation on scyther, seven roles are implemented to cover all protocol aspects.

The analysis will be for each protocol phase to make sure that the data and metadata are still protected while being transferred which in turn fulfills protocol goals in terms of data privacy. Fig. 2 states proposed protocol analysis results.

For registration and anonymization phase, the object sends to network manager to be added and approved. The manager checks eligibility for that object request to be approved or rejected. If the request is approved, the object starts to request its fake identity via identity servers, minimum two servers to be requested as stated in Algorithm 1.



Claim				Status		Comments	Patterns
IOTDatrPrivacy	Edge	IOTDatrPrivacy,Edge1	Reachable	Ok	Verified	Exactly 2 trace patterns.	2 trace patterns
	Identity2	IOTDatrPrivacy,Identity21	Reachable	Ok	Verified	Exactly 1 trace pattern.	1 trace pattern
	Identity1	IOTDatrPrivacy,Identity11	Reachable	Ok	Verified	Exactly 1 trace pattern.	1 trace pattern
	Object	IOTDatrPrivacy,Object1	Reachable	Ok	Verified	Exactly 1 trace pattern.	1 trace pattern
	Manager	IOTDatrPrivacy,Manager1	Reachable	Ok	Verified	Exactly 1 trace pattern.	1 trace pattern
	Gateway	IOTDatrPrivacy,Gateway1	Reachable	Ok	Verified	Exactly 4 trace patterns.	4 trace patterns
	Cloud	IOTDatrPrivacy,Cloud1	Reachable	Ok	Verified	Exactly 8 trace patterns.	8 trace patterns

Fig. 2. Scyther tool results.

Once fake identity is got, the objects start to send collected data to cloud. Fig. 3 shows attack trace for registration and anonymization phase, the object is modeled as role Object and manager is modeled as role Manager.

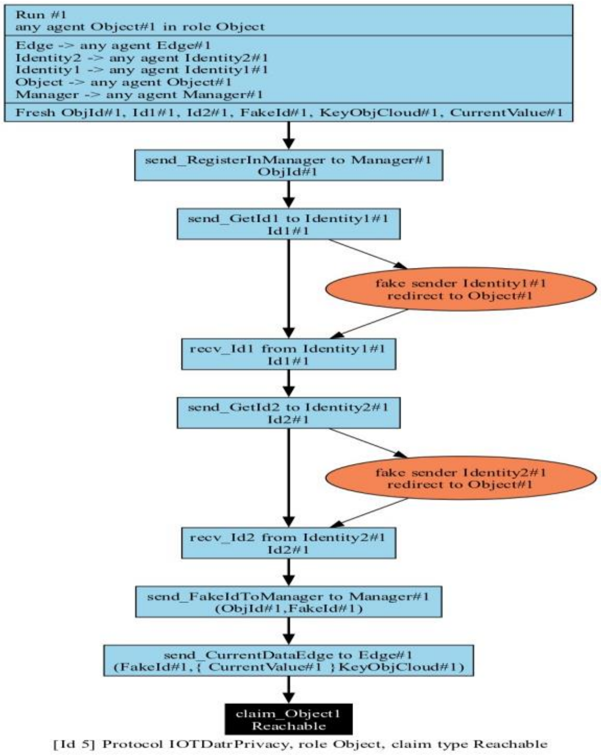

Run #1
any agent Object#1 in role Object
Edge -> any agent Edge#1
Identity2 -> any agent Identity2#1
Identity1 -> any agent Identity1#1
Object -> any agent Object#1
Manager -> any agent Manager#1
Fresh ObjId#1, Id1#1, Id2#1, FakeId#1, KeyObjCloud#1, CurrentValue#1
send_RegisterInManager to Manager#1
ObjId#1
send_GetId1 to Identity1#1
Id1#1
fake sender Identity1#1
redirect to Object#1
recv_Id1 from Identity1#1
Id1#1
send_GetId2 to Identity2#1
Id2#1
fake sender Identity2#1
redirect to Object#1
recv_Id2 from Identity2#1
Id2#1
send_FakeIdToManager to Manager#1
(ObjId#1,FakeId#1)
send_CurrentDataEdge to Edge#1
(FakeId#1,{ CurrentValue#1 }KeyObjCloud#1)
claim_Object1
Reachable
[Id 5] Protocol IOTDatrPrivacy, role Object, claim type Reachable

Fig. 3. Registeration and anonymization phase analysis.

For data transfer to cloud phase, this phase provides end to end secured channel between objects and cloud to safely transfer data from objects to cloud. Objects start communication by encrypting needed data by pre-established session key and then send encrypted data to edge with fake identity constructed. The edge will add needed metadata (timestamp, edge id, gateway id) to data received and then encrypt data and metadata with cloud public key and add message authentication code to the request to be verified in order to ensure data integrity. The protocol provides two levels of protection to fulfill data privacy:

- From object to cloud which mainly encrypts data using lightweight Speck-R due to nature of objects as being constrained devices. This level protects data from active/passive attacks.
- From edge to cloud which mainly encrypts previously encrypted data and metadata using public keys for cloud. This is considered a wrapper for the first level which means after encrypting object's data with speck-R, the edge adds necessary metadata and encrypt the whole data which already have encrypted read with public key for cloud which in turn gives more protection level.

The edge will send the whole request to GT to check whether the sent id is on list updated by managers or not. If yes, the request will be sent to cloud to be verified and stored.

V. Evaluation

In this section, a comprehensive evaluation of proposed protocol is presented by comparing it against xiluo [25] protocol. The criteria will be divided into two parts:

1) Strength of protocols to fulfill the following:

- Object Registration: The protocol must have the ability to make sure that no unauthorized object can be added to the network to avoid any injection and impersonation attack.
- Identity Protection: The protocol must have the ability to provide way to protect object identities to avoid impersonation and tracking attacks.
- Key Management: The protocol must have the ability to provide a way of constructing session keys without depending on trusted third parties or physically exchange it to make sure that the key will still be secured until changing it.
- Data Protection Level: The protocol must have the ability to protect data and metadata exposed from objects.
- Confidentiality: The protocol must have the ability to protect transferred data from eavesdropping and sniffing.
- Integrity: The protocol must have the ability to protect transferred data from tampering and modification
- Mutual Authentication: The protocol must have the ability to provide a way for communication parties to authenticate each other before starting communication.

Table III demonstrates the comparison in terms of protocol strength criteria.

TABLE III. COMPARISON BETWEEN XILOU AND PROPOSED PROTOCOL IN TERMS OF STRENGTH OF PROTOCOL

Criteria	XI LUO Protocol [25]	Proposed Protocol	Comment
Object Registration	Depends on Control Center which already had predefined records for objects (ID, Key) then after verification, the join request is accepted or rejected.	Depends mainly on cluster manager which already has a predefined records for objects (ID) then after validation, the join request is accepted or rejected	XI Luo Protocol is more powerful on registration phase since the verification is done through decryption of message using key that is already stored on control center. This operation is done once and may cost additional performance but fulfill more security against injection attacks
Identity Protection	Does not provide identity protection, in communication, it works directly with objects real identity	Provides identity protection via anonymization phase which depends on identity servers to change object real identity to another one	The proposed protocol is more powerful on identity protection. No communication is done unless object changed its real identity to another one to prevent identity tracking and impersonation attacks. The new identity must be frequently changed due to its expiration which in turn provide extra layer of security to prevent eavesdropping and inference attacks. The new identity is communicated to cloud to be able to receive information and correctly map it to correct object.
Key Management	Depends mainly on Control center to create the session key between communication parties. Therefore, in order to create a key between two objects, the first one will send a request to control center to create shared key then control center will back to requester and other party with needed key to start communication.	Depends mainly on lightweight key agreement protocol which aims to establish session key between communication parties only without need for any third parties. The communication parties only have the constructed session keys which provides more protection against key leakage	Proposed protocol employs a powerful key management approach via a lightweight key agreement protocol which ensures perfect secrecy and mutual authentication unlike Xi luo protocol, it depends on control center to provide that via persistent encryption keys stored on control center.
Data protection Level	Focuses mainly on objects' data	Focuses on objects' data and metadata related.	Proposed protocol focuses not only on data but metadata as well due to valuable information that can be revealed from.
Confidentiality	Partially fulfilled. Usage of cryptography to send and receive any message fulfill confidentiality but ability of control center to decrypt any sent message due to have access to all session keys is considered violation.	Fulfilled, being a TTP-Free and relying on End-to-end encryption starting from key construction which relies on lightweight key agreement protocol that is totally constructed between communication parties only then use of lightweight cryptography for any message sent or receive to ensure that no party whether internal or external network can read the message except authorized parties	Proposed protocol is considered more powerful in providing confidentiality. Xi luo protocol depends on key construction totally on control center which gives it ability to read any message sent due to have access to all keys used which in turn considered violation for confidentiality unlike proposed protocol which key construction is totally between communication parties only to forbid unauthorized access to data.
Integrity	Fulfilled via message authentication code used while exchanging messages	Fulfilled via message authentication code used while exchanging messages	Both fulfill integrity to make sure that any tampered message will be detected which in turn resist active attacks.
Mutual Authentication	Does not provided, communication parties does not have ability to securely authenticate each other due to being trusted to control center to authenticate each party before session key construction.	Fulfilled via key agreement protocol. Key agreement protocol aims to establish a session key between communication parties to securely exchange messages between each other. The first stage of adopted protocol is mutual authentication to make sure that each party authenticated each other before starting session key construction.	This is one of most critical part to be considered. Proposed protocol gives ability to communication parties to mutually authenticate each other before starting communication to resist any impersonate attacks and make sure that no one pretends to be other which in turn constitutes to confidentiality indirectly.

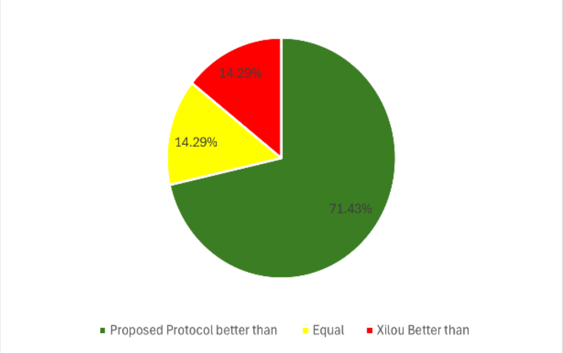


Fig. 4. Evaluation of proposed protocol in terms of protocol strength.

Although Xi Lou protocol tried to fulfill data privacy, it mainly depends on control center as a trusted third party to fulfill key exchange and object registration which in turn is considered security concern. If any attacker gain access to the control center, he will have access to all security keys for all individual objects and shared keys between objects. Fig. 4 summarizes results in terms of percentages with percentage of superiority of proposed protocol over XiLuo's one.

2) Attacks can resist which are the following:

- Man In the middle.
- Eavesdropping and Interference

- False Data Injection attack.
- Advanced Persistent Threat
- Active attacks.
- Reply attacks
- Tracking attacks.

Table IV demonstrates the comparison in terms of stated attacks.

Therefore, the proposed protocol has the ability to fully resist mentioned attacks and provide full data privacy. Fig. 5 summarizes the attacks resistance in terms of fully, partially and not resistant.

TABLE IV. COMPARISON BETWEEN XILOU AND PROPOSED PROTOCOL IN TERMS OF ATTACKS RESISTANCE

Attacks	**XI LUO Protocol [25]**	**Proposed Protocol**	**Comment**
Man in the middle attack (MiM)	Partially Protected	Fully Protected	For Xi lou protocol: dependency on control center as a trusted third party which makes the whole system vulnerable to MiM attack if attacker gain access to that. Proposed Protocol: provides end-to-end protection without relying on trusted third party. Objects try to establish their key via lightweight key agreement protocol instead of direct exchanging them which in turn provides fully protection against MiM.
Eavesdropping and Interference	Fully Protected	Fully Protected	Traffic on network is always encrypted which provide protection against Eavesdropping and inference
False Data Injection attack	Fully Protected	Fully Protected	Xi lou and proposed protocol fulfill protection against false data injection via encryption decryption used so no one can inject any data unless have a valid key on network. For Proposed Protocol: The protocol capable of resisting injection attacks by relying on end-to-end encryption which oblige all communication parties to establish session key before communication which ensures mutual authentication for communication parties. On the other side, on both protocols, registration phase works effectively against that attack as non-node can be injected on network without having a key registered already on control center in case of xi lou or have cluster manager approval in case of proposed protocol.
Advanced Persistent Threat	Partially Protected	Fully Protected	For Xi lou protocol: if attacker gain access to control center, attacker can be silent and has ability to extract all session keys constructed inside control center between communication parties. Proposed Protocol: it provides end to end security so gaining access to any part of network and stay silent will not reveal any information to attacker.
Active attacks	Fully Protected	Fully Protected	All messages sent are equipped with message authentication codes to resist any tampering to messages sent which in turn protect data from active attacks
Reply attacks	Fully Protected	Fully Protected	All messages sent are equipped with timestamps, these timestamps are validated against receiving party to make sure that no reply attack has been carried out.
Tracking attacks	Not protected	Fully Protected	For Xi lou protocol: the protocol deals with objects using their real identities without any masking or anonymization which in turn enables attacker to track any object inside network. Proposed Protocol: it provides anonymization for objects identities by replacing real object identities with fake one via identity servers which in turn change periodically object identities while data transfer which forbid and avoid any tracking attacks.

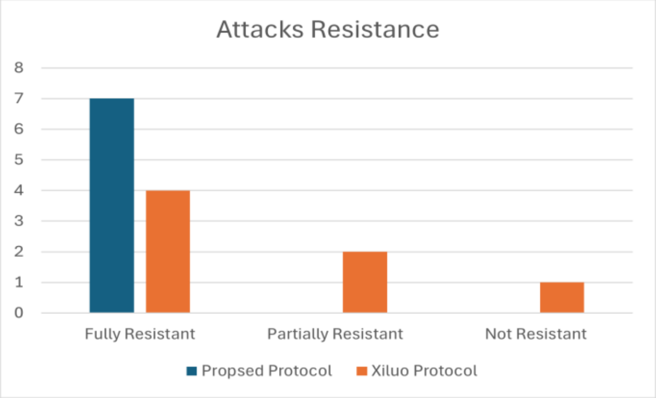


Fig. 5. Evaluation of proposed protocol in terms of attacks resistance.

VI. CONCLUSION

In this paper, a lightweight TTP-Free privacy preservation protocol is presented in order to provide complete data privacy via end to end protection for data and metadata while being transferred to cloud. The protocol depends on providing end to end encryption with a powerful lightweight key agreement protocol to fulfill full data privacy. The proposed protocol was analyzed using scyther tool to guarantee that no vulnerabilities are found. Furthermore, it was evaluated against others protocol in terms of protocol design criteria and attacks resistance. The results showed that the proposed protocol is well designed against the mentioned criteria and surpasses to other protocol by five out of seven which represents 71.4% of overall criteria and equalized in one criteria which is represented by 14.3% which makes the proposed protocol overall fulfillment is 85.7%

against 28.6% for others protocol. Moreover, the proposed protocol has ability to fully resist mentioned attacks which in turn makes it more suitable to fulfill data privacy objective.

In the future work, the proposed protocol will be extended to fulfill device to device (D2D) data privacy to affirm that all communications whether D2D or D2C are protected.

REFERENCES

[1] Atzori, Luigi, Antonio Iera, and Giacomo Morabito. "The internet of things: A survey." Computer networks 54.15 (2010): 2787-2805.

[2] INFSO, D. "Networked Enterprise & RFID INFSO G. 2 Micro & Nanosystems." Co-operation with the Working Group RFID of the ETP EPOSS, Internet of Things in (2020).

[3] Naresh, Vankamamidi Srinivasa, et al. "Internet of Things in Healthcare: Architecture, Applications, Challenges, and Solutions." Comput. Syst. Sci. Eng. 35.6 (2020): 411-421.

[4] Kadhim, Kadhim Takleef, et al. "An Overview of Patient's Health Status Monitoring System Based on Internet of Things (IoT)." Wireless Personal Communications 114.3 (2020).

[5] Sangeethalakshmi, K., U. Preethi, and S. Pavithra. "Patient health monitoring system using IoT." *Materials Today: Proceedings* 80 (2023): 2228-2231.

[6] Kumar, Mohit, et al. "Healthcare Internet of Things (H-IoT): Current Trends, Future Prospects, Applications, Challenges, and Security Issues." *Electronics* 12.9 (2023): 2050.

[7] Onasanya, Adeniyi, and Maher Elshakankiri. "Smart integrated IoT healthcare system for cancer care." *Wireless Networks* 27 (2021): 4297-4312.

[8] Onasanya, Adeniyi, and Maher Elshakankiri. "Secured cancer care and cloud services in IoT/WSN based medical systems." Smart Grid and Internet of Things: Second EAI International Conference, SGIoT 2018, Niagara Falls, ON, Canada, July 11, 2018, Proceedings 2. Springer International Publishing, 2019.

[9] Singh, Ravi Pratap, et al. "Internet of things (IoT) applications to fight against COVID-19 pandemic." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14.4 (2020): 521-524.

[10] Bhardwaj, Vaneeta, Rajat Joshi, and Anshu Mli Gaur. "IoT-based smart health monitoring system for COVID-19." *SN Computer Science* 3.2 (2022): 137.

[11] Ruan, Junhu, et al. "Agriculture IoT: Emerging trends, cooperation networks, and outlook." *IEEE Wireless Communications* 26.6 (2019): 56-63.

[12] Kitpo, Nuttakarn, et al. "Internet of things for greenhouse monitoring system using deep learning and bot notification services." 2019 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2019.

[13] Santhosh, N., M. Srinivsan, and K. Ragupathy. "Internet of Things (IoT) in smart manufacturing." *IOP Conference Series: Materials Science and Engineering*. Vol. 764. No. 1. IOP Publishing, 2020.

[14] Serror, Martin, et al. "Challenges and opportunities in securing the industrial internet of things." IEEE Transactions on Industrial Informatics 17.5 (2020): 2985-2996.

[15] Onesimu, J. Andrew, J. Karthikeyan, and Yuichi Sei. "An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services." *Peer-to-Peer Networking and Applications* 14 (2021): 1629-1649.

[16] Zhen, X. U. E., and L. I. U. Xingyue. "Providing a Framework for Security Management in Internet of Things." International Journal of Advanced Computer Science and Applications 13.11 (2022).

[17] Javaid, Uzair, Muhammad Naveed Aman, and Biplab Sikdar. "Blockpro: Blockchain based data provenance and integrity for secure iot environments." *Proceedings of the 1st Workshop on Blockchain-enabled Networked Sensor Systems*. 2018.

[18] Alzoubi, Yehia Ibrahim, et al. "Internet of things and blockchain integration: security, privacy, technical, and design challenges." *Future Internet* 14.7 (2022): 216.

[19] Othman, Soufiene Ben, et al. "Privacy-preserving aware data aggregation for IoT-based healthcare with green computing technologies." *Computers and Electrical Engineering* 101 (2022): 108025.

[20] Jayaraman, Prem Prakash, et al. "Privacy preserving Internet of Things: From privacy techniques to a blueprint architecture and efficient implementation." Future Generation Computer Systems 76 (2017): 540-549.

[21] Soldatos, John, et al. "Openiot: Open source internet-of-things in the cloud." Interoperability and Open-Source Solutions for the Internet of Things: International Workshop, FP7 OpenIoT Project, Held in Conjunction with SoftCOM 2014, Split, Croatia, September 18, 2014, Invited Papers. Springer International Publishing, 2015.

[22] Abu-Tair, Mamun, et al. "Towards secure and privacy-preserving IoT enabled smart home: architecture and experimental study." *Sensors* 20.21 (2020): 6131

[23] Li, Shancang, et al. "Lightweight privacy-preserving scheme using homomorphic encryption in industrial Internet of Things." *IEEE Internet of Things Journal* 9.16 (2021): 14542-14550.

[24] Ahmed, Mohammed Imtyaz, and Govindaraj Kannan. "Secure and lightweight privacy preserving Internet of things integration for remote patient monitoring." *Journal of King Saud University-Computer and Information Sciences* 34.9 (2022): 6895-6908.

[25] Luo, Xi, et al. "A lightweight privacy-preserving communication protocol for heterogeneous IoT environment." IEEE Access 8 (2020): 67192-67204.

[26] Soni, Mukesh, and Dileep Kumar Singh. "LAKA: lightweight authentication and key agreement protocol for internet of things based wireless body area network." Wireless personal communications 127.2 (2022): 1067-1084.

[27] Sleem, Lama, and Raphael Couturier. "Speck-R: An ultra light-weight cryptographic scheme for Internet of Things." *Multimedia Tools and Applications* 80 (2021): 17067-17102.

[28] Van, Dang Hai, and Nguyen Dinh Thuc. "A privacy preserving message authentication code." IT Convergence and Security (ICITCS), 2015 5th International Conference on. IEEE, 2015.

[29] Herzog, Jonathan. "A computational interpretation of Dolev–Yao adversaries." *Theoretical Computer Science* 340.1 (2005): 57-81.

[30] Cremers, Cas JF. "Unbounded verification, falsification, and characterization of security protocols by pattern refinement." *Proceedings of the 15th ACM conference on Computer and communications security*. 2008.

Basira: An Intelligent Mobile Application for Real-Time Comprehensive Assistance for Visually Impaired Navigation

Amal Alshahrani, Areej Alqurashi, Nuha Imam, Amjad Alghamdi, Raghad Alzahrani
College of Computing-Computer Science and Artificial Intelligence Department,
Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Individuals with visual impairments face numerous challenges in their daily lives, with navigating streets and public spaces being particularly daunting. The inability to identify safe crossing locations and assess the feasibility of crossing significantly restricts their mobility and independence. The profound impact of visual impairments on daily activities underscores the urgent need for solutions to improve mobility and enhance safety. This study aims to address this pressing issue by leveraging computer vision and deep learning techniques to enhance object detection capabilities. The Basira mobile application was developed using the Flutter platform and integrated with a detection model. The application features voice command functionality to guide users during navigation and assist in identifying daily items. It can recognize a wide range of obstacles and objects in real-time, enabling users to make informed decisions while navigating. Initial testing of the application has shown promising results, with clear improvements in users' ability to navigate safely and confidently in various environments. Basira enhances independence and contributes to improving the quality of life for individuals with visual impairments. This study represents a significant step towards developing innovative technological solutions aimed at enabling all individuals to navigate freely and safely.**

Keywords—Visual impairment; mobility application; computer vision; object detection; obstacle detection

I. Introduction

People with visual impairments suffer from many challenges in their daily lives, and among these basic challenges is the difficulty of crossing streets and moving around in public places. The visually impaired have difficulty identifying safe crossing locations and assessing the possibility of crossing. The street is an unfamiliar and risky environment for these individuals, which affects their freedom of movement and independence. There are about 285 million people who suffer from visual impairment all over the world, including 39 million people whose vision is limited (blind), and 246 million people whose vision is impaired, according to statistics from the World Health Organization [1]. The number of blind people in Saudi Arabia is about 159 thousand blind people, according to undocumented statistics Official.

The impact of visual problems on people's daily activities is profound, with simple tasks such as detecting obstacles and finding their stuff becoming difficult. The issue of mobility for visually impaired people is a major concern to this day.

In Table I, we summarized the limitations of the current systems, namely Envisn [2], Glimpse [3], ChirpAR [4], Be My Eyes [5], and BlindSquare [6]. These drawbacks have been identified through a comparison table between these systems and Basira by carefully examining and evaluating these systems, we have identified specific areas where improvements are necessary. With the introduction of Basira, we aim to address these limitations and provide a comprehensive solution that effectively resolves these issues.

TABLE I. COMPARISON BETWEEN BASIRA AND CURRENT SYSTEMS

Application/function	**Glimpse**	**ChirpAR**	**Be My Eyes**	**BlindSquare**	**Basira**
Crossroads safely				●	●
detection the obstacles	●				●
Search for objects with voice feedback		●			●
Arabic language support		●	●	●	●

II. Previous Studies

In recent years, significant advancements have been made in applications and digital devices designed to assist individuals with visual impairments. A study conducted by Senjam et al. [6] in 2021 highlighted that smartphones are now widely accepted and less stigmatized compared to traditional assistive devices. The number of apps tailored for people with visual impairments is also on the rise, including options like VoiceOver, Aipoly Vision, TapTapSee, Be My Eyes, Seeing AI, and Seeing Assistant Move. However, many of these applications are not adequately designed to support safe navigation and road crossing.

In 2022, Mehmood et al. [7] explored the needs and challenges faced by blind and visually impaired individuals in Saudi Arabia concerning the availability and use of digital devices. Their online survey, which included 164 participants, revealed that tools like the White Cane, mobile phones, Envision, Seeing AI, VoiceOver, and Google Maps were commonly used. Mobility was identified as the primary reason for using personal devices, with white canes and mobile phones reported by 49% and 84% of respondents, respectively.

Additionally, Montezuma et al. [8] conducted research in 2021 comparing the performance of Orcam MyEye 1 and Seeing AI. Both applications demonstrated over 95% accuracy in recognizing plain text documents, but their accuracy fell to between 13% and 57% for text on curved surfaces. Participants completed 71% of tasks with Orcam MyEye 1 and 55% with Seeing AI.

In the same year, Salunkhe et al. [9] developed an Android-based object recognition app that leverages the smartphone camera to capture real-time images. This app processes images using TensorFlow's object detection API, specifically the SSD algorithm, achieving an accuracy of around 90% in experimental evaluations.

Furthermore, in 2022, See A et al. [10] proposed a system that combines obstacle detection and object recognition in a single application. Utilizing a deep learning model with the YOLO v3 framework for multi-object detection, this system employs the ARCore Depth Lab API from Google to create a 3D depth map. It identifies obstacles and provides audio alerts while recognizing over 90 different classes of objects.

Lastly, Patil et al. [11] in 2022 discussed a mobile application that integrates multiple functionalities, utilizing artificial intelligence and machine learning techniques. This includes the YOLOv3 algorithm for object recognition, a currency recognition model built with TensorFlow and a dataset from Kaggle, which contains over 1,000 different objects.

III. METHODOLOGY

Basira application has been designed to assist blind and visually impaired individuals in navigating their surroundings. The application aims to develop a portable and flexible navigational solution that meets their needs. To achieve this goal, a smartphone with a depth camera is used. The chosen smartphone for this research is the Honor COR-L29, which operates on the Android V9 operating system with a HiSilicon Kirin 970 processor. The phone features 8 GB of RAM and a large 6.3-inch IPS LCD display. The screen includes a notch at the top housing the front camera, and it offers an FHD+ resolution with a 19.5:9 aspect ratio. The screen-to-body ratio is 83%, indicating very slim bezels and the pixel density is approximately 409 pixels per inch. The rear camera setup is dual, consisting of a 16 MP primary sensor and a 2 MP secondary sensor, with a narrow f/2.2 aperture and LED flash. The camera setup supports depth sensing for portrait mode effects.

The development of the Basira application followed a structured methodology aimed at creating a comprehensive solution to aid individuals with visual impairments in navigating their environments. Central to this methodology was the meticulous design of the user interface, which prioritized flexibility and ease of use. The interface was tailored to accommodate various levels of visual impairment, with a focus on integrating intuitive voice commands for seamless interaction.

A. System Requirements

System requirements are a description of what a system should do, how it should behave, what properties it must have, and what are the limitations of the system, and it is divided into functional requirements and non-functional requirements.

- Functional and Data Requirements

1) The user shall be able to switch between the three functions by swiping the screen, whether it is “detection for obstacles/free Walk ","Cross the road" or "Search for objects".

a) The system shall provide different button for each service or provide swiping the screen.

b) The system shall display the interface of the chosen service.

2) The user should have the ability to search for objects in their surroundings using voice commands.

a) The system should be equipped with voice recognition capabilities to accurately process user voice commands and perform object searches based on the provided voice input.

3) The user should receive voice alerts through the app when using the camera feature to cross the road, providing assistance and guidance.

a) The system should integrate computer vision technology to analyze the live camera feed and provide real-time assistance for street crossing. It should detect objects (traffic light colors, crossing Line existing, whether there are cars or bicycles or not) and generate voice alerts to guide the user safely across the road.

b) The system should have a voice alert feature that can convert text-based information into voice output. It should deliver clear and concise instructions to the user regarding the street crossing process.

4) The user should receive audio feedback that provides information about the names of detected obstacles.

a) The system must utilize appropriate technologies to detect and detection for obstacles (free walk) within the camera's field of view. It should analyze the camera feed and accurately recognize obstacles.

b) The system must provide detection of the obstacles to enhance user safety while moving.

5) the user is presented with a welcome interface along with audio instructions.

a) The system displays three welcome interfaces with audio instructions.

b) The system displays the welcome interfaces only once when the application is opened for the first time.

c) The system allows the user to navigate through the welcome interfaces by swiping the screen.

- Non-Functional Requirements

1) Look and Feel Requirements:

a) Basira App will have an interface that is easy to use and accessible to all users. It will be designed with simplicity in mind and will be compatible with screen readers commonly used by blind people.

2) Usability Requirements:

a) Basira App will be designed to be easy to use for blind and visual impairments, providing clear and concise instructions.

b) The app will be organized to save users' time and effort, with intuitive navigation and a logical flow of tasks.

c) The app will be self-explanatory and not require excessive instructions or guidance.

d) The system will accurately and clearly describe objects to the user using a clear voice.

3) Performance Requirements:

a) Basira App's interface will load quickly, providing a seamless user experience.

4) Portability Requirements:

a) Basira App will be compatible with both Android and iOS platforms, ensuring its availability to a wide range of users.

Fig. 1 shows the relationship between user with different use cases and how they interact with the system. Displays the services provided by system, where the user accesses the following services: Search in the surrounding for Objects, Crossing the street, Obstacles Detecting (free walk) functions.

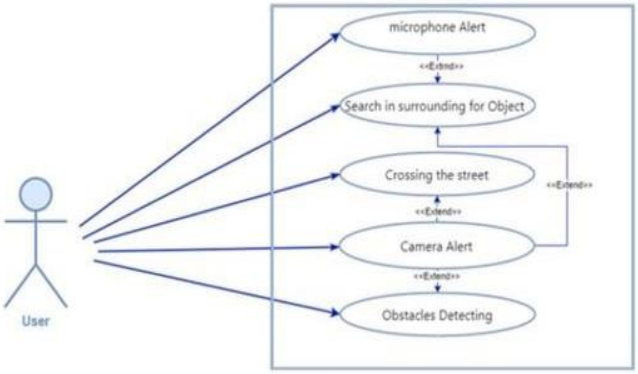


Fig. 1. Use case diagram.

B. Workflow Diagrams

We used the Data Flow Diagrams (DFDs), the data flow diagram is a graphical representation of the data stream within Basira's system. It gives a general overview of the data and system functionality.

Fig. 2 shows context diagram (Level 0) as shown in is a general overview or abstraction view of the whole Basira's system as a single process without any details. It shows the relationship between Basira's system and the external entities User, Text-to-Speech tool, Speech-toText tool and Computer vision tool.

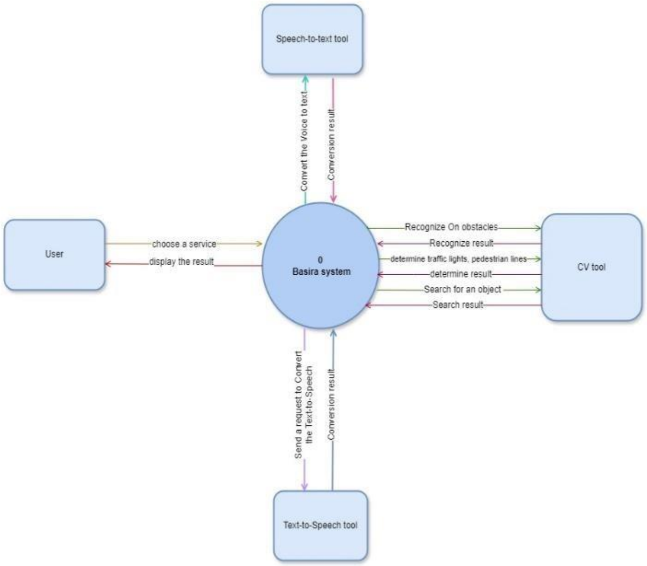


Fig. 2. Data flow diagram Level 0.

Fig. 3 shows the Data Flow Diagram (Level 1), which is a more detailed version of a context diagram. It shows the main processes of a system and the data flows between them. Here, the system has three main processes Obstacles Detecting\free Walk, Crossing the street, Search for objects.

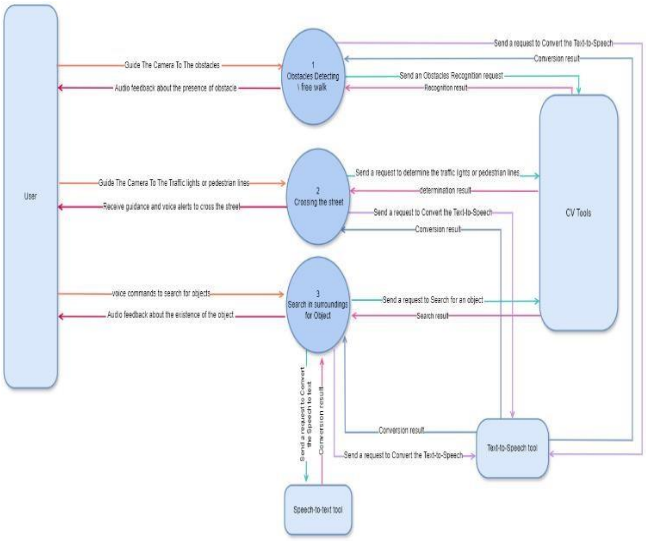

User
Guide The Camera To The obstacles
Audio feedback about the presence of obstacle
1
Obstacles Detecting \ free walk
Send a request to Convert the Text-to-Speech
Conversion result
Send an Obstacles Recognition request
Recognition result
Guide The Camera To The Traffic lights or pedestrian lines
Receive guidance and voice alerts to cross the street
2
Crossing the street
Send a request to determine the traffic lights or pedestrian lines
determination result
Send a request to Convert the Text-to-Speech
Conversion result
CV Tools
voice commands to search for objects
Audio feedback about the existence of the object
3
Search in surroundings for Object
Send a request to Search for an object
Search result
Send a request to Convert the Speech to text
Conversion result
Conversion result
Send a request to Convert the Text-to-Speech
Text-to-Speech tool
Speech-to-text tool

Fig. 3. Data flow diagram level 1.

C. Mobile Development and Implementation

The mobile application was developed utilizing Flutter, a recent and effective mobile framework to build iOS and Android applications from a single code base [12]. The Basira application included the creation of three educational audio interfaces within the application. These interfaces served as informative guides, offering users insights into the features and functionalities of Basira. By providing this educational foundation, users were empowered to navigate the application with confidence and proficiency. The Basira application includes three main functions: Crossing the Road, Free Walking, and Object Searching as shown in Fig. 4.

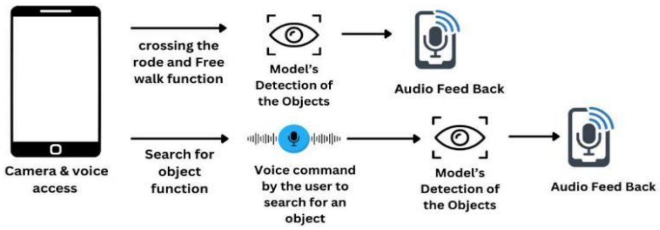

crossing the rode and Free walk function
Model's Detection of the Objects
Audio Feed Back
Camera & voice access
Search for object function
Voice command by the user to search for an object
Model's Detection of the Objects
Audio Feed Back

Fig. 4. Shows how the functions of the Basira app work.

A crucial aspect of the Basira application is the implementation of intuitive navigation mechanisms that allow users to easily traverse the app by swiping the screen. This ensures accessibility and enhances the user experience. The user-friendly interface provides quick access to features, offering real-time auditory feedback and voice commands in the object search function. This enhances users' ability to navigate safely and make informed decisions.

D. Development of AI Model

The crossing the road and street obstacle (free walking) dataset was collected from various sources, including Roboflow [13] and the Google Image Search engine. These datasets were carefully categorized to include classes such as pedestrian crossings that alert and guide the visually impaired to the presence of the street, the vehicles and cars encountered during the crossing, and traffic lights with their colours indicating stop, caution, and go. The street obstacle dataset included potholes, vehicles, traffic cones, road barriers, and natural obstacles like branches and trees. Each category in both datasets consisted of 500 carefully selected images to provide a diverse and balanced dataset for model training.

The object detection model was trained on the COCO (Microsoft Common Objects in Context) dataset, which encompasses a wide range of common objects in daily life [14]. The two models were developed and evaluated using the You Only Look Once (YOLOv5) model in study [15].

The crossing the road and free walking model achieved an accuracy of 85.4% and a size of 13.6MB in Pytorch format. The object detection model achieved an accuracy of 94.3% and a size of 14.2 MB in Pytorch format. Both models were built to operate in real-time without any user intervention.

We chose the MVC architecture in Fig. 5 for our system because it provides a clear separation of responsibilities and supports our system's action and reaction approach.

E. Search for Object Function

The 'Object Search function in the Basira application aims to assist blind and visually impaired individuals in finding the items they need. It has been implemented using the YOLOv5s model [16] for real-time object detection. When the user vocally specifies the item they are looking for, the system verifies the detected objects in real time and compares them with the desired item. Upon detecting the desired item, the user is notified via an audio alert. This function enhances the independence of blind users, helping them navigate safely and comfortably by easily identifying items without the need for assistance from sighted individuals.

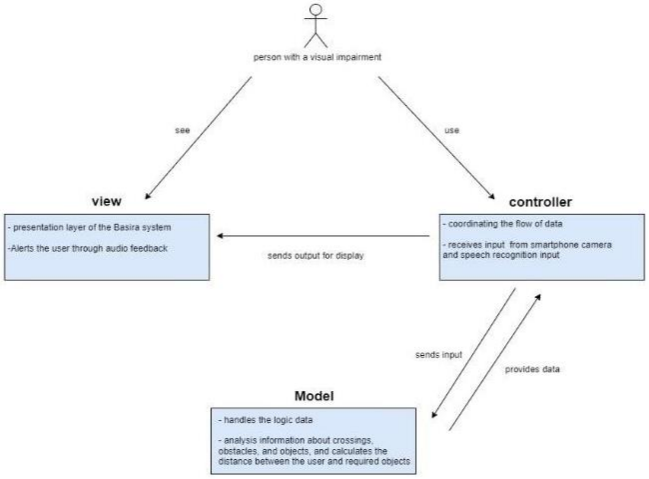

person with a visual impairment
see
use
view
- presentation layer of the Basira system
-Alerts the user through audio feedback
sends output for display
controller
- coordinating the flow of data
- receives input from smartphone camera and speech recognition input
sends input
provides data
Model
- handles the logic data
- analysis information about crossings, obstacles, and objects, and calculates the distance between the user and required objects

Fig. 5. MVC Diagram.

F. Free Walk / Obstacles Detection Function

The free walk function in Basira application is designed to help visually impaired users navigate their environment by detecting obstacles in their path. This function relies on advanced computer vision algorithm, which is YOLOv5s model, to identify various hazards such as traffic cones, potholes, cars, barriers, and trees through the smartphone's camera feed. When an obstacle is detected, the system provides immediate auditory feedback; alerting the user to the presence of an obstacle and helping them avoid potential hazards.

G. Crossing the Road Function

The crossing the street function in the Basira application is designed to ensure that visually impaired users can safely navigate pedestrian crossings. This function involves detecting traffic signals and crosswalks and providing real-time guidance about when it is safe to cross the street. The system identifies traffic lights and gives auditory guidance, such as indicating when the light is green, and it is safe to cross. Additionally, it detects the presence of vehicles near the crossing area and advises users to wait if necessary.

IV. Results and Discussion

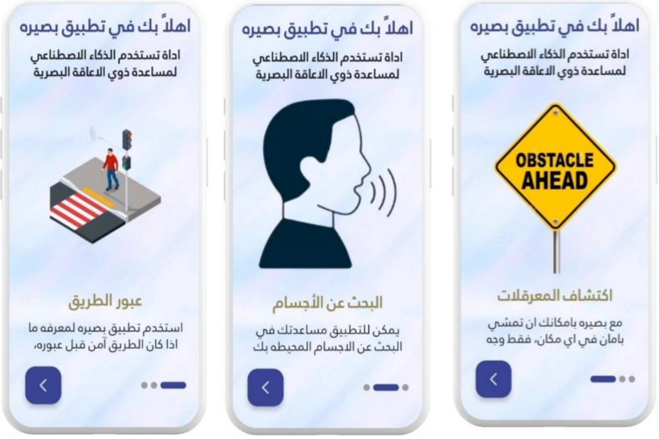
A. User Interface for Basira

Since Basira's target users are visually impaired Arabic native speakers, the language used for both the developed interfaces and audio feedback was Arabic. The interfaces were meticulously designed with a robust emphasis on accessibility and usability for visually impaired users. Drawing upon foundational design principles such as visibility, feedback, constraints, consistency, and affordance [17], these concepts were pivotal in crafting interfaces that are user-friendly and efficient.

As shown in Fig. 6, upon launching the app, users are greeted with the app logo, followed by a welcome interface only if it is their first time using the application. User onboarding is considered important because it helps new users understand how to use the application and navigate its features correctly. It aids in improving the user experience and increasing the likelihood of long-term application usage.

The welcome interface encompasses three distinct screens. Fig. 6(a) showcases the Crossing the Road feature, while Fig. 6(b) presents the Search for Objects interface. Finally, Fig. 6(c) introduces the Free Walking feature. Each interface includes detailed definitions and explanations of the respective application functionalities, accompanied by audio instructions.

Regarding the application's main interface is composed of three straightforward and simple screens. Upon launching the application (excluding the initial launch of the application), the user is presented with the Free Walking interface as the primary screen. This interface displays the camera and includes a sliding visual element that appears as a notification when obstacles are detected, as shown in Fig. 7(a). Swiping left navigates the user to the second interface which is Crossing the Road, It also displays the camera, accompanied by a sliding visual element that appears as a message when a pedestrian lines or traffic light is detected, as depicted in Fig. 7(b). The Search for Objects interface, shown in Fig. 7(c).



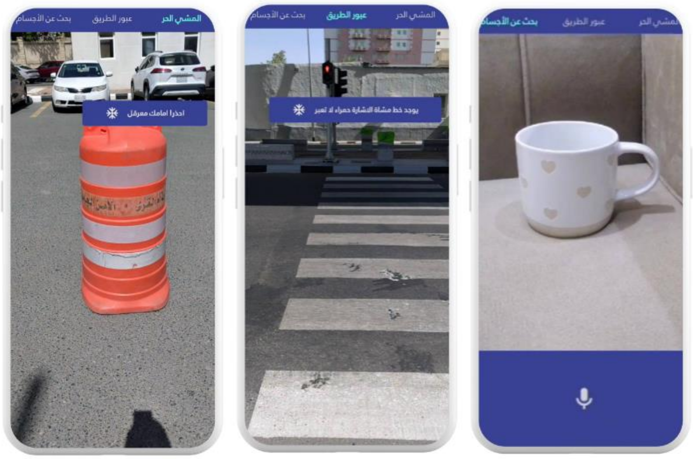
اهلاً بك في تطبيق بصيره
اداة تستخدم الذكاء الاصطناعي
لمساعدة ذوي الاعاقة البصرية
عبور الطريق
استخدم تطبيق بصيره لمعرفه ما
اذا كان الطريق آمن قبل عبوره،

اهلاً بك في تطبيق بصيره
اداة تستخدم الذكاء الاصطناعي
لمساعدة ذوي الاعاقة البصرية
البحث عن الأجسام
يمكن للتطبيق مساعدتك في
البحث عن الاجسام المحيطه بك

اهلاً بك في تطبيق بصيره
اداة تستخدم الذكاء الاصطناعي
لمساعدة ذوي الاعاقة البصرية
اكتشاف المعرقلات
مع بصيره بامكانك ان تمشي
بامان في اي مكان، فقط وجه

(a) (b) (c)

Fig. 6. Basira welcome interfaces: (a) Crossing the Road interface, (b) Search for Objects interfaces, (c) Free Walking interfaces.



احذرا امامك معرقل

يوجد خط مشاة الاشارة حمراء لا تعبر

(a) (b) (c)

Fig. 7. Basira main interfaces: (a) Free Walking interfaces, (b) Crossing the Road interface, (c) Search for Objects interfaces.

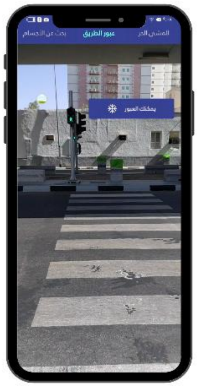
Features a split design: the upper portion displays camera, while the lower part containing a microphone button. By pressing this button, user can voice command to search for specific objects. Each interface provides audio feedback and notifications to the user, enhancing the overall user experience.

B. Testing Results

The Basira application has been tested on a single user from the target audience of individuals with visual impairments (complete or partial blindness), and the individual who has utilized the Basira to undergo testing is a person who is completely blind. Three key functionalities need to be tested by this user: crossing the road, free walking, and searching for objects. We have conducted unit testing, which involves evaluating individual components to ensure their correct operation, to detect and resolve bugs early in the development process.

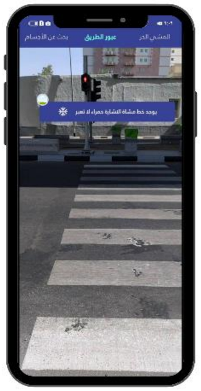
And it's worth noting that in all the scenarios mentioned below, the user will have their camera open and pointed in the direction they are walking. The application operates in real-time, so the user's gaze will be tracked accordingly.

1) Testing for “crossing the road” function: In the first scenario (Fig. 8) that was tested, a designated crosswalk is present, and a green traffic signal is displayed, indicating that vehicles are required to yield to pedestrians. Additionally, there are no cars or bicycles in the immediate vicinity of the crossing area, it will provide audio feedback in Arabic language stating "***يمكنك العبور***” (you can cross).



يمكنك العبور

Fig. 8. Testing result for “Safe to Cross” condition.



يوجد خط مشاة الاشارة حمراء لا تعبر

Fig. 9. Testing result for “Red Traffic Light at Crosswalk” condition.

In the second scenario (Fig. 9), both a red traffic light and a designated crosswalk are detected simultaneously. This combination of conditions indicates that it is not safe for pedestrians to cross the road at this time, it will provide audio feedback in Arabic language stating “ يوجد خط مشاة الاشارة حمراء لا "تعبر”(there is a crosswalk, the light is red, do not cross).

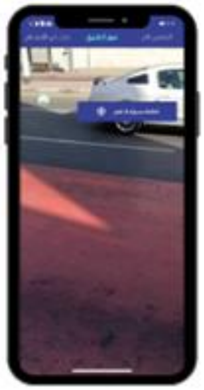


Fig. 10. Testing result for “Cars or Bicycles Present” condition.

In third scenario (Fig. 10), a car is detected, indicating that it is not safe to cross the road. This condition applies regardless of whether the detected vehicles are cars, bicycles, or both. The presence of any moving vehicles near the crossing area poses a significant risk to pedestrians, it will provide audio feedback in Arabic language stating”امامك سيارة لا تعبر ” (There is a car ahead, do not cross).

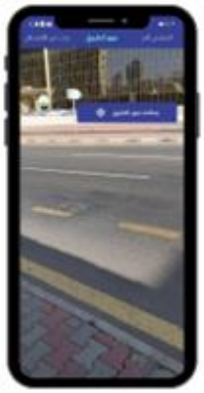


Fig. 11. Testing result for “Safe to Cross (default)” condition.

In fourth scenario (Fig. 11), there are no visible red or yellow traffic signals, and there are no cars or bicycles in the vicinity of the crossing area. Pedestrians are able to cross the crosswalk safely without any anticipated risk, it will provide audio feedback in Arabic language stating”يمكنك عبور الطريق ” (you can cross the road).

2) Testing for “free walk / obstacles detection” function: Fig. 12, if Basira detects any obstacles such as: traffic cone, pothole, car, tree or barrier in path of the user, it will provide audio feedback in Arabic language stating ***“امامك معرقل”*** (obstacle ahead).

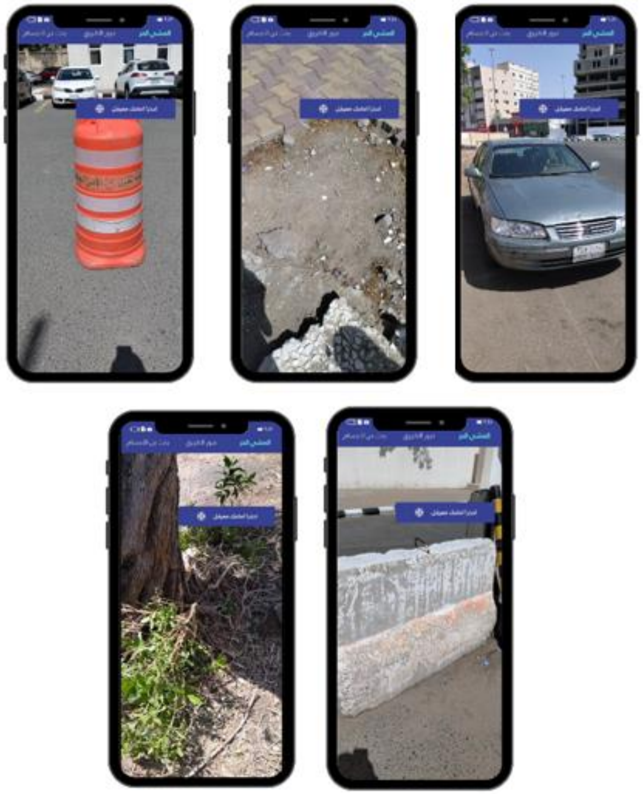


Fig. 12. Testing results for “free walking (obstacles detection)” function.

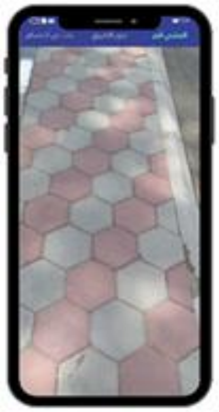


Fig. 13. Testing result if no obstacle ahead.

In second scenario in (Fig. 13), if Basira does not detect any obstacles in the user's path, it will not provide any feedback.

3) Testing for “searching for object” function: In the following functionality, the user will utilize their voice. They will press the voice recognition icon and then state the name of the object they are searching for in Arabic.

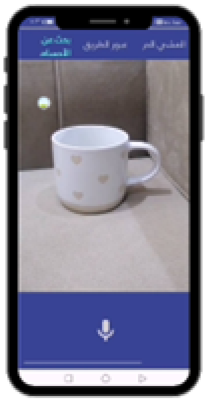


Fig. 14. Testing result for searching about “كأس” cup.

Fig. 14 the user said, **“كأس”** (cup). Then, Basira it will attempt to detect whether the cup is present in front of the camera or not. In this case, the cup is present, so it will provide audio feedback saying **“الشيء الذي تبحث عنه امامك”** (the thing you are looking for is in front of you).

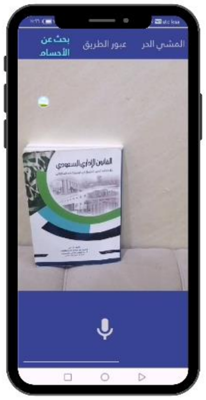


Fig. 15. Testing result for searching about “كأس” cup.

Fig. 15, we will repeat the same experiment but with a different object. The user said, **“كأس”** (cup). Then, Basira will attempt to detect whether the cup is present in front of the camera or not. In this case, the cup is present, so it will provide audio feedback saying **“لم يتم العثور عن الشيء الذي تبحث عنه”** (The object you were looking for was not found). Because the actual object in front of the camera is a book.

V. Limitations and Future Work

Throughout the development phase of our project, we faced several challenges that required innovative solutions. Notably, Flutter lacks built-in support for real-time detection using the YOLO model. To address this, we converted the YOLO model into a TorchScript extension, enabling efficient real-time object detection through frame transmission within our Flutter application. Moreover, time constraints hindered our ability to prepare more data and enhance model accuracy. Initially, we planned to connect Python code with Flutter interfaces and the camera via Flask; however, this setup encountered real-time performance issues, prompting us to rewrite the program entirely in Flutter for better integration.

VI. Conclusion

This study introduces the Basira application, leveraging Deep Learning technology to aid blind and visually impaired individuals. The application offers three primary functions. Firstly, it provides auditory alerts to users regarding obstacles such as trees, barriers, and other impediments on the street. Secondly, it helps users safely cross roads through auditory notifications based on various scenarios, such as pedestrian crossings, traffic light signals (green for crossing, red for waiting), and vehicle presence. Lastly, Basira includes an object search feature, enabling users to vocally inquire about objects in their vicinity. The system responds with auditory alerts confirming the object's presence or absence.

The mobile interface of Basira was designed to look friendly and easy to use. Developed using Flutter, the application is compatible with Android devices. When the application is opened, a welcome interface will appear, showcasing the features offered by Basira. The user is then taken to the free walking (obstacle detection) interface, where they can swipe the screen to smoothly navigate to the road crossing page or the object search page. By seamlessly integrating deep learning models with a user-friendly mobile app, Basira aims to enhance the independence and mobility of blind and visually impaired individuals.

References

[1] World Health Organization. (2010, February 12). Health systems financing: the path to universal coverage. Retrieved from: https://www.who.int/whr/2010/en/

[2] Envision,App,"Envision,[Online].Available:https://www.letsenvision.com/app.

[3] GlimpseAR,Visual,Aid,"AppStore,[Online].Available: https://apps.apple.com/sa/app/glimpse-ar-visual-aid/id1311012359?l=ar.

[4] ChirpAR,"App,Store,[Online].Available:https://apps.apple.com/us/app/chirp-ar/id1439199416.

[5] Be My Eyes Be My Eyes - See the world together. Retrieved from https://www.bemyeyes.com/

[6] Senjam, S. S., Manna, S., & Bascaran, C. (2021). Smartphones- Based Assistive Technology: Accessibility Features and Apps for People with Visual Impairment, and its Usage, Challenges, and Usability Testing. Clinical Optometry, 13, 311-322. DOI: 10.2147/OPTO.S336361. Available at: https://www.tandfonline.com/doi/full/10.2147/OPTO.S336361

[7] Busaeed, S., Mehmood, R., & Katib, I. (2022). Requirements, Challenges, and Use of Digital Devices and Apps for Blind and Visually Impaired. NOT PEER-REVIEWED. Preprints [Online]. Available at: https://www.preprints.org/manuscript/202207.0068/v1

[8] Granquist, C., Sun, S. Y., Montezuma, S. R., Tran, T. M., Gage, R., & Legge, G. E. (2021). Evaluation and Comparison of Artificial Intelligence Vision Aids: Orcam MyEye 1 and Seeing AI. Journal of Visual Impairment & Blindness, 115(4), 277-285. Available at: https://journals.sagepub.com/doi/abs/10.1177/0145482X211027492

[9] Salunkhe, A., Raut, M., Santra, S., & Bhagwat, S. (2021). Android-based object recognition application for visually impaired. *ITM Web of Conferences*, 40, 03001. [Online]. Available: https://doi.org/10.1051/itmconf/20214003001

[10] See, A. R., Sasing, B. G., & Advincula, W. D. (2022). A Smartphone-Based Mobility Assistant Using Depth Imaging for Visually Impaired and Blind. *Applied Sciences*, 12(6), 2802. [Online]. Available: https://doi.org/10.3390/app12062802.

[11] Patil, R., Modi, R., Parandekar, A., & Deone, J. B. (2022). Designing mobile application for Visually Impaired and Blind Persons. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4108763.

[12] BlindSquare Retrieved from https://www.blindsquare.com/

[13] Flutter-Build apps for any screen." Google, [Online]. https://flutter.dev/

[14] Roboflow. (n.d.). What's New in YOLOv8? Roboflow Blog. Retrieved from https://blog.roboflow.com/whats-new-inyolov8/#yolov8-architecture-a-deep-dive

[15] COCO Consortium. COCO - Common Objects in Context. Retrieved from https://cocodataset.org/#home.

[16] Al-Shahrani, A., Alghamdi, A., Alqurashi, A., Alzahrani, R., & Imam, N. (2024). Real-time comprehensive assistance for visually impaired navigation. International Journal of Computer Science and Network Security, 24(5), 3-4. https://doi.org/10.22937/IJCSNS.2024.24.5.1I. S.

[17] Solawetz, J. (2020). Yolov5 new versionimprovements and evaluation. Roboflow.Seachdate.Retrieved,fromhttps://blog.roboflow.com/yol ov5-improvementsand- evaluation/

Machine Translation-Based Language Modeling Enables Multi-Scenario Applications of English Language

Shengming Liu*
School of Foreign Languages and Culture, Panzhihua University, Panzhihua 617000, China

***Abstracts*—Traditional machine translation models suffer from problems such as long training time and insufficient adaptability when dealing with multiple English language scenarios. At the same time, some models often struggle to meet practical translation needs in complex language environments. A translation model that combines the feed-forward neural network decoder and the attention mechanism is suggested as a solution to this problem. Additionally, the model analyzes the similarity of the English language to enhance its translation ability. The resulting machine translation model can be applied to different English scenarios. The study's findings showed that the model performs better when the convolutional and attention layers have a higher number of layers relative to one another. The highest average value of the bilingual evaluation study for the research use model was 29.65. The research use model can machine translate different English language application scenarios and also the model performed better. The new model performed better than the traditional model and was able to translate the English language well in a variety of settings. The model used in the study had the maximum parameter data size of 4586, which is 932 higher than the lowest statistical machine translation model of 3654. The metric value was 3.96 higher than the statistical machine translation model. It is evident that investigating the use of the model can enhance the English language scene translation effect, with each scene doing well in translation. This provides new ideas for the direction of multi-scene application of machine translation language model afterwards.**

Keywords*—*Machine translation; decoder; English language; multi-scene; attention mechanism

I. INTRODUCTION

Machine translation (MT) refers to the text translation process that uses only machines to translate natural language. MT can translate a large number of English language (EL) text messages more conveniently and quickly, and reduce the communication barriers and other problems that occur in people's work [1]. As science and technology have advanced, machine translation (MT) has grown in importance as a text translation tool. But the traditional MT model can no longer meet the increasingly complex language environment, so how to improve the efficiency of MT, so that it is more suitable for a number of different scenarios has become an important direction of the current research [2]. The traditional MT model is mainly based on the Transformer to carry out multi-level stacked translation to realize the purpose of MT. The high complexity of traditional models makes them take a lot of time in the training process, which makes it difficult to meet the requirements of real-time and high efficiency in multi-scene applications. In addition, the traditional model relies on a multi-level stacked translation mechanism, which causes problems such as dimension explosion when dealing with long text sequences [3]. A type of deep-structure neural network model called a convolutional neural network uses the feed-forward neural network's (FNN) attention mechanism (AM) to shorten training times and enhance training results [4]. FNN is often used in complex data models because it is easy to be trained and can handle high-dimensional data. By processing sequential data more effectively, AM can enhance the accuracy and performance of machine learning models. The AM module can effectively reduce the dimension explosion problem when translating long sequences of text and improve translation accuracy and efficiency. The FNN module can reduce the overall complexity of the model and shorten the training time. The meta-learning module can ensure the translation effect of the model in complex scenes. The encoder-decoder structure enables efficient sequence-to-sequence translation. Due to the addition of the attention mechanism, the encoder-decoder structure can process and translate long sequence texts more accurately. On this basis, the study improves the Transformer based on the traditional MT by adding AM and FNN, and reduces the complexity of the model by extracting the local semantic information of words through AM. The research uses a decoder to decode and analyse linguistic terms and filter out closer results. The new system is able to integrate the attention mechanism and feed-forward neural network, optimising the decoder part of the traditional Transformer model. To improve the system's ability to capture the local semantic information in the language sequence and reduce the complexity of the model. At the same time the model improves the adaptation effect to different English scenarios by introducing a meta-learning approach. The study is broken down into six sections. Introduction is given in Section I. Section II gives detail about the related work. English language model for machine translation is given in Section III. Section IV provides analysis of English language. Discussion is given in Section V and finally, Section VI concludes the paper.

II. RELATED WORK

MT is a translation model for linguistic text translation through software, which is currently widely used in different translation fields. Using MT to translate independently translated sentences, Maruf S et al. were able to enhance the model's translation quality by adding a neural network model. The study also evaluated the current applicable models and

*Corresponding Author.

analyzed them using both training and test sets. The study's findings demonstrated that applying the approach can improve research in this area [5]. Translation effort is complicated by ambiguity, as discovered by Yang M et al. when certain uncommon words are frequently substituted with tokens in machine translation (MT). Therefore, a hierarchical clustering approach was proposed, by building a dataset of rare words up to be able to be assembled into the MT framework, which can be successfully translated. The results of the study indicated that the quality of MT can be improved by using this method [6]. Dabre R et al. found that the use of multilingual neural MT can improve the quality of translation, and this translation model is more convenient and faster compared with the traditional MT. Therefore the study investigated this MT method in more depth and explored the future research direction of this method by analyzing the current literature and use cases, which provides new ideas for future research in this direction [7]. Khaw L L et al. found a new model for enhancing students' English context through writing. To improve the model's capacity for contextual learning, a RA engineering model was used in the development of the new model. The study's findings demonstrated how well this methodology worked to improve students' contextual learning [8].

According to Ban H et al., as Internet technology has advanced, so too has the necessity for interlanguage communication, and language models used in automatic translation have gotten more efficient. In order to create a novel MT model using an encoder and decoder, the research applied a deep learning MT model. By mapping between languages, the new model can enhance performance by immediately converting data into vectors. The study's findings demonstrated that the new model might outperform the conventional model [9]. Bharti N et al. concluded that the ranking of translated languages can improve the performance of MT and also train the MT effectively. Therefore, the study used a new mechanism for MT by sorting the MT output statements. The study's findings showed that the novel approach can enhance machine translation (MT) performance and is helpful for translating linguistic studies [10]. Wu H found that interactive translation of MT using multimedia is an effective way of translating a large number of utterances and programming languages. Therefore, the study proposed a fuzzy learning-based MT system that uses multimedia software to improve the performance and effectiveness of MT. Finally, the study pointed out the defects and advantages of the current multimedia translation and proposed a new translation program. Additionally, the study's findings demonstrated that employing the new model can improve English translator training [11]. Maimaiti M et al. found that there are many ways to enhance the data in MT, but the traditional methods are difficult to ensure the quality of the data after enhancement. Therefore a new evaluation framework was used in the study for data enhancement and the new approach used a discriminator model in order to reduce syntactic and semantic errors. The outcomes demonstrated that the data enhancement performance of this approach is significantly better than other approaches [12].

In summary, most of the studies in MT are mainly to enhance the effect of translated data and MT performance, followed by the improvement of MT through machine learning, but this approach often fails to analyze and study more complex EL environments. To improve MT's capacity for scene adaption, the study superimposes multi-layer FNN and attention to incorporate AM and FNN into the MT model. Lastly, to improve the model's capacity to feature extract data, the data is feature extracted from the model using conditional random fields (CRF) and bidirectional encoder representations from Transformers.

III. English Language Modeling for Machine Translation

This section mainly focuses on the model construction of EL based on the MT model, firstly, the AM will be added to the model, and then the structure of the decoder and other structures will be introduced and analyzed, followed by the analysis and elaboration for the purpose of improving the model effect and coping with the translation of EL in different scenarios.

A. *Language Model for Machine Translation Incorporating the Attention Mechanism*

Language model is a processing model to convert the traditional natural language, the current commonly used language model is mainly using neural networks for language processing and model design. When creating the target language, machine translation (MT) is a crucial model of language processing that can typically convert languages, decode and analyze data from the source language, and treat the previous result as the output of the latter item to obtain the source language model formula for MT, as indicated in Eq. (1) [13].

$$P(Y \mid X, \theta) = \prod_{t=1}^{T^Y+1} P(y_t \mid y_{t-1}, X; \theta) \tag{1}$$

In Eq. (1), X denotes the source language sequence, T^Y is the length of the sequence, and Y is the final language sequence obtained. θ denotes the set of data parameters and T denotes the length of translated words. Generally the use of parameter estimation to train the machine for translation enables to get the size of the loss function value of the current formula model. Usually in the use of MT it is necessary to convert the linguistic data into dimensional vectors that can be used by the machine, so it is necessary to use an encoder to analyze the MT process. The encoder-decoder's construction is depicted in Fig. 1.

In Fig. 1, the encoder adds word data such as "I", "LOVE", "I", "HOME "The data is converted by the decoder, and then the converted data is expressed in machine language by the decoder, such as "HOME" is expressed as [1,0,1], this kind of data transcoding can solve the problem of semantic conversion, but usually there is also a dimensional explosion in the conversion process problem. Therefore, the study adds the attention module to the traditional model to solve the problem of long language sequences. Because in the translation process of EL, it is impossible to compare and analyze each word, and in fact, the translation only needs to translate the adjacent part of the word, so the mapping of the local word can get the result of all the words translated. For this reason, AM is added to the translation process of EL, which is used to capture the information of global features. As shown in Fig. 2.

In Fig. 2, the attention module mainly includes multiple attention and convolutional attention. Multiple attention includes parameters such as temporal expression of EL sequences, word expression, and so on. Convolutional attention is mainly the linguistic dimension segmentation of the convolutional layer, and the parameter data of the two kinds of attention are segmented and processed to obtain the latest model attention output. The sequence expression of AM is shown in Eq. (2) [14].

$$S = \{W_1, W_2, \cdots, W_n\} \tag{2}$$

In Eq. (2), S denotes the output of the language sequence and W_n denotes the word expression in the sequence. The sequence vector encoding at this point is shown in Eq. (3).

$$E = Input(S) \tag{3}$$

In Eq. (3), E belongs to the set of word sequence vectors, and the obtained vectors are subjected to the dimensional splitting operation of the attention input, and the different dimensions of attention are correlated to obtain the correlation shown in Eq. (4) [15].

$$d_S = d_X + d_{X_2} \tag{4}$$

In Eq. (4), d_S is the dimension size of the language sequence and d_X is the dimension size of the number of attention heads. The attention heads is set to h_T and the convolutional AM is set to h_C in Eq. (5) to obtain the AM allocation formula as shown in Eq. (5) [16].

$$h_T * d_T^h + h_C * d_C^h = d_T + d_C = d_S \tag{5}$$

In Eq. (5), d_T^h denotes the dimension assigned to each number of attention for multiple, and d_C^h denotes the dimension assigned to each number of attention for convolution. After analyzing the dimensional information of the attention, the initial attention dimension (AD) is used to build the encoder and decoder as shown in Fig. 3.

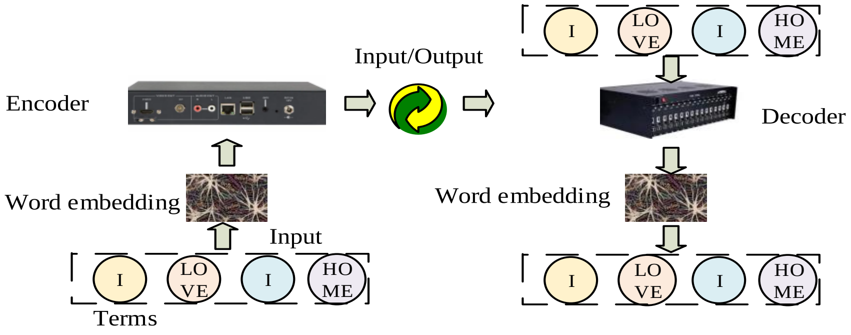


Fig. 1. Encoder decoder structure.

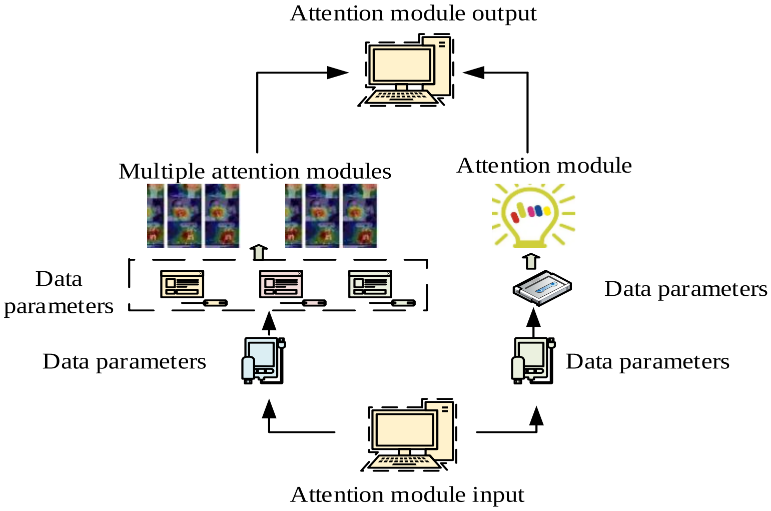


Fig. 2. Information capture process of global features.

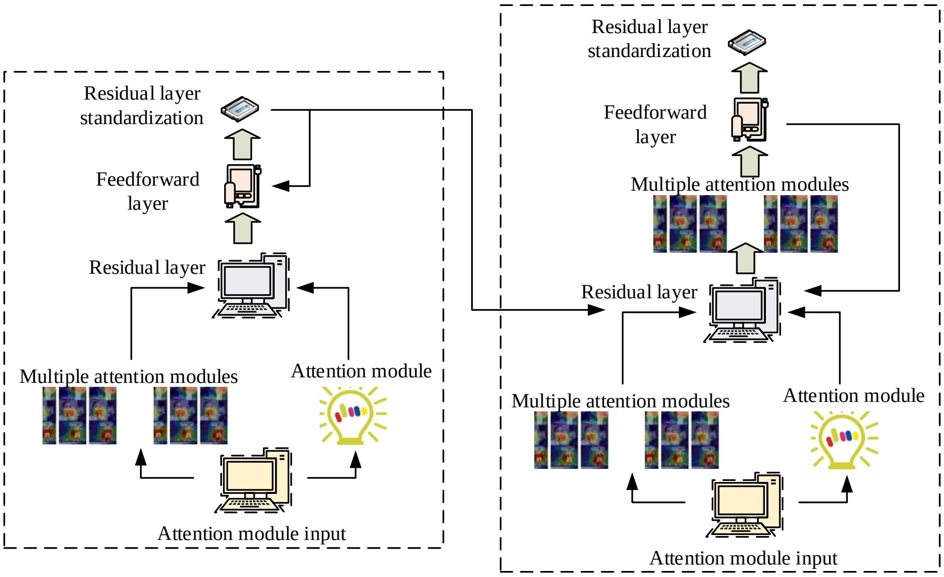

Residual layer standardization
Feedforward layer
Residual layer
Multiple attention modules
Attention module
Attention module input
Residual layer standardization
Feedforward layer
Multiple attention modules
Residual layer
Multiple attention modules
Attention module
Attention module input

Fig. 3. Initial attention dimension construction process.

Fig. 3 shows the addition of a front-loaded feedback layer and a residual connection layer during the pre-modeling stage of the encoder. The feedback layer primarily modifies the linear data to increase the attention model's computational capacity. The decoder's entire structure is the same as the encoder's, but with additional multiple attention added. The model utilized is the same as the traditional model module, and the EL data obtained after encoding in the encoder is transferred into the encoder for recomputation. In the English data input part of the model, the existing vector positions and inflectional lengths are set, the input data vectors are aligned by means of data complementary zeros, and then the initialization calculation of English word vectors is carried out by the function and the positions are encoded in the data vectors, and the final data obtained are analyzed as the input dimensions.

The encoder stage of the model first splits the data vectors to divide the different dimensions of the attention model, and the assigned dimensions of the two attention models are calculated to obtain the dimension size of the encoder attention convolution [17]. Finally the obtained results and dimensions are merged by computation and later fed into the pre-feedback network to enhance the data. The decoder output stage requires the addition of the same attentional model that enables the current model to obtain more EL text data. The obtained text data is used as input to the attention matrix of the decoder [18]. Again same as the encoder stage the result of the previous layer is augmented by the feed forward network before outputting and the final result obtained is the current desired result.

B. *Machine Translation Model for English Modal Scene Application System*

In the translation of EL scenarios, the main problem is the EL translation and usage problem, which can be solved by analyzing the MT encoder and decoder in the previous section to study the problems such as English translation. The main composition of the model used in the study is composed of AM and Transformer, mainly to enhance the module's feature extraction (FE) ability for nonlinear vectors and words in low latitude space. The upper part of the model used belongs to the feed forward network module as shown in Fig. 4.

In Fig. 4, the feedforward network module is mainly composed of multiple attention models and feedforward network module, while the network module includes data linear change operation, multiple perceptual layers and neural network layer optimization. The inputs and outputs of the feedforward network layer are expressed to obtain the formula for the FNN layer as shown in Eq. (6) [19].

$$FNN(D) = relu(DW_D + b_D)W_R + b_R \tag{6}$$

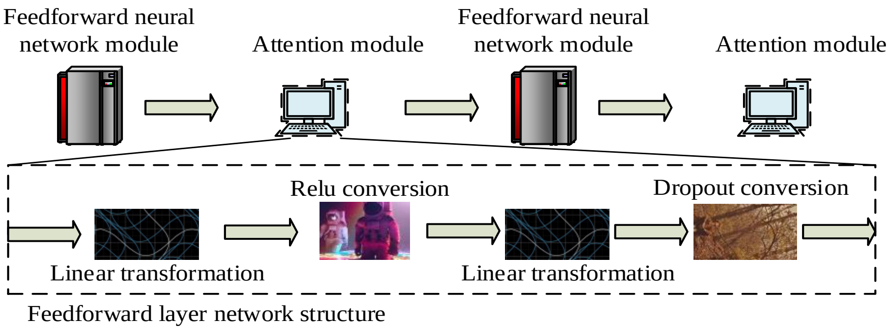


Fig. 4. Feedforward network module.

In Eq. (6), D denotes the input of data and W_R denotes the output of data. b_R is the AD of the output and b_D denotes the AD of the input. The MT model of AM has been able to translate specific ELs, but the information and terminology of the domains are not fixed for different English scenarios, and in order to further make the model adapt to more scenario applications, the model needs to be trained by meta-learning methods. Fig. 5 displays a schematic diagram of the training model.

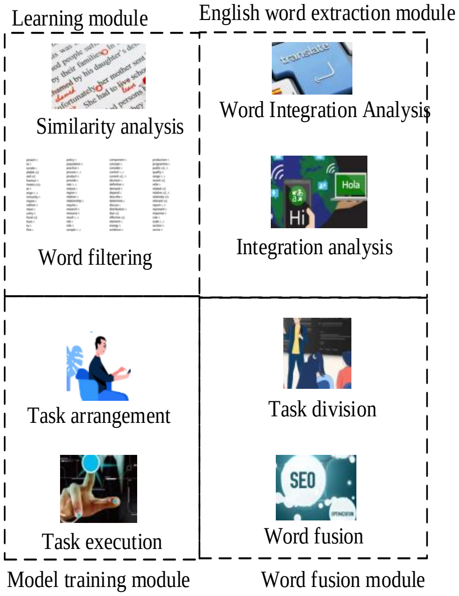


Fig. 5. Schematic diagram of training model.

In Fig. 5, the training for the model is mainly divided into several modules for training, learning module, English word extraction module, word fusion module, and model training module. The learning module is mainly responsible for word similarity analysis of the current application scenarios. The English word extraction module performs word incorporation analysis for different English scenes. The word fusion module divides the tasks for meta-learning. The model training module arranges training tasks for the training process of the model. It is vital to extract words from the language used in the current scene during the model training process in order to integrate additional information from the model training data. This is done by marking the words in the sentence and assessing them in relation to various vocabularies and scenarios for comprehension. Extracting data through the method of extracting words requires FE of data from the feature expression layer and the conditional random distribution layer [20]. Fig. 6 depicts the extracted words FE procedure.

To acquire the feature representation of the EL sentence using the model, the input sequence in Fig. 6 must first be feature extracted. Secondly, the deeper neural network is input as a conditional randomized feature layer. In the stochastic conditional feature layer it is necessary to define the state feature function and map the input feature sequence to the probability distribution of the sequence, and each state function is subjected to positional correlation calculation. If a feature function that can be identified is labeled, the value of the feature function at the current labeled position is used as the current input, and the model FE is able to return a corresponding real number. In addition to defining the function labeled for FE, it is also necessary to define the position of the feature function. The distribution of word features in the model is calculated for the best marking case given by the sequence, and the probability of the sequence is calculated using the function, as shown in Eq. (7) [21].

$$p(y\mid x)=$$

$$\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1},y_i,x_i)+\sum_{i=1}^{n}\sum_{j=1}^{K}\theta_j g(y_{i-1},y_i)\right) \tag{7}$$

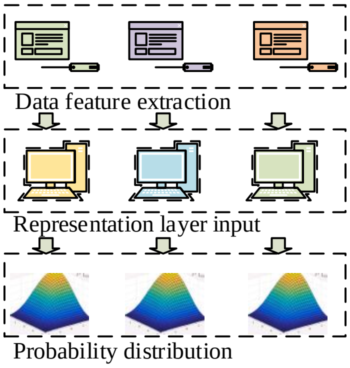


Fig. 6. Extraction process of word features.

In Eq. (7), $Z(x)$ denotes the normalization factor, which is capable of summing up all labeled sequence probabilities, guaranteeing that the distribution of the current probabilities are all integer 1. λ_k and θ_j denote the parameters of the model. $f_k(y_{i-1},y_i,x_i)$ denotes the current defined feature function. $g(y_{i-1},y_i)$ denotes the value of another defined feature function. Following the model's training, a comparable function serves as the loss function. The model's parameters are then determined by gradient descent minimization of the loss function, and the best position for the sequences is ultimately determined by decoding the data following training [22].

In addition to the extraction of feature data it is also necessary to translate the sentences in the target context, but since there are already multiple aligned words in the current environment, the study requires automatic acquisition of aligned sentences in the context and translation of the target sentence (TS). As shown in Eq. (8) [23].

$$A=(x_i,y_i):x_i\in X,y_i\in Y \tag{8}$$

In Eq. (8), $X=x_1,x_2,\cdots,x_n$ denotes a random sentence in the scene and $Y=y_1,y_2,\cdots y_m$ denotes a parallel sentence corresponding to the TS. (x_i, y_i) denotes a word pair where two words are semantically similar in the same sentence in the same context. When extracting word embedding for multiple words, the word embedding need to be obtained through the hidden state of the model, and the similarity of the words is calculated after the word embedded words are obtained. In this way the dot product of contextual principal vectors of word embedding can be computed for each feature of the target word and the aligned word. After that all the features exceeding the threshold set by the model are filtered and then the similarity matrix is obtained through the contextual word embedding as shown in Eq. (9) [24].

$$S = h_x h_y^T \tag{9}$$

In Eq. (9), S denotes the probability distribution of the similarity matrix, and both h_x and h_y denote the word embedding data. The matrix of initial phrases is obtained by this method, and then the final alignment matrix is obtained by ensemble interaction between the initial sentence (IS) and the sentences obtained from the TSs, as shown in Eq. (10) [25].

$$Align = (S_{xy} > c) \cap (S_{yx}^T > c) \tag{10}$$

In Eq. (10), S_{xy} denotes the IS matrix, S_{yx}^T denotes the TS matrix, and c denotes the threshold value. When the threshold value is set to 1 and the rest of the parameter items are set to 0, then the value of the alignment matrix in the matrix is equal to 1 means that at this time the two TSs and the IS are aligned with each other [26]. As some texts are processed for word alignment in different scenarios, multiple subwords need to be aligned [27]. However, in different scenarios where a term includes multiple words and a word contains multiple segmented sub-words, it is only necessary to align a sub-word in the scenario with each other, and then the words can be considered to be aligned with each other. However, this method will greatly increase the error generated during the training of the model, so it is also necessary to filter the aligned sentences. The method used in the model is to represent the two sentences as vectors and calculate the cosine similarity between the two words as shown in Eq. (11) [28].

$$similarity = \cos(\theta) = \frac{A * B}{|A||B|} \tag{11}$$

In Eq. (11), $similarity$ denotes the magnitude of cosine similarity between two sentences and $\cos(\theta)$ denotes the cosine value. A is the IS word and B is the TS word. For the model the loss function is calculated as shown in Eq. (12) [29].

$$L(\phi) = \sum_{t=1}^{T} l^t(\hat{\theta}^t) \tag{12}$$

In Eq. (12), $L(\phi)$ denotes the value of the loss function and $\hat{\theta}^t$ denotes the specific network in the model that performs the initialization operation. T denotes the number of executed tasks, and $l^t(\hat{\theta}^t)$ denotes the loss value of the network performing the task in the model. When the network task is executed at initialization, its execution network is the same as the initial network, then after the model training update, the execution parameters of the current network are different from those of the initial network. At this time, the loss value is calculated, and the loss value is aggregated so that all the loss values of the network model can be obtained [30]. The final flow of the obtained multi-scene translation decoding model is shown in Fig. 7.

In Fig. 7, the initial phase of the model involves first matching the EL file from the system, loading the EL, after which the model is trained through the model. It is determined whether there is a request or not, if yes then the initial EL is obtained from the text, if not then it waits for the instruction to be issued. After obtaining the IS the sentence language is pooled to choose whether to use the decoding pool or to sample the data. After that the EL sentences from different scenarios are modeled and segmented, the resulting ELs are translated into text using MT, and finally the translated results are output as text.

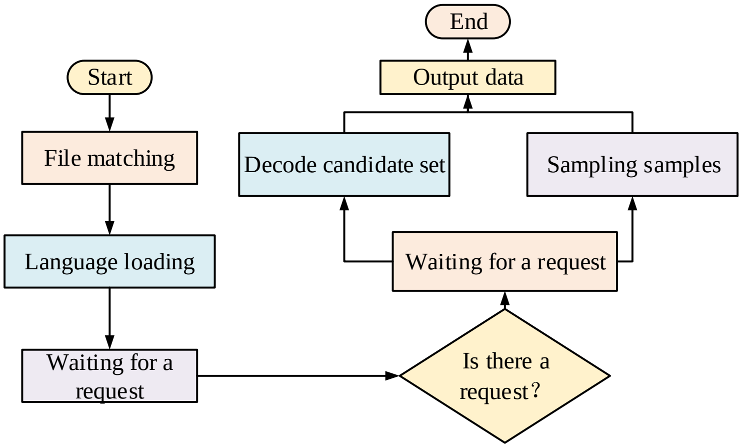


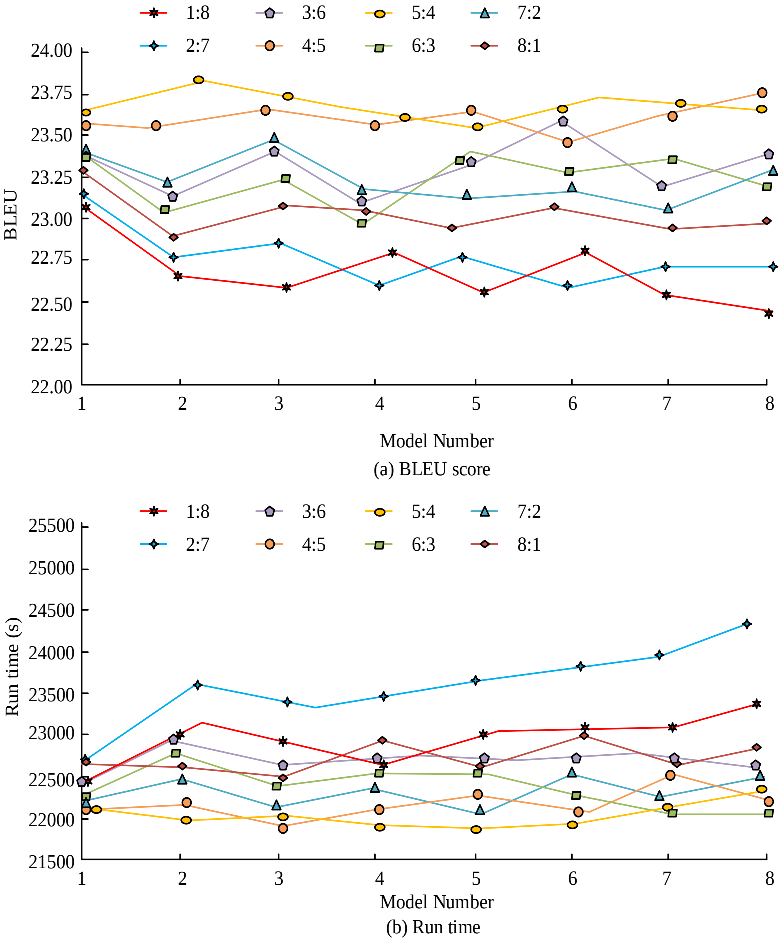
Fig. 7. Multi scene translation decoding model process.

IV. Application Analysis of English Language Scenarios Based on Machine Translation Language Modeling

This section's primary goal is to validate the advanced nature of the model utilized in the current study by conducting an experimental analysis of the model and comparing and analyzing the impact of model translation. Second, in order to confirm the model application effect even more, experiments are conducted to confirm the translation effect of the model currently in use in various circumstances.

A. *Machine Translation Model Effect Analysis*

The study uses the WMT-20 English dataset published by MT Association for data processing. The size of the test set in this dataset is 7.9k and the size of the training set is 3.0 M. The processor used for the simulation model of the data used is Intel i9-10920X, the operating system is selected as Ubuntu 20.04, the size of the RAM is 64GB, the graphics card is selected as RTX3090, and the dimensionality of the word embedding is set to 512, and the dimensionality of the hidden layer is set to 2048. The attention and convolutional layers were analyzed in different ratios, and several different ratios were experimentally analyzed to find the best ratio. Eight different model comparison ratios are selected for the study. For example, the convolutional layer comparison attention module is 1:8, 2:7, 3:6, 4:5, 5:4, 6:3, 7:2, and 8:1. These eight attention ratios are compared to get the model ratio training parameter data as shown in Table I. The evaluation index is evaluated with the score of bilingual evaluation understudy (BLEU), which is the evaluation index of MT results, and its score is able to evaluate and analyze the MT model. Comparison of different proportions of the model effect is obtained as shown in Fig. 8.



(a) BLEU score

(b) Run time

Fig. 8. Comparison of running speed and BLEU values of models with different scales.

In Fig. 8(a), in several model ratios when the number of convolutional and attentional layers are closer to each other, the BLEU value of model performance is larger, which indicates that when the number of layers of convolutional and attentional layers are closer to each other the model performance results are better, so that the highest BLEU mean value of the model with the ratio of 4:5 and 5:4, respectively, is 23.48 and 23.67. Comparing the lowest ratio model 1: 8 mean value of 22.67, it is higher by 0.81 and 1.00 respectively. In Fig. 8(b), in the running time comparison, the closer the ratio is to its performance, the shorter the running time is, the shortest running time is 5:4 ratio at this time the average running time is 22065s, the highest running time is 2:7 running time average is 23587s, the difference between the two ratios of the model running time is 1522s. It can be seen that the closer the ratio of the two modules is to each other, the better the model's effect is. This might be because the attention model runs faster in terms of parameter values and gives the model a higher dimensionality, which translates into a shorter running time. To examine how the number of attention layers affects the model's data processing effect, a variety of distinct attention layers were examined and tested, yielding the comparison parameters displayed in Table I

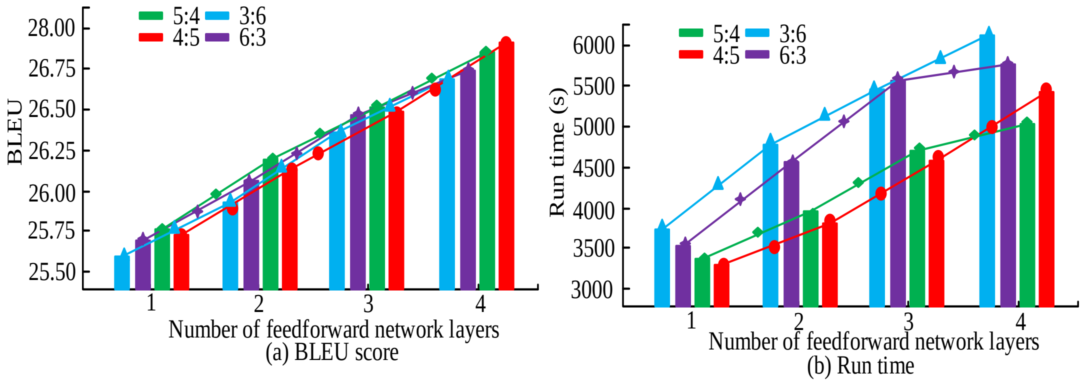
The model's parameter values, BLEU values, and running time all grow as the number of attention layers does in Table I, suggesting that while adding more attention layers can improve the model's MT index, doing so also lengthens the model's computation time. This suggests that increasing the attention model to boost the model's efficiency is not feasible after the number of layers is chosen and at the same time, the running time and the running efficiency of the model need to be taken into account, so as seen in the figure, the relatively better number of attention layers is 3, and the running time of this layer is 22548s, which has a relatively shorter running time, and at the same time, the BLEU value is at a higher value. To analyze the effect of feedforward network on the model effect, choose the above model in which the attention layer and convolutional layer are close to each other, 3:6, 6:3, 4:5 and 5:4 these four models are compared and tested, and the different feedforward network layers are analyzed to get as shown in Fig. 9.

The BLEU value of the model in Fig. 9(a) increases as the feedforward layers increase, but the four models' BLEU values change relatively little from one another. The model with the ratio of 5:4 has a larger overall change in BLEU value, but the model as a whole joins a feedforward network and its BLEU value does not change significantly. This may be due to the different ratios of the models and has little effect on how many feedforward layers are added. As the feedforward layers is increased, as shown in Fig. 9(b), the model's running speed rises and reaches its maximum when there are four feedforward layers. It can be seen that in the selection of feedforward layers, using higher feedforward layers has a good MT index but at the same time brings slower running speed, so it is more appropriate to choose 3 feedforward layers in the selection of feedforward layers. The parameter data, which shows the number of parameters used in the model during operation, will be compared with the data as shown in Fig. 10 to test the model's current use after adding various feedforward layers and pay attention to the model's effect. The larger the parameter data, the more parameters the model requires during operation, and the better the model's overall performance.

In Fig. 10(a), the parameter data of the model increases as the layers used in the model increases in several models, and the research uses the highest amount of model parameter data, which indicates that the research uses the model with the best results in model processing. The lowest value of the model parameter data for the FNN-only model indicates that the FNN-only model does not enhance the data processing of the model well. In Fig. 10(b), the model increases the running time with the increase of the model layers, and the running time of the research use model has relatively less time in the comparison of the five models, with an average time of 2854 s. The FNN-only model has the shortest running time, which might be because it can significantly cut down on the amount of time it spends using data parameters. In Fig. 10(c), the model used by the study has the highest BLEU value with a mean value at 29.65 and the model using only AM has the lowest BLEU value with a mean value of 21.84. The difference between the mean values of the two models is 7.81. In order to compare the effectiveness of the different algorithmic models with the method used by the study, the traditional statistical machine translation (SMT), rule-based machine translation (RBMT) and neural machine translation (NMT) are compared with the research use model obtained as shown in Table II.

TABLE I. COMPARISON OF DATA WITH DIFFERENT ATTENTION LEVELS

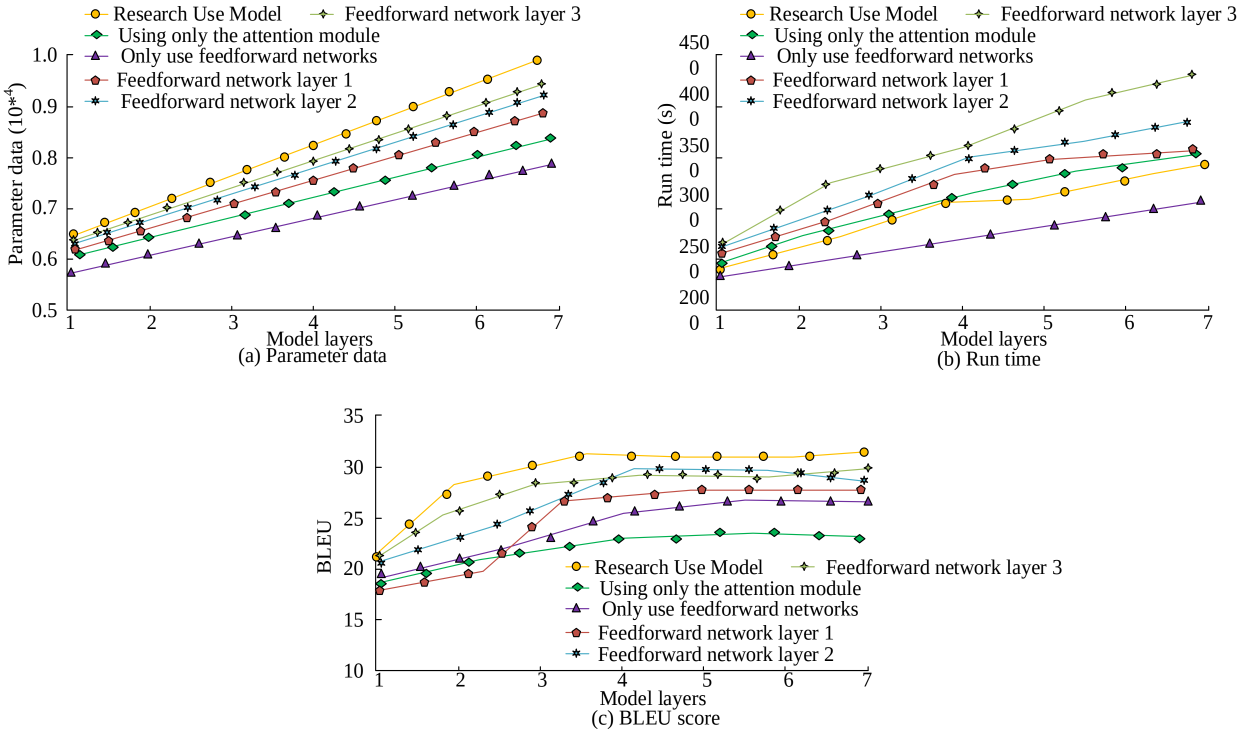
Model	Parameter values	BLEU	Run time (s)
Attention level 1	48756.00	20.65	19587.00
Attention level 2	50325.00	22.36	20549.00
Attention level 3	52364.00	23.54	22548.00
Attention level 4	55368.00	25.41	24596.00
Attention level 5	58642.00	27.54	25368.00



(a) BLEU score

(b) Run time

Fig. 9. Comparison of data from different feedforward layers.



(a) Parameter data

(b) Run time

(c) BLEU score

Fig. 10. Comparison of data from different network models.

TABLE II. COMPARISON OF DATA FROM DIFFERENT MODELS

Data set	Model	Parameter data	BLEU	Run time (s)
Dataset 1	SMT	3654.00	22.58	2456.00
	RBMT	3685.00	22.98	2635.00
	NMT	4210.00	24.59	2678.00
	Transformer	4326.00	24.76	2754.00
	Research Use Model	4586.00	26.54	2448.00
Dataset 2	SMT	4256.00	21.65	2546.00
	RBMT	4365.00	22.68	2635.00
	NMT	4686.00	23.54	2485.00
	Transformer	4758.00	24.85	2635.00
	Research Use Model	4962.00	25.67	2483.00

In Table II, in dataset 1, the model used in the study has the largest parameter data size of 4586, which is 932 higher compared to the lowest SMT model of 3654. The BLEU value of the model also has the highest value of 26.54, which is 3.96 higher compared to the lowest SMT model. In the run time comparison, the model used in the study exhibits a lower runtime data compared to several models, as long as 2448s. This suggests that the study's model outperforms other models in terms of overall performance. Additionally, comparing the data in dataset 2, the model employed in the study has parameter data that is 706 times greater than the SMT model, a BLEU value that is 4.02 times higher, and the model with the shortest run time, 2483s, than the SMT model. This shows that the model used in the study has a better performance than the traditional models.

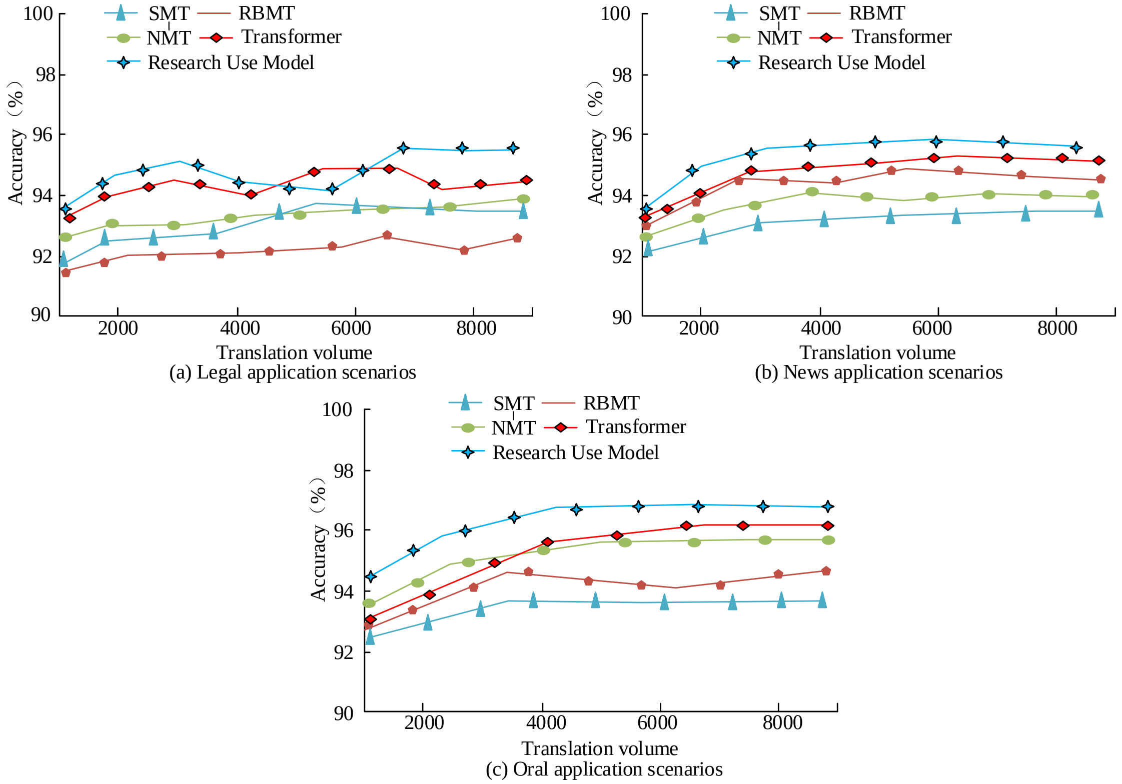
B. *Effectiveness of Machine Translation Modeling in Practice*

To test the effectiveness of the current research using the model in the application of English translation in multiple scenarios, English data from different scenarios are used, and data from three different scenarios such as legal application scenarios, news application scenarios, and speaking application scenarios are selected. 20,000 of these data are sampled and analyzed. The translation recognition accuracies of several traditional models in different scenarios in the previous subsection are compared and obtained as shown in Fig. 11.

In Fig. 11(a), the change in accuracy of several models used by the research in the legal application scenario increases gradually with the increase in the amount of data, after which it tends to a relatively stable state. The accuracy rate of the research-used model is higher among the five models, with the highest value of 95.6%, which is 3.1% higher relative to the smallest model, RBMT, at 92.5%. In Fig. 11(b), the accuracy of the models varies similarly to Fig. 11(a), with the research use model having a higher accuracy relative to the other models. The accuracy of the research use model is 95.7% higher than the accuracy of the SMT model 93.1% by about 2.6%. Similarly in Fig. 11(c) the accuracy of the research use model 96.5% is higher than the accuracy of the SMT model 93.2% about 3.3%. This shows that the research used model has higher accuracy and better modeling. Many scenarios are modeled and analyzed to produce the model test results displayed in Table III in order to assess the application effect of the research usage model in various English settings.

From Table III, in several models used, the study uses models in different scenarios with better MT effect higher indicators. Among them, the highest indicator value in the scenario of engineering analysis has 27.95, which is 6.90 higher compared to the lowest SMT model. Looking at different scenarios, the MT effect of the model in different scenarios varies, and some scenarios show higher indicator values, which may be due to the fact that the model in that scenario is more suitable for MT. To examine the MT effect of the currently used model in different scenarios, the translation indexes of several scenarios were compared as shown in Fig. 12.

In Fig. 12, the model's BLEU metrics changes in different scenarios all increase with the number of scenario samples, which may be due to the fact that more data samples can improve the model's translation efficiency during the training process. However, the news scene has the lowest metric value among several scene models, which may be due to the fact that the news scene contains more sentences and words about the emotional expression of the scene, which is more challenging for the MT. The model used in the study is able to translate ELs from different scenes and at the same time can achieve better translation results.



(a) Legal application scenarios

(b) News application scenarios

(c) Oral application scenarios

Fig. 11. Comparison of application accuracy in different model scenarios.

TABLE III. COMPARISON OF MACHINE TRANSLATION METRICS FOR DIFFERENT SCENARIO MODELS

Scene	SMT	RBMT	NMT	Transformer	Research Use Model
Law	21.65	22.65	24.65	24.68	26.48
News	22.54	23.54	24.84	24.68	27.65
Spoken language	21.85	23.48	25.03	25.36	27.30
Organism	22.64	22.68	25.16	24.79	26.54
Analysis	21.05	23.75	24.26	25.67	27.95

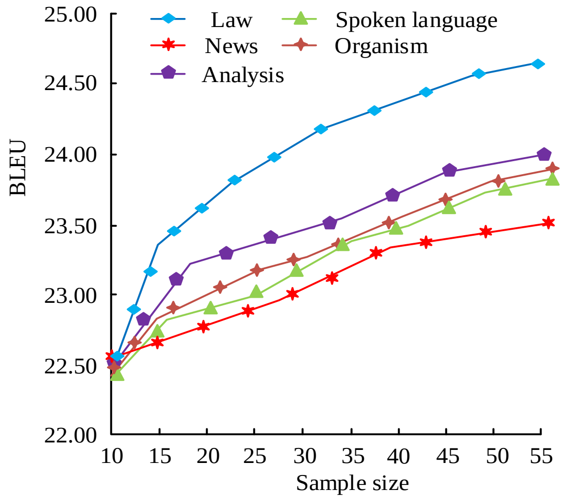


Fig. 12. Machine translation metrics for different scenarios of the model.

V. Discussion

In the bilingual evaluation scoring test, the actual running effect of the module is best when the ratio of convolutional and attention layers is 4:5 and 5:4, which may be due to the reason that adding more convolutional modules and attention layers can improve the running effect of the model. In the analysis of different attention layer parameters, the more attention layers the better the model runs, which may be due to the fact that adding more attention layers can improve the data processing effect and running efficiency of the model. However, the increase of attention layers also causes the model running speed to decrease, which indicates that the addition of attention layers needs to guarantee the running speed of the model. In the comparison of the different feedforward layers of the model, the higher the number of feedforward layers, the lower the running speed, but at the same time, the higher the running score of the model. This may be due to the fact that the addition of feedforward layers improves the model running effect but reduces the model running speed. Therefore, the study needs to select the appropriate run feedforward layer. In the comparison of the number of layers used in the model, the more the number of layers used in the model the larger the data parameters of the model, which may be due to the fact that the increase in the number of layers used enhances the loading of the model with the used parameters. Also using only a single network model is not effective in enhancing the efficiency of the model, this may be due to the reason that a single module is not effective in enhancing the model. The research use model has the highest BLEU value of 29.65, which indicates the reason that the research use model enhances the model effectiveness with the addition of different modules. In the comparison of the models of different methods, the research use model has the largest data parameter content of 4586, the highest BLEU value of 26.54, and the shortest running time of only 2448 s. This may be due to the fact that the research use model adds more modules to enhance the model running efficiency and parameter loading. This is similar to the results obtained by Shao M et al. [20]. In the practical application, the running time of the research use model can reach up to 95.6%, which may be due to the fact that the research use model adds the attention mechanism to improve the accuracy of the model running. In the application of different scenarios, the research use model was able to run in different scenarios and the model's run-up scores were all at high values, which may be due to the addition of the meta-learning module.

In summary, the research use model showed better model capabilities in model run effectiveness, run time and translation scores. This indicates that the model used in the study has a good prospect for practical application in English translation in multiple scenarios. However, the research can also consider integrating multimodal information, such as pictures and videos, in the model to combine visual and textual information to improve the accuracy of translation and semantic understanding. Meanwhile, on the basis of English multi-scene translation, the model's multi-language translation ability can be further investigated. And the existing models may have the problem of context incoherence when dealing with long text translation. The introduction of a global context mechanism can be explored to enhance the coherence and accuracy of the model in long text translation, thus improving the user experience. Therefore the use of the model in English translation may have good practical application ability and can provide better research value for English translation in multiple scenarios. It also has a better guiding significance for English translation and English learning in multiple scenes.

VI. Conclusion

The study suggests a new model based on MT and primarily addressed the issue of multi-scene application in English translation today. To enhance the translation effect of the model, the new model used BERT and CRF to extract words from the English scene's starting words. The results of the study showed that the 4:5 and 5:4 models had the highest BLEU mean values of 23.48 and 23.67, respectively. This was 0.81 and 1.00 higher than the lowest model, the 1:8 mean value of 22.67. The shortest run time was the 5:4 ratio at this point in time with a run time mean value of 22065s and the highest run time was the 2:7 run time mean value of 23587s. The number of attention layers where the model works better was 3. The run time for this layer was 22548s and the BLEU value was at a higher value. The highest mean BLEU value for the model used in the study was 29.65. The largest parameter data size for the model used in the study was 4586, which was 932 higher compared to the lowest SMT model 3654. The BLEU value for the model used in the study was also the highest at 26.54, which was 3.96 higher compared to the lowest SMT model. The research use model had a higher accuracy among the five models, with the highest value of 95.6%, which is 3.1% higher compared to the smallest model RBMT with 92.5%. It can be concluded that the model used in the study is able to translate EL for different scenarios very well, and its performance is also better than the traditional MT model. Although the study has achieved a lot of results, there are still many problems, such as the data used in the study is relatively small, more and larger data will be analyzed in the future, and different decoders will be added in the subsequent study to achieve further improvement of the model. The decoder may affect the broad applicability of the system in diverse scenarios when the system is dealing with more complex syntactic structures or rare words. Therefore subsequent research will explore the applicability of the system. Finally the research system may also suffer from computational complexity leading to high deployment and maintenance costs of the system in certain application scenarios. Therefore subsequent research will further reduce the system operating costs.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

RESEARCH INVOLVING HUMAN PARTICIPANTS AND / OR ANIMALS

Not applicable.

INFORMED CONSENT

The author agreed that the paper could be published.

REFERENCES

[1] Bacca-Acosta J., Fabregat R., Baldiris S., Kinshuk. "Determinants of student performance with mobile-based assessment systems for English as a foreign language courses," J. Comput. Assist. Learn., vol. 38, no. 3, pp. 797-810, 2022, DOI: 10.1111/jcal.12783.

[2] Chen J., Moradi H., Widodo H. P., Wood A., Gupta D. "ASIAN ENGLISH LANGUAGE CLASSROOMS: WHERE THEORY AND PRACTICE MEET," Appl. Linguist., vol. 44, no. 3, pp. 592-595, 2020, DOI: 10.1093/applin/amaa040.

[3] Gonzalez F. A., Elgeti S., Behr M., Auricchio F. "A deforming-mesh finite-element approach applied to the large-translation and free-surface scenario of fused deposition modeling," Int. J. Numer. Methods Fluids, vol. 95, no. 2, pp. 334-351, 2023, DOI: 10.1002/fld.5151.

[4] P. Y. Hao, "Asymmetric Possibility and Necessity Regression by Twin Support Vector Networks," IEEE Transactions on Fuzzy Systems, vol. 29, no. 10, pp. 3028-3042, 2020, DOI: 10.1109/TFUZZ.2020.3011756.

[5] Maruf S., Saleh F., Haffari G. "A Survey on Document-level Neural Machine Translation: Methods and Evaluation," ACM Comput. Surv., vol. 54, no. 2, Article 36, 2021, DOI: 10.1145/3441691.

[6] Yang M., Liu S., Chen K., Zhang H., Zhao E. "A Hierarchical Clustering Approach to Fuzzy Semantic Representation of Rare Words in Neural Machine Translation," IEEE Trans. Fuzzy Syst., vol. 28, no. 5, pp. 992-1002, 2020, DOI: 10.1109/TFUZZ.2020.2969399.

[7] Dabre R., Chu C., Kunchukuttan A. "A Survey of Multilingual Neural Machine Translation," ACM Comput. Surv., vol. 53, no. 5, Article 38, 2020, DOI: 10.1145/3406095.

[8] Khaw L. L., Tan W. W. "Creating Contexts in Engineering Research Writing Using a Problem-Solution-Based Writing Model: Experience of Ph.D. Students," IEEE Trans. Prof. Commun., vol. 63, no. 2, pp. 155-171, 2020, DOI: 10.1109/TPC.2020.2988758.

[9] Ban H., Ning J. "Design of English Automatic Translation System Based on Machine Intelligent Translation and Secure Internet of Things," Mobile Inf. Syst., vol. 2021, no. 4, Article 9, 2021, DOI: 10.1155/2021/8670739.

[10] Bharti N., Joshi N., Mathur I., Katyayan P. "Quality-Based Ranking of Translation Outputs," IT Prof., vol. 22, no. 4, pp. 21-27, 2020, DOI: 10.1109/MITP.2020.2976009.

[11] Wu H. "Multimedia Interaction-Based Computer-Aided Translation Technology in Applied English Teaching," Mobile Inf. Syst., vol. 2021, no. 2, pp. 1-10, 2021, DOI: 10.1155/2021/5578476.

[12] Maimaiti M., Liu Y., Luan H., Sun M. "Data augmentation for low-resource languages NMT guided by constrained sampling," Int. J. Intell. Syst., vol. 37, no. 1, pp. 30-51, 2021, DOI: 10.1002/int.22616.

[13] Bommidi B. S., Teeparthi K., Kosana V. "Hybrid wind speed forecasting using ICEEMDAN and transformer model with novel loss function," Energy, vol. 265, no. 15, Article 120234, 2023, DOI: 10.1016/j.energy.2020.120234.

[14] Beck M. B. D. A. C., Beck D. A. C. "Efficient 3D Molecular Design with an E(3) Invariant Transformer VAE," J. Phys. Chem. A, vol. 127, no. 37, pp. 7844-7852, 2023, DOI: 10.1021/acs.jpca.3c04188.

[15] Nourizadeh H., Mosallanejad A., Setayeshnazar M. "Optimal placement of fixed series compensation and phase shifting transformer in the multi-year generation and transmission expansion planning problem at the pool-based market for maximizing social welfare and reducing the investment costs," IET Gener. Transm. Distrib., vol. 16, no. 15, pp. 2959-2976, 2022, DOI: 10.1049/gtd2.12488.

[16] Sima W., Peng D., Yang M., Sun P. "Reversible Wideband Hybrid Model of Two-Winding Transformer Including the Core Nonlinearity and EMTP Implementation," IEEE Trans. Ind. Electron., vol. 68, no. 4, pp. 3159-3169, 2021, DOI: 10.1109/TIE.2020.2977544.

[17] Yadav S., Mehta R. K. "Modelling of magnetostrictive vibration and acoustics in converter transformer," IET Electr. Power Appl., vol. 15, no. 3, pp. 332-347, 2021, DOI: 10.1049/elp2.12025.

[18] Li Z., Jiao Z., He A. "Knowledge-based artificial neural network for power transformer protection," IET Gener. Transm. Distrib., vol. 14, no. 24, pp. 5782-5791, 2020, DOI: 10.1049/iet-gtd.2020.0542.

[19] Nicolae P. M. T., Nicolae I. D. V. D., Nitu M. C., Nicolae M.-S. P. M. "Analysis and Experiments Concerning Surges Transferred Between Power Transformer Windings Due to Lightning Impulse," IEEE Trans. Electromagn. Compat., vol. 65, no. 5, pp. 1476-1483, 2023, DOI: 10.1109/TEMC.2023.3299630.

[20] Shao M., Qiao Y., Meng D., Wangmeng Z. "Uncertainty-guided hierarchical frequency domain Transformer for image restoration," Knowl.-Based Syst., vol. 263, no. 5, Article 110306, 2023, DOI: 10.1016/j.knosys.2023.110306.

[21] Lu G., Zheng D., Zhang Q., Zhang P. "Effects of Converter Harmonic Voltages on Transformer Insulation Ageing and an Online Monitoring Method for Interlayer Insulation," IEEE Trans. Power Electron., vol. 37, no. 3, pp. 3504-3514, 2022, DOI: 10.1109/TPEL.2021.3118020.

[22] Li H., Li C., Zheng A., Tang J. "MsKAT: Multi-Scale Knowledge-Aware Transformer for Vehicle Re-Identification," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 19557-19568, 2022, DOI: 10.1109/TITS.2022.3166463.

[23] Sun Q., Li Y., Ma D., Zhang Y. "Model Predictive Direct Power Control of Three-port Solid-State Transformer for Hybrid AC/DC Zonal Microgrid Applications," IEEE Trans. Power Deliv., vol. 37, no. 1, pp. 528-538, 2021, DOI: 10.1109/TPWRD.2021.3064418.

[24] Conway T. "An Isolated Active Balancing and Monitoring System for Lithium Ion Battery Stacks Utilizing a Single Transformer Per Cell," IEEE Trans. Power Electron., vol. 36, no. 4, pp. 3727-3734, 2021, DOI: 10.1109/TPEL.2020.3024904.

[25] Qiu H., Wang S., Sun F., Wang Z., Zhang N. "Computational Investigations on the Four-Stage MA-Class Fast Linear Transformer Driver With Sharing Cavity Shell," IEEE Trans. Plasma Sci., vol. 49, no. 9, Article 1, 2021, DOI: 10.1109/TPS.2021.3093561.

[26] Guo Z., Yu R., Xu W., Feng X. "Design and Optimization of a 200-kW Medium-Frequency Transformer for Medium-Voltage SiC PV Inverters," IEEE Trans. Power Electron., vol. 36, no. 9, pp. 10548-10560, 2021, DOI: 10.1109/TPEL.2021.3059879.

[27] Medeiros R. P., Costa F. B., Silva K. M., Muro J. de J. C. "A Clarke-Wavelet-Based Time-Domain Power Transformer Differential Protection," IEEE Trans. Power Deliv., vol. 37, no. 1, pp. 317-328, 2021, DOI: 10.1109/TPWRD.2021.3059732.

[28] Kucka J., Dujic D. "Current Limiting in Overload Conditions of an LLC-Converter-Based DC Transformer," IEEE Trans. Power Electron., vol. 36, no. 9, pp. 10660-10672, 2021, DOI: 10.1109/TPEL.2021.3060106.

[29] Paul A. K. "Structured Protection Measures for Better Use of Nanocrystalline Cores in Air-Cooled Medium-Frequency Transformer for Induction Heating," IEEE Trans. Ind. Electron., vol. 68, no. 5, pp. 3898-3905, 2021, DOI: 10.1109/TIE.2020.2984978.

[30] Usman A. M., Abdullah M. K. "An Assessment of Building Energy Consumption Characteristics Using Analytical Energy and Carbon Footprint Assessment Model," Green Low-Carbon Econ., vol. 1, no. 1, pp. 28-40, 2023, DOI: 10.47852/bonviewGLCE3202545.

Detecting Malware of Windows OS Using AI Classification for Image of Extracted Behavior Features

Kang Dongshik, Noor Aldeen Alhamedi
University of the Ryukyus, Okinawa, Japan

***Abstract*—Malware detection is crucial for protecting digital environments. Traditional methods involve static and dynamic analysis, but recent advancements leverage artificial intelligence (AI) to enhance detection accuracy. This study aims to improve malware detection by integrating dynamic malware analysis with AI-driven techniques. The primary challenge addressed is accurately classifying and detecting malware based on behavior extracted from isolated virtual machines. By analyzing 50 malware samples and 11 benign programs, we extract ten behavioral features such as process ID, CPU usage, and network connections. We employ text-based classification using feedforward neural networks (FNN) and recurrent neural networks (RNN), achieving accuracy rates of 56% and 68%, respectively. Additionally, we convert the extracted features into grayscale images for image-based classification with a convolutional neural network (CNN), resulting in a higher accuracy of 70.1%. This multi-modal approach, combining behavioral analysis with AI, not only enhances detection accuracy but also provides a comprehensive understanding of malware behavior compared to competing methods.**

***Keywords*—Malware analysis; dynamic-based analysis; image classification; malware behavior extraction; text**

I. Introduction

Recently, the number, severity, sophistication of malware attacks, and cost of malware inflicts on the world economy have been increasing exponentially. Attacks with these kinds of software have disastrous effects and cause considerable material damage to individuals, private companies, and governments' assets. Thus, malware should be detected before damaging the important assets in the company [1]. The primary motivation for this research stems from the need to enhance existing detection mechanisms to keep pace with the constantly changing threat landscape. With traditional analysis methods, we aim to significantly improve the detection and classification accuracy of malicious software. One of the key advantages of our approach is the combination of dynamic-based malware analysis with AI-driven techniques. This allows for a more comprehensive understanding of malware behavior. This hybrid approach not only improves detection rates but also enhances the ability to accurately classify and understand the nature of malware.

There are two main techniques for analyzing malware static and dynamic-based analysis. Static-based analysis examines the malware code without actually executing it. This by integrating advanced artificial intelligence (AI) techniques can provide information about suspicious functions, network activity, impacted files, etc. Dynamic-based analysis executes the malware code in an isolated environment to observe its runtime behavior. This provides insight into the full impact of the malware. A key benefit of static-based analysis is the ability to thoroughly inspect malware code using techniques like disassembly and decompilation to identify suspicious functions related to replication, propagation, payload activation, and more [2]. The static techniques help reveal overall structure, dependencies, triggers for malicious events, and obfuscation attempts. However, lacking runtime behavior, static-based analysis cannot confirm the real impact of suspected capabilities. Complex packing or encryption techniques also limit code inspection. Other hand, the dynamic-based analysis provides direct observation of malware behavior in action by executing it and monitoring the resulting activity.

Dynamic-based analysis confirms suspected functions based on static clues and captures full infection chains showing the progression and end objectives of malware according to case studies by [3]. Dynamic monitoring of memory access, network calls, system API usage, and more creates a comprehensive picture. Additionally, dynamic-based analysis is particularly effective in identifying and analyzing newly emerging malware strains. As it focuses on the runtime behavior, it is better equipped to handle polymorphic and metamorphic malware that may change its form to evade static-based analysis techniques. Leveraging AI models for the analysis of malware code or the study of malware behavior has significantly contributed to the detection of malware in recent years. Numerous AI models have been integrated into static or dynamic approaches to augment both the malware detection rate and feature extraction processes. Despite the notable progress in the field of AI, these models still face various challenges. This research will use many models of AI to detect malware.

Robust malware analysis faces numerous obstacles. The sheer volume of malware proliferating at a rapid pace presents a formidable challenge in comprehensively examining this ever-expanding threat landscape. Additionally, malware authors employ sophisticated obfuscation tactics, such as code interchange, amalgamation, register reassignment, null insertion, and subroutine reordering [3], purposefully designed to evade detection by anti-malware systems. Despite decades of development, these security solutions still exhibit high false positive rates, undermining their accuracy.

Moreover, certain malware strains possess the ability to identify virtualized environments, resulting in altered or ceased execution, hindering effective analysis. The evasion techniques employed by malware necessitate lengthy detection times, potentially ranging from minutes to hours depending on the specific malware variant, during which systems remain vulnerable to compromise. Furthermore, the ambiguity surrounding API calls, as both malicious and benign software may legitimately invoke common APIs, complicates the process of distinguishing malware based on API usage patterns.

These factors, including the immense scale, obfuscation methods, virtual environment detection capabilities, delayed identification timelines, and the dual usage of APIs, collectively contribute to the arduous nature of robust malware analysis, necessitating the development of advanced techniques to overcome these challenges effectively. The juxtaposition of text classification and image classification in the analysis of extracted behavior. It underscores that a nuanced understanding of program nature, distinguishing between benign and malicious entities, can be achieved through thorough behavior analysis. The model primarily relies on the extraction of malware features. Within the developed script, two distinct observers play a crucial role. The first observer extracts the entirety of the process, encompassing its characteristics, as well as details related to internet connections. The second observer is tasked with monitoring any file creation specifically linked to the malware. The experimental framework involves the extraction of 10 distinct features through the monitoring of behaviors within an isolated Virtual Machine. Python libraries such as psutil, subprocess, wmi, watchdog, time, json, and os were employed to develop functions responsible for observing malware behavior and subsequently extracting pertinent information to a JSON file. The extracted features encompassed critical aspects such as process ID, process name, username, CPU percentage.

II. Related Work

Artificial Intelligence (AI) has emerged as a powerful tool in this ongoing struggle to detect and classify malware offering advanced capabilities in identifying and mitigating malware threats.

In a study [4], the third paper analyzes different classical machine learning algorithms for malware detection - Random Forest, Support Vector Machine (SVM), grid search optimized SVM, and K-Nearest Neighbors (KNN). The goal is to validate the effectiveness of these models for detecting zero-day malware attacks. The dataset from Kaggle contained 19,611 PE files, with 14,599 malicious samples and 5,012 benign files with 77 numeric features. Three training/test splits were used. Various accuracy metrics were calculated: accuracy, F1-score, confusion matrix, precision, recall and Type I/II errors. Random Forest performed the best with 96% accuracy and 93% F1score, with low errors and fastest training time. Optimized SVM improved results significantly but slowed down execution. KNN also performed decently with simpler implementation. Analysis showed Random Forest has good prospects for realtime zero-day malware detection. The model can process 25,000 files per second. For deployment, more diverse input data covering different malware families is needed.

In study [5], the authors used convolutional neural networks (CNNs) for malware classification by visualizing malware programs as grayscale images. The images are generated from the bytecode of malware programs and classified using CNN architectures. They evaluate several well-known CNN models like AlexNet, ResNet, and VGG16 using transfer learning on a malware image dataset. They also propose a custom shallow CNN architecture that achieves 96% accuracy, but is faster to train than the other complex models. The customized CNN and transfer learning models are also tested as feature extractors, with the features fed into SVM and KNN classifiers. This achieves even better performance up to 99.4% accuracy. They set a new benchmark on the public BIG 2015 malware dataset. The proposed system combining CNN feature extraction + SVM classifier obtains state-of-the-art 99.4% accuracy in distinguishing between nine malware classes. Visualization and CNN-based classification is shown to be effective for malware detection. The approach is computationally efficient compared to static/dynamic-based analysis. Fusing different CNN model predictions can further improve performance.

In study [6], the authors used Support Vector Machines (SVMs) for malware analysis and classification. SVMs are supervised learning models that can analyze high-dimensional, sparse data and recognize patterns. The authors collect a heterogeneous malware dataset from a real threat database. The data has features like time, format, domain, and IP address. They visualize the dataset using techniques like scatter plots and radius visualization to understand correlations and structure before classification. An SVM model with a polynomial kernel is trained on the dataset to classify malware vs normal software. The model is validated using cross-validation, leave-one-out and random sampling. The SVM classifier achieves 93-95% accuracy, 97-98% sensitivity and 86-90% specificity on the malware dataset. Validation shows the model generalizes very well. The high-performance highlights that SVMs can effectively classify heterogeneous malware data gathered from computer networks and security systems.

In study [7], the paper proposes a deep learning framework for malware visualization and classification using convolutional neural networks (CNNs). The key aspects are: Malware files are converted into three image types - grayscale, RGB color, and Markov images. Markov images help retain global statistics of malware bytes. A Gabor filter approach is used to extract textures and discriminative features from the malware images. Two CNN models are used for classification – a custom 13-layer CNN and a pretrained 71-layer Xception CNN fine-tuned for malware images. The framework is evaluated on two public Windows malware image data sets, a custom Windows malware dataset, and a custom IoT malware dataset. Markov images provide the best results, with the fine-tuned Xception CNN achieving over 99% accuracy on multiple datasets. The computational efficiency is also better compared to prior works. The approach demonstrates effectiveness for real-time malware recognition and classification. The visualization and deep learning framework extracts features automatically without extensive feature engineering.

The framework's resilience against adversarial attacks is also analyzed by adding noise to test images. Some drop in accuracy is noticed, indicating scope for improvement. The current landscape underscores the significance of AI models as powerful tools for the analysis, classification, and detection of malware. These models can seamlessly integrate with both static and dynamic-based analysis, yielding noteworthy results that underscore their pivotal role in shaping the future of this field.

Arabo et al. [8] analyzed CPU and RAM usage patterns as potential indicators for detecting ransomware processes. Their findings suggested that while not the primary factors, monitoring CPU and RAM could complement other behavioral characteristics in identifying malicious processes. Regarding CPU usage, they observed variations that showed potential for distinguishing ransomware activities. Specifically, for the ViraLock ransomware sample, the maximum CPU usage peaked at 25% [1]. Such CPU spikes could potentially signify the initiation of encryption or other malicious operations by the ransomware. As for RAM consumption, the study found that ransomware samples generally exhibited low and relatively stable memory usage patterns. In the case of ViraLock, the maximum RAM usage was only around 2% [1]. However, the authors noted that while low RAM usage alone may not be a definitive indicator, it could be considered in combination with other behavioral factors. The researchers highlighted that while CPU and RAM usage showed some differences between ransomware and benign processes, the most significant distinguishing factor was abnormally high disk read/write activity [1]. Nonetheless, incorporating CPU and RAM monitoring alongside disk usage analysis could potentially enhance the accuracy and robustness of ransomware detection systems based on process behavior analysis.

III. Methodology

The current investigation is centered on the behavioral analysis within an isolated Windows environment in virtual machine for the purpose of detecting malware. To achieve this, a combination of Recurrent Neural Network (RNN) for text classification and Convolutional Neural Network (CNN) for image classification is employed to analyze the extracted data. Diverging from the methodologies outlined in previous studies [3], [6], and [7], the classification approach adopted here focuses on the inherent characteristics of the malware file itself. This is achieved through a comprehensive analysis of the malware binary file and, notably, by representing the malware file as an image utilizing various visualization techniques. In this research, the emphasis is on visualizing the malware's behavior and, subsequently, conducting analyses based on these visual representations and also analysis the extracted features as a text. The presented model offers a juxtaposition of text classification and image classification in the analysis of extracted behavior. It underscores that a nuanced understanding of program nature, distinguishing between benign and malicious entities, can be achieved through thorough behavior analysis.

The model primarily relies on the extraction of malware features. Within the developed script, two distinct observers play a crucial role. The first observer extracts the entirety of the process, encompassing its characteristics, as well as details related to internet connections. The second observer is tasked with monitoring any file creation specifically linked to the malware.

The experimental framework involves the extraction of 10 distinct features through the monitoring of behaviors within an isolated Virtual Machine. Python libraries such as psutil, subprocess, wmi, watchdog, time, json, and os were employed to develop functions responsible for observing malware behavior and subsequently extracting pertinent information to a JSON file. The extracted features encompassed critical aspects such as process ID, process name, username, CPU percentage.

The modules for this research were developed using TensorFlow and Keras, leveraging the Sequential model architecture. These tools enabled efficient construction and training of neural networks for malware detection, facilitating both text-based and image-based classification with enhanced accuracy through deep learning techniques. Fig. 1 shows proposed processing model.

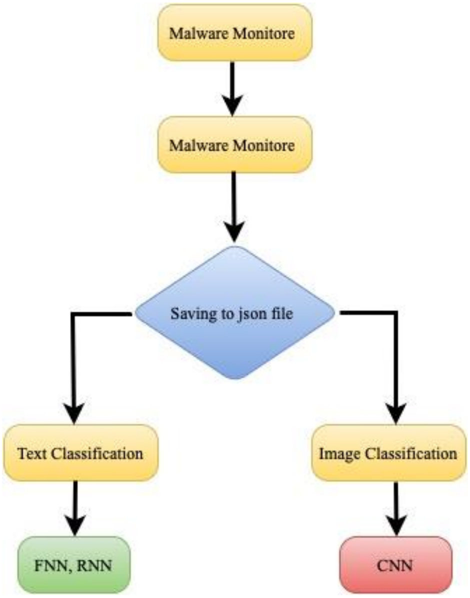
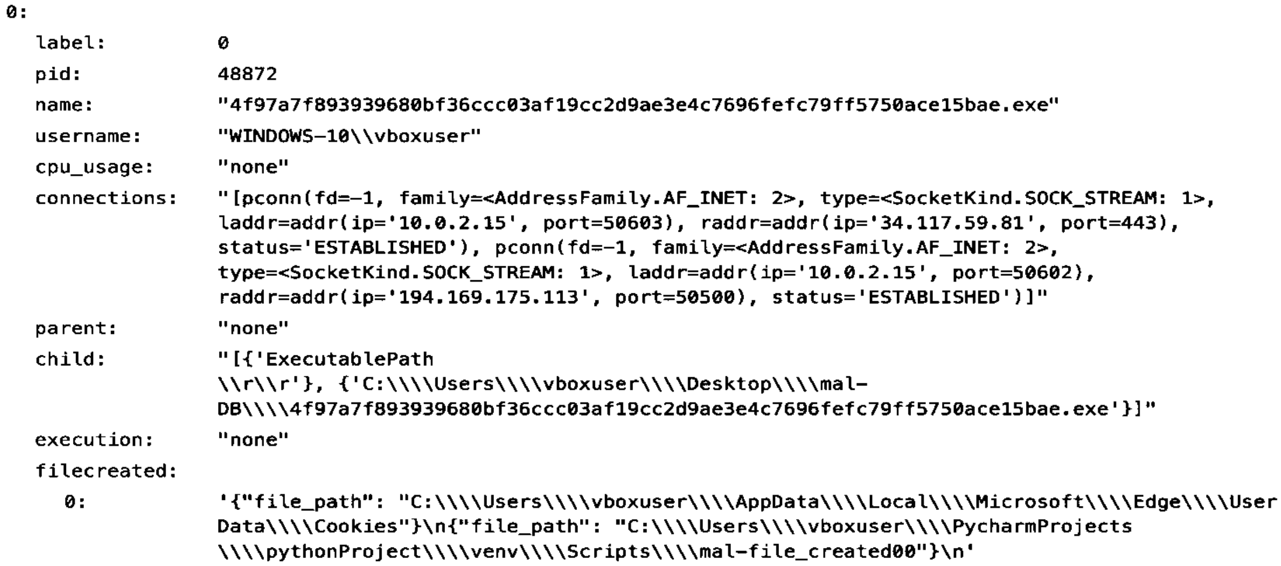


Fig. 1. Proposed processing.

Following the extraction of these features, the gathered information is stored in a JSON file (see Fig. 2) for further next step.

A. Text Analysis

The analytical process for the extracted features unfolded across two phases. Initially, the data underwent textual analysis, leveraging a simple feedforward neural network (FNN) model designed for binary classification using the Keras library to create a fully connected dense layer with 128 nodes. The output layer has 1 node and uses 'sigmoid' activation for binary classification. Subsequently, a recurrent neural network (RNN) model was employed to classify the same textual data, creates an embedding layer that transforms integer word indices to dense word vector representations.



```
0:
  label:          0
  pid:            48872
  name:           "4f97a7f893939680bf36ccc03af19cc2d9ae3e4c7696fefc79ff5750ace15bae.exe"
  username:       "WINDOWS-10\\vboxuser"
  cpu_usage:      "none"
  connections:    "[pconn(fd=-1, family=<AddressFamily.AF_INET: 2>, type=<SocketKind.SOCK_STREAM: 1>,
                  laddr=addr(ip='10.0.2.15', port=50603), raddr=addr(ip='34.117.59.81', port=443),
                  status='ESTABLISHED'), pconn(fd=-1, family=<AddressFamily.AF_INET: 2>,
                  type=<SocketKind.SOCK_STREAM: 1>, laddr=addr(ip='10.0.2.15', port=50602),
                  raddr=addr(ip='194.169.175.113', port=50500), status='ESTABLISHED')]"
  parent:         "none"
  child:          "[{'ExecutablePath
                  \\r\\r'}, {'C:\\\\Users\\\\vboxuser\\\\Desktop\\\\mal-
                  DB\\\\4f97a7f893939680bf36ccc03af19cc2d9ae3e4c7696fefc79ff5750ace15bae.exe'}]"
  execution:      "none"
  filecreated:
    0:            '{"file_path": "C:\\\\Users\\\\vboxuser\\\\AppData\\\\Local\\\\Microsoft\\\\Edge\\\\User
                  Data\\\\Cookies"}\n{"file_path": "C:\\\\Users\\\\vboxuser\\\\PycharmProjects
                  \\\\pythonProject\\\\venv\\\\Scripts\\\\mal-file_created00"}\n'
```

Fig. 2. Sample of Json file content connection details, parent process, child process, execution path, and created files.

B. *Image Analysis*

By transforming data into images, researchers can leverage the vast body of knowledge and advancements in image processing techniques, readily applicable to the analysis of the transformed data. This data-to-image transformation unlocks the power of CNNs for a wider range of analysis tasks, promoting deeper insights into complex datasets. So this research implements the power of CNN alongside with the behavior analysis Subsequent to the behavioral analysis, the extracted features underwent further evaluation through an image classification paradigm. A dedicated function was developed to transform these feature data into grayscale images. This transformative process involved the removal of associated labels, conversion of the data into binary numerical representations, subsequent transformation of these binary values into hexadecimal equivalents, and, finally, depiction of these hexadecimal values onto a 30*30 grayscale canvas.

The 30x30 size was empirically determined to balance information preservation and computational efficiency. Representing features as images enabled the utilization of convolutional neural networks (CNNs), which excel at capturing spatial patterns the extracted features underwent further evaluation through an image classification paradigm. This visual representation approach offered several key advantages. Firstly, it enabled leveraging powerful deep learning techniques like convolutional neural networks, adept at capturing spatial patterns invaluable for malware characterization. Secondly, transforming features into images facilitated uncovering intrinsic relationships and patterns obfuscated in the original data's raw representation. Thirdly, the image domain allowed seamless integration of transfer learning and pre-trained models, expediting the analysis process. Lastly, the visually interpretable nature of images could provide insights into the discriminative characteristics learned by the models, aiding explain ability. By combining dynamic monitoring with visual analytics, this multi-pronged approach offered a potent framework for comprehensive malware analysis and classification.

The dataset employed for experimentation comprised 50 instances of .EXE malware sourced from diverse families, obtained from the Malware Bazaar database, a freely accessible online repository. Additionally, 11 benign programs were included for comparative analysis. The monitoring process lasted three seconds for every malware instance, during which the monitoring code ran in the background, observing the processes and file creation activities of the malware. After the monitoring period, the code produced a JSON file containing the captured information. The dataset has been divided into 40 malware behavior and six benign program behavior for the training and 10 malware behavior and five benign program behavior for testing. Fig. 3 shows converting text to image process.

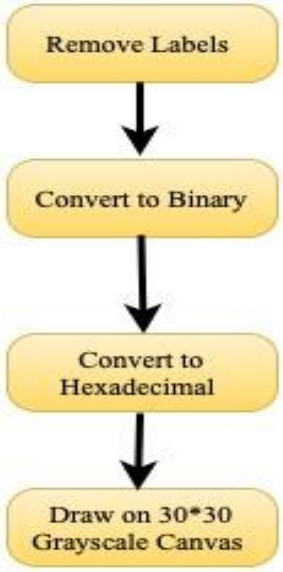


Fig. 3. Converting text to image.

IV. Experiments

A. *Text Analysis*

The described FNN model exhibited an accuracy rate of 56% with a corresponding loss rate of 0.78. For the RNN model: It takes the vocabulary size equal to 32 and output dimensionality as arguments. Also LSTM layer models the sequential nature and long-range context of text. The output dense layers act as classifiers on top of LSTM representations. The model is compiled with binary cross entropy loss, adam optimizer and accuracy metric.

With epoch 100, yielding an improved accuracy rate of 68% with a reduced loss rate of 0.67.

B. *Image Analysis*

Convolutional Neural Networks (CNNs) have revolutionized image analysis due to their ability to extract intricate spatial features. However, their power can be extended to non-image data by transforming it into a suitable image representation. This approach offers several advantages: CNNs excel at automatically learning relevant features from images, circumventing the need for manual feature engineering, a time-consuming and potentially error-prone step in traditional analysis. Data transformation allows for the visualization of complex relationships between data points within the image domain. This empowers CNNs to identify subtle patterns that might be obscured in the raw data format. The experiment was done using two suggested models. The first model (Fig. 5) is simple and the second model is more complex both models are based on CNN. The simple model consists of:

- Conv2D layer: Performs 2D convolution with 32 filters and 3x3 kernel. Extracts spatial features from input image.
- MaxPool2D: Max pooling layer reduces dimensions to summarize the features detected by the convolution layer.
- Flatten: Flattens the pooled feature map into a 1D vector to prepare for fully-connected layers.
- Dense layers: Fully-connected layers that act as classifier on top of the extracted features. 64 nodes in first dense layer.

Output layer contains single node with 'sigmoid' activation for binary classification. This model takes input images of shape (30, 30, 1) indicating 30x30 grayscale images. Using this simple model over these grayscale pictures gives accuracy rate 70.1% with loss 0.67.

The second model also based on CNN with more complex architecture: The model then uses several convolutional layers (Conv2D) to extract features from the image. These layers apply filters (also called kernels) that slide across the image, detecting patterns and edges.

The first Conv2D layer has 256 filters, each of size 3x3. As the filter slides across the image, it performs element-wise multiplication between the filter weights and the corresponding pixel values in the image. The results are then summed and passed through an activation function (relu in this case) to introduce non-linearity. This process helps identify low-level features like edges, corners, and simple shapes. The subsequent Conv2D layers follow the same principle but with a different number of filters (128 and 64 in this example). These layers extract progressively more complex features based on the lower-level features detected earlier.

MaxPooling2D layers are inserted after some convolutional layers. These layers downsample the feature maps by taking the maximum value within a specific window (2x2 in this example). This helps reduce the number of parameters and computational cost while potentially capturing the most important features. Fig. 4 shows sample representation of the resultant images.

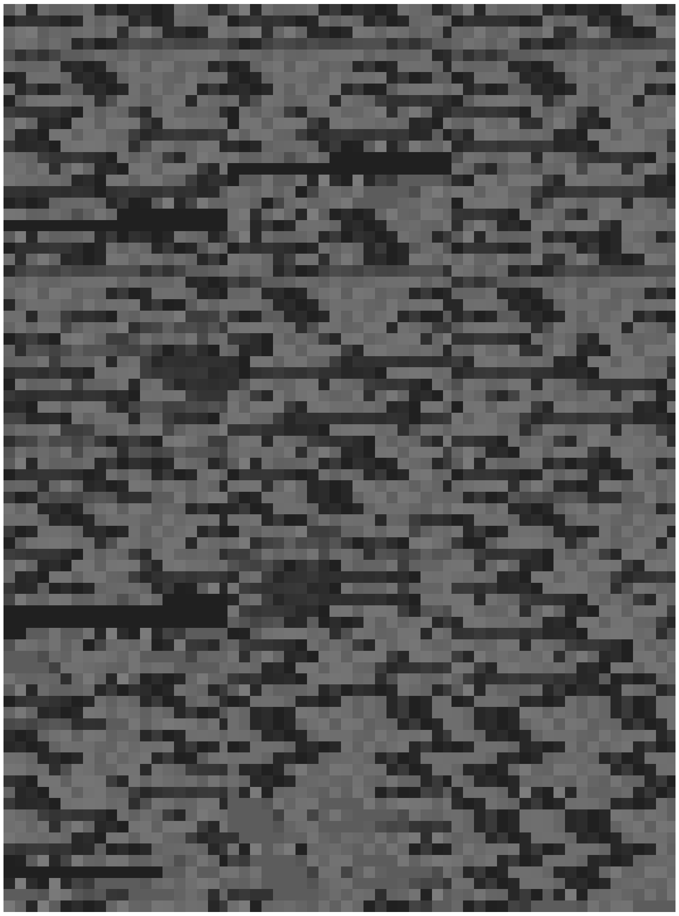


Fig. 4. A sample representation of the resultant images, offering a glimpse into their visual characteristics.

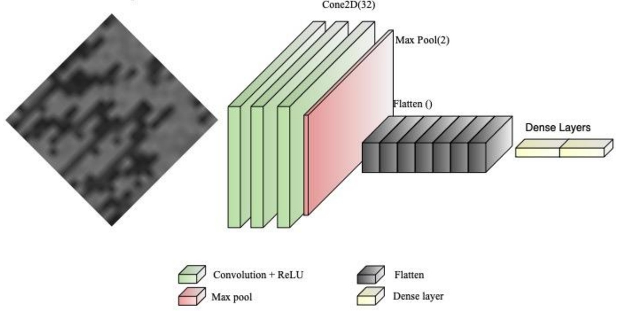


Fig. 5. The structure of the first model.

The Dropout layer (commented out) randomly drops a certain percentage (25% in this example) of activations during training. This helps prevent the model from overfitting to the training data by forcing it to learn more robust features.

After the convolutional and pooling layers, the model uses a Flatten layer to convert the 3D feature maps into a 1D vector (see Fig. 6). This allows the fully-connected layers to process the extracted features. The model then uses several fully-connected layers (Dense) to classify the image. These layers work similarly to traditional neural networks, where each neuron receives input from all neurons in the previous layer, performs weighted sums, and applies an activation function. The first three fully connected layers (4096, 2048, and 1024 neurons) are responsible for learning complex, high-level representations based on the extracted features. The relu activation allows these layers to learn non-linear relationships between the features.

The final Dense layer has only one neuron with a sigmoid activation function. This neuron outputs a value between 0 and 1, representing the probability of the image belonging to a specific class. As a summary of this model. The convolutional layers act as feature detectors, extracting progressively more complex features from the input image. The pooling layers reduce the dimensionality of the data while retaining important information. The dropout layer helps prevent overfitting. The fully-connected layers learn high-level representations and produce the final classification probability.

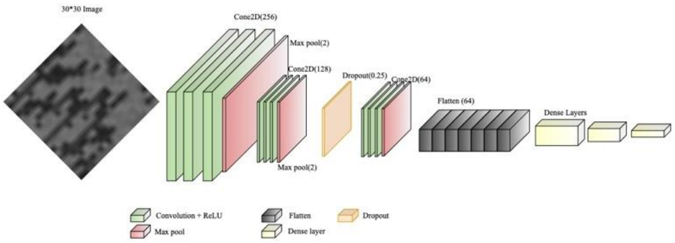


Fig. 6. The structure of the second CNN model.

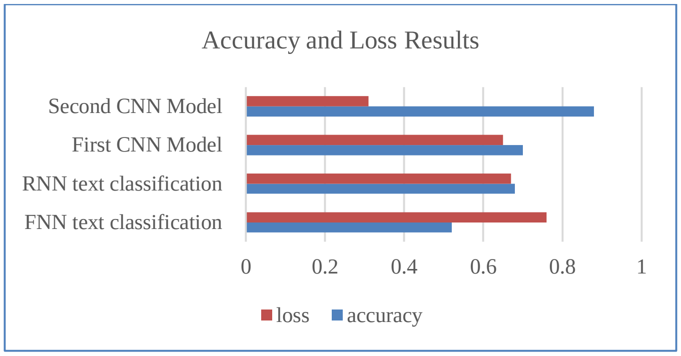


Fig. 7. Bar chart for accuracy and loss.

Using this complex model over these grayscale pictures gives accuracy rate 88% with loss 0.31. Comprehensive performance evaluation through bar charts (Fig. 7) illustrates accuracy and loss metrics for both text and image classification. The findings suggest that combining behavioral analysis with AI models, particularly in the image domain, holds promise for effective malware detection. This multimodal approach provides a holistic understanding of malware behavior, potentially enhancing overall detection capabilities in the evolving cybersecurity landscape. The study contributes to advancing malware detection methodologies by leveraging the synergy between static and dynamic analyses, bolstered by AI integration, and offers insights into the promising potential of image-based classification for improved accuracy in identifying malicious behavior.

The Second Model with numerous convolutional and fully-connected layers grants high capacity for learning intricate features. While advantageous for complex datasets, it can lead to overfitting, particularly with limited training data. The model memorizes training data too well, hindering performance on unseen examples. Furthermore, training and running this deep model can be computationally expensive due to the high number of parameters. This translates to significant processing power and memory requirements, potentially limiting its use in resource-constrained environments. The results from the text classification and image classification shows that these methods of analyzing malware might be a good way to detect the malware using the extracted behavioral features.

V. Conclusion

This study successfully employs dynamic-based analysis within a virtual machine (VM) to extract crucial behavioral features from Windows malware. Integrating these features with advanced text and image classification models (RNN and CNN) shows promise for malware detection. Image classification, based on transformed feature data, achieves a superior accuracy of 88% compared to 68% in text classification. This multi-modal approach, combining behavioral analysis with AI models, provides a nuanced understanding of malware behavior. To enhance model robustness, we recommend increasing the number of malware and benign samples, including a wider range of malware families, and exploring additional features like registry changes. Experimenting with different visualization techniques for image generation and testing more complex CNN architectures or pre-trained models with fine- tuning could further improve accuracy. Addressing adversarial attacks is crucial; incorporating noise resilience mechanisms is suggested for future work. These enhancements contribute to advancing malware detection methodologies, ensuring adaptability in the evolving cybersecurity landscape.

References

[1] Aslan, Ö., & Samet, R. (2019). A comprehensive review on malware detection approaches. IEEE Access, Advance online publication. https://doi.org/10.1109/ACCESS.2019.2963724

[2] Roundy, K.A. and Miller, B.P., 2013, August. Binary-code obfuscations in prevalent packer tools. In Proceedings of the 2013 ACM workshop on Software PROtection (pp. 3-14).M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[3] Rossow, C., Dietrich, C. J., Grier, C., Kreibich, C., Paxson, V., Pohlmann, N. & van Steen, M. (2012). Prudent practices for designing malware experiments: Status quo and outlook. In 2012 IEEE Symposium on Security and Privacy (pp. 65-79). IEEE.

[4] Nafiiev, A., Kholodulkin, H., & Rodionov, A. (2022). Comparative analysis of machine learning methods for detecting malicious files. Algorithms and Methods of Cyber Attacks Prevention and Counteraction.

[5] V. S. P. Davuluru, B. N. Narayanan and E. J. Balster, "Convolutional Neural Networks as Classification Tools and Feature Extractors for

Distinguishing Malware Programs," 2019 IEEE National Aerospace and Electronics Conference (NAECON), 2019, pp. 273-277.

[6] M. Kruczkowski and E. Niewiadomska-Szynkiewicz, "Support Vector Machine for malware analysis and classification," 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014, pp. 415-420.

[7] Sharma, O., Sharma, A., & Kalia, A. (2022). Windows and IoT malware visualization and classification with deep CNN and Xception CNN using Markov images. Journal of Intelligent Information Systems. Advance online publication.

[8] Arabo, A., Dijoux, R., Poulain, T., & Chevalier, G. (2020). Detecting Ransomware Using Process Behavior Analysis. Procedia Computer Science, 168, 289-296.

Dynamic Path Planning for Autonomous Robots in Forest Fire Scenarios Using Hybrid Deep Reinforcement Learning and Particle Swarm Optimization

N.K.Thakre[1], Divya Nimma[2], Anil V Turukmane[3], Akhilesh Kumar Singh[4], Divya Rohatgi[5], Balakrishna Bangaru[6]

Department of Humanities, YCCE, Nagpur (M.S.), India[1]

Phd in Computational Science, University of Southern Mississippi, Data Analyst in UMMC, USA[2]

Professor, School of Computer Science & Engineering, VIT- AP University, Amaravati, Vijayawada, Andhra Pradesh, India[3]

Professor, Department of Mechanical Engineering, Aditya College of Engineering & Technology, Surampalem,
Andhra Pradesh, India[4]

Associate Professor, Dept. of CSE, Bharati Vidyapeeth Deemed to be University, Department of Engineering and Technology,
Navi Mumbai, Maharashtra, India[5]

Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur Dist., Andhra Pradesh, India[6]

***Abstract*—The growing frequency of forest area fires poses critical challenges for emergency response, necessitating progressive solutions for effective navigation and direction planning in dynamic environments. This study investigates an adaptive technique to enhance the performance of autonomous robots deployed in forest area fireplace scenarios. The primary objective is to develop a hybrid methodology that integrates advanced studying strategies with optimization techniques to enhance route planning beneath unexpectedly changing situations. To reap this, a simulation-based total framework became hooked up, in which self-reliant robots were tasked with navigating diverse forest fire eventualities. The method includes schooling a model to dynamically adapt to environmental modifications at the same time as optimizing direction choice in real time. Performance metrics together with direction efficiency, adaptability to obstacles, and reaction time been analyzed to assess the effectiveness of the proposed solution. Results indicate an enormous improvement in path planning performance as compared to traditional methods, with more suitable adaptability main to faster response instances and extra effective navigation. The findings underscore the functionality of the proposed method to cope with the complexities of forest area fire environments, demonstrating its potential for real-world applications in disaster response. The results are shown in the conceived DRL-PSO framework where execution time is reduced up to 95% and the success rate of 95 % for the proposed method compared to the conventional ones. Python is used to implement the proposed work. Compared to the proposed method's execution time of 68. 3 seconds and the highest success rate among evaluated strategies, so it can be used as a powerful solution for autonomous drone navigation in dangerous situations. In the end, this research contributes precious insights into adaptive route planning for self-sufficient robots in unsafe situations, providing a strong framework for destiny advancements in disaster management technologies.**

Keywords*—Adaptive path planning; deep reinforcement learning; disaster environments; drone rescuing; particle swarm optimization; forest fire

I. INTRODUCTION

As a result of climate change and human activity, forest fires have become more severe and frequent threats posing serious threats to human safety and ecosystems These flames spread rapidly, and their unpredictable nature necessitates using state-of-the-art technological solutions for efficient monitoring and rescue efforts RL) using those tasked with rescuing people in active forest fire situations The platform offers a new approach develop a flexible system for drones by using fire detection information and real-time data to inform and optimize the autonomous drone approach planning processes [1]. Drones can operate safely and effectively in a dangerous environment thanks to fire detection data, which provides vital information about fire locations and severity spread. Path planning is important in robotics, with the aim of determining the optimal path of material movement from start to finish, which can be used in aerospace, military, manufacturing, agriculture. A subset of autonomous guidance requires dynamic decisions as a robot moves forward toward its goal. Recent advances in participatory navigation and UAV technologies highlight their high scalability, scalability and adaptability for various applications such as search and rescue [2], agriculture, and inspection. UAVs are designed to operate without human intervention, making them ideal for projects such as visual inspection of large buildings. An important approach in this area is Coverage Path Planning (CPP), which focuses on efficient, collision-free paths that cover all important paths in an area Path planning can be offline, online, or hybrid, in which online channels are important for dynamic, unfamiliar environments and where robots must continuously collect and process distance data to safely navigate Challenges such as optimal, collision avoidance Method a three-dimensional planning is adopted in UAVs often using global solutions for complex problems and local solutions for dynamic constraints visibility graphs, fast searching random trees, probabilistic routing, . algorithms There are, although reasoning methods do not always find the best methods [3].

Intelligent transportation systems (ITS) [4] are aimed at increasing road capacity, reducing accidents, improving efficiency and reducing congestion, as well as reducing energy consumption and environmental impact Vehicles as it automatically develop key components of the ITS, including environmental concepts, road design, tracking and monitoring. The key to participatory transportation planning is to identify efficient, collision-free routes from the starting point to the destination [5]. Various path-planning algorithms have emerged, including geometric, graph search, intelligent bionic, artificial potential field, and sampling-based algorithms like RRT and probabilistic roadmap, which excel in complex environments. Traditional RRT methods focus on finding paths but lack efficiency in convergence, search speed, and path optimality. Improvements like Biased RRT, Bi-RRT, RRT-connect, and RRT address these issues but often neglect vehicle-specific constraints. For dynamic environments, algorithms like potential field combinations and enhanced Bi-RRT adapt paths in real-time, accounting for dynamic obstacles. Recent advancements integrate reinforcement learning and heuristic methods to enhance RRT-based planning for smooth, collision-free paths in complex scenarios [6]. Path optimization techniques such as cubic B-splines, Dubins curves, and path pruning further refine paths, though challenges remain in maintaining curvature consistency and minimizing control difficulty. Disaster events like fires, floods, and landslides demand urgent and efficient rescue measures to minimize economic losses and threats to human life. Drones have become essential in disaster rescue, offering advantages such as 3D map reconstruction, emergency mapping, and environmental assessment. Multi-drone systems are particularly useful in complex terrains where traditional rescue methods struggle. However, effective mission planning for drones involves addressing environmental challenges and drone performance constraints. Heuristic algorithms like genetic algorithms (GA) and particle swarm optimization (PSO) are commonly used for mission planning but face issues like slow convergence and local optima. To overcome these limitations, this paper proposes improved GA and PSO algorithms for mission planning in complex 3D environments [7].

Autonomous vehicle systems, especially mobile robots, and autonomous vehicles have received considerable attention and commercialization, especially in the field of logistics and services in the service industry. The widespread use of reinforcement learning is limited by long training periods and limited computational resources Despite the advantages of DRL, such as intensive learning and large sensor dependence reduced, training in challenging environments is time-consuming and can lead to poor performance due to localized practice played a role. The high-quality research on DRL-based active obstacle avoidance and road systems from 2018 to 2022, identifying gaps and proposing future research directions to improve safety , efforts, and potential growth in this field [8]. Existing systems for route planning in autonomous robotic and mobile systems exhibit several shortcomings. Traditional RRT methods work well for challenging terrain but perform poorly in speed matching and road optimization, with enhanced versions such as Bi-RRT and RRT-Connect often ignoring vehicles specifically limited. They struggle with slow convergence and local change. Heuristic algorithms such as GA and PSO also face problems of slow convergence and local adaptability, especially in complex 3D environments. DRL methods, although powerful, have long training times and high computational resource requirements and result in poor performance due to localized action execution AMRs, although simpler than AGVs, need to be limited by movement in for safety, and limits their operating range. The proposed DRL model addresses these issues by using real-time data for informed decision-making, integrating DRL into advanced road planning processes to increase efficiency and safety, and DRL and integrating PSO to improve robustness and performance This approach ensures flexibility and resilience scenarios are appropriate, and achieves high accuracy and reliability through the traditional DRLC limitations of training time and estimation by the effortful handling.

The major key contribution are as follows:

- The study introduces a novel hybrid approach combining Deep Reinforcement Learning (DRL) with Particle Swarm Optimization (PSO), enhancing the efficiency and adaptability of independent robots in navigating complicated, dynamic catastrophe environments.
- The proposed approach permits real-time adaptation to unexpectedly changing environmental situations and obstacles normally determined in disaster zones, improving the robots' capacity to make well timed and optimal decisions
- The research contributes to multi-agent structures through demonstrating how multiple independent robots can collaborate successfully to gain common desires, which includes search-and-rescue operations, by leveraging DRL and PSO coordination abilities
- The model optimizes computational aid utilization through the integration of DRL and PSO, permitting robots to carry out path-making plans with decreased computational overhead while retaining high-performance stages.

The research validates the model by using the simulations in practical dynamic disaster environments, showing extensive enhancements in pathfinding efficiency, robustness, and undertaking of entirety in comparison to traditional techniques.

The proposed research is arranged as follows: The current models are reviewed in Section II. In Section III, the drawbacks of the existing frameworks are briefly reviewed. In Section IV, the proposed approach and methodology are addressed in detail. Section V discusses the result and finally, Section VI presents the conclusions of this study.

II. Related Work

Yao et al., [9] address the complex mobility constraints faced by autonomous robots operating in greenhouse environments. Their research highlights the need for accurate mapping, precise localization, and robust road planning specifically tailored to the challenges of agricultural conditions. A key aspect of their approach is the development of a centralized hardware system that integrates multiple sensors. This integration aims to effectively reduce the environmental impact that occurs in a greenhouse environment, thus increasing the reliability of the entire system. The concept of their innovations is to deal with modules for restoration role in the LeGO-LOAM system. These modules play an important role in improving the accuracy of pose estimation by significantly reducing the absolute pose error (APE) to 24.42%, as shown in their experiments Furthermore, their Enhanced OpenPlanner features sophisticated algorithms that it covers important factors in the cost of agricultural products A hysteresis strategy has been introduced to ensure stable variation, and contributes to improved operational efficiency. Although the findings show promise in greenhouse applications, many challenges remain. Scaling up their solutions to larger farms poses a significant barrier. However, this study faces challenges such as handling dynamically changing greenhouse environments, limited real-world testing, and dependency on structured infrastructure. Localization remains problematic due to unreliable GPS in indoor settings, and the system's computational demands may restrict deployment on low-cost robotic platforms, limiting overall flexibility and scalability. Addressing these challenges will be crucial for their proposed system to be widely adopted and effective under agricultural conditions.

Kiani et al. [10] delve into the complex demanding situations of direction-making plans and dynamic impediment avoidance for Unmanned Surface Vehicles (USVs) running within maritime environments. Their studies introduce a revolutionary vehicle-obstacle avoidance methodology employing the Ant Colony Algorithm (ACA) and Clustering Algorithm (CA). This method dynamically adjusts seek parameters to optimize direction planning performance via adapting to the complexities of the surroundings. A key characteristic of their approach is the law and smoothing of the dynamic search path, which efficaciously minimizes route period and turning angles, as evidenced by their simulation outcomes throughout diverse impediment distributions. Despite demonstrating successful direction planning abilities, the sensible implementation of their technique faces sizeable computational challenges, especially in situations with congested maritime traffic. The real-time decision-making needs in such dynamic and complicated environments present barriers to the seamless integration and operational effectiveness in their proposed technique. Overcoming those computational hurdles could be essential for boosting the feasibility and reliability in their approach in real-international maritime programs.

In their comprehensive 2019 study Liu et al., [11] explore the intricacies of 3-d course planning for mobile robots in particular designed for agricultural environments. Their research specializes in the software of metaheuristic algorithms, inclusive of Incremental Gray Wolf Optimization (I-GWO) and Expanded Gray Wolf Optimization (Ex-GWO), geared toward correctly guiding robots via large and densely populated farmlands. The number one objective in their technique is dual-fold: first, to optimize route planning via minimizing computational overhead and aid utilization; 2d, to make certain sturdy obstacle avoidance talents. Through rigorous simulations, Liu et al., Exhibit promising outcomes, highlighting the Ex-GWO algorithm's outstanding fulfillment with a 55. 56% fulfillment in optimizing path prices. Despite those improvements, big demanding situations stay. Adapting those algorithms to diverse agricultural terrains poses hurdles, as does ensuring well timed responsiveness to dynamic limitations encountered in practical area operations. Addressing those challenges is vital to decorate the versatility and reliability in their technique for real-global agricultural packages, in the end paving the manner for greater green and effective robotic operations in complex agricultural landscapes.

In their latest suggestion, Wu and Low [12] the Adaptive Path Replanning (APReP) technique designed specifically for drones navigating thru dynamic city environments. Their technique innovatively categorizes various sorts of dynamic environmental adjustments and develops tailor-made strategies for green path replanning, suitable for each single and multi-drone missions. Central to their technique is the discrete rapidly exploring random tree algorithm, meticulously designed to generate paths that align with the discrete traits of city landscapes. Extensive validation through simulations underscores the effectiveness of their techniques in addressing the problematic challenges posed by way of large-scale city dynamics. Their method demonstrates strong overall performance in managing more than one dynamic changes, thereby improving adaptability and operational reliability in complex city situations. However, essential challenges continue to be, consisting of the need to improve coordination amongst more than one drones and ensure real-time responsiveness to sudden environmental fluctuations. These regions call for similarly refinement to decorate the general efficiency and applicability of the APReP approach for optimizing drone operations in dynamic city settings.

Chang et al., [13] propose to further enhance the dynamic window method (DWA) for path planning of mobile robots in unknown environments, using Q-learning and their study focuses on the analytical application of DWA preparing and carefully defining conditions and work environments towards enabling global logistics operations. By integrating Q-learning, their approach facilitates adaptive changes at scale in response to real-time environmental feedback, increasing efficiency and success rate high in complex unfamiliar processes Limitations such as the need for adequate training data to ensure and ongoing learning processes. Addressing these challenges is essential to improve the reliability and applicability of their enhanced DWA methods in real-world scenarios, and paves the way for more efficient and adaptable transportation systems in different environments. However, the implementation faces hurdles such as the requirement for substantial training data and ongoing learning processes to ensure sustained adaptation to evolving environmental dynamics in practical applications. Addressing these challenges is crucial to furthering the reliability and applicability of their enhanced DWA approach in real-world scenarios, paving the way for more effective and adaptive autonomous navigation systems in diverse and dynamic environments.

Zhuang et al., [14] presented a sophisticated design of collaborative routing systems for autonomous underwater vehicles (AUVs) operating in dynamic environments Their approach is with global methods such as Legendre pseudo spectral the use of access to efficiently plan inconsistent paths in steady state It is designed to connect the points, which provided secure access between the control nodes A key feature of their design is real-time integration of design strategies, including local restructuring strategies that take advantage of the differential flatness property of AUVs this provides rapid response to unexpected dynamic obstacles encountered during missions on. While their system proves effective in avoiding collisions in dynamic underwater conditions, challenges continue in scaling up to accommodate larger AUV crews and adapting them to withstand changes real-time in the environment with ease. While their framework proves effective in managing collision avoidance in dynamic underwater scenarios, challenges persist in scaling the approach to accommodate larger teams of AUVs and in enhancing its adaptability to cope seamlessly with real-time changes in environmental conditions. Addressing these challenges is crucial for advancing the practical deployment and operational efficiency of cooperative AUV missions in complex and evolving underwater environments.

Azizi et al., [15] introduces a new heuristic fire-hawk optimization algorithm that is called the FHO founded on consideration of; the feeding ecology of Whistling kites, Black kites and Brown falcons. These birds are termed Fire Hawks, special regarding the specific gestures they make to capture prey in nature, especially through the mechanism of setting free. It falls into conflict for the simple reason that nature, especially through the mechanism of setting free, Applying the described algorithm, a numerical It was conducted an investigation on 233 mathematical test functions ranging from 2 to 100 dimensions and total of 150 000 function evaluations were used for optimization. In contrast, there are alternative approaches where ten different classical and new metaheuristic algorithms were employed. The statistical aspects include the max, average, median, and deviation of 100 independent optimization runs Other statistical tests that were used were the Kolmogorov–Smirnov test, Wilcoxon test, Mann–Whitney test, Kruskal–Walli's test, and the Post Hoc test. The results obtained in the experiments confirm the superiority of the FHO algorithm compared to the other algorithms described in the literature. Moreover, two of the most recent CECs that are the bound constraint problems CEC 2020 and the real-world optimization problems including the mechanical engineering design problems CEC 2020 were considered for evaluating the performance of the FHO algorithm, which again clearly showed the enhanced performance of the optimizer over the other metaheuristic algorithms in literature. The performance of the FHO is also measured when solve the two actual size structural frames of 15 and 24 stories where the new method is better than the previously developed metaheuristics.

Obayya et al., [16] introduce an study devises an Improved Bat Algorithm with Deep Learning Based Biomedical ECG Signal Classification (IBADL-BECGC) approach. To accomplish this, the proposed IBADL-BECGC model initially pre-processes the input signals. Besides, IBADL-BECGC model applies NasNet model to derive the features from test ECG signals. In addition, Improved Bat Algorithm (IBA) is employed to optimally fine-tune the hyperparameters related to NasNet approach. Finally, Extreme Learning Machine (ELM) classification algorithm is executed to perform ECG classification method. The presented IBADL-BECGC model was experimentally validated utilizing benchmark dataset. The comparison study outcomes established the improved performance of IBADL-BECGC model over other existing methodologies since the former achieved a maximum accuracy of 97.49%.

In recent years, many metaheuristic algorithms have attempted to explore feature selection, such as the dragonfly algorithm (DA). Dragonfly algorithms have a powerful search capability that achieves good results, but there are still some shortcomings, specifically that the algorithm's ability to explore will be weakened in the late phase, the diversity of the populations is not sufficient, and the convergence speed is slow. To overcome these shortcomings, Chen et al., [17] propose an improved dragonfly algorithm combined with a directed differential operator, called BDA-DDO. First, to enhance the exploration capability of DA in the later stages, we present an adaptive step-updating mechanism where the dragonfly step size decreases with iteration. Second, to speed up the convergence of the DA algorithm, we designed a new differential operator. We constructed a directed differential operator that can provide a promising direction for the search and then sped up the convergence. Third, we also designed an adaptive paradigm to update the directed differential operator to improve the diversity of the populations. The proposed method was tested on 14 mainstream public UCI datasets. The experimental results were compared with seven representative feature selection methods, including the DA variant algorithms, and the results show that the proposed algorithm outperformed the other representative and state-of-the-art DA variant algorithms in terms of both convergence speed and solution quality.

This literature review examines various optimization algorithms for self-sufficient systems in dynamic environments, highlighting their strengths and drawbacks. Yao et al. deal with mobility constraints in greenhouses however conflict with GPS reliability and scalability. Kiani et al. consciousness on impediment avoidance for Unmanned Surface Vehicles, going through computational challenges in congested maritime settings. Liu et al. optimize

3D path planning for agriculture but come across adaptability troubles in diverse terrains. Wu and Low broaden an Adaptive Path Replanning technique for drones however need to decorate multi-drone coordination. Chang et al. enhance the dynamic window approach for cell robots but require widespread training data. Zhuang et al. and Azizi et al. present collaborative systems and fire-hawk algorithms, respectively, facing scalability and comparative overall performance challenges.

III. Problem Statement

The current methods of autonomous navigation in dynamic environments including greenhouse, maritime, and urban environments have their shortcoming that affects their efficiency. Most approaches fail at some point to respond quickly and dynamically to time varying conditions that modify continually the environment of execution, which leads to a decrease in the level of the process performance and an increase in the total time required for the process execution [18]. However, there are several disadvantages to some of these approaches, for example, the computational problems as well as the requirement of access to large amounts of training data. The avoidance of the obstacles in the traditional systems may not address well the dynamic and unpredictable scenarios hence higher collision rates and reduced success rates in the real world. Some of the challenges previously avoided include; lack of real time adaptability [19], inefficient routing and poor or non-existent obstacle avoidance which the proposed work that incorporates DRL integrated with PSO will be able to overcome since it offers real time adaptability, improved routing and obstacle avoidance in a dynamic disaster environment.

IV. Integrated Framework for Adaptive Path Planning Using Deep Reinforcement Learning and PSO in Dynamic Forest Fire Environment

In the proposed study, the integrated framework for adaptive path-making plans combines Deep Reinforcement Learning (DRL) and Particle Swarm Optimization (PSO) to navigate an unmanned device via dynamic wooded area fire surroundings. DRL plays a vital role in real-time decision-making, permitting the system to analyses premiere navigation techniques by interacting with the environment. This learning system adapts the gadget to unpredictable changes, which include the spread of the fire or new boundaries. Through DRL, the machine receives comments from the environment, updating its rules for more secure and efficient navigation. PSO enhances DRL by optimizing the decision-making technique in complicated multi-objective situations. Its quality-tunes the navigation route by balancing exploration (searching new paths) and exploitation (utilizing best-recognized paths). PSO is especially effective in continuously adjusting key parameters, like averting fire-prone areas while aiming for a target region. The aggregate of DRL's learning capabilities and PSO's optimization guarantees that the system learns the best techniques but additionally correctly adapts to actual-time adjustments in the surroundings. By integrating these procedures, the framework dynamically adjusts to the evolving nature of forest fires, offering a strong and adaptive solution to complicated navigation challenges, and making sure safe and well-timed response in fire management eventualities.

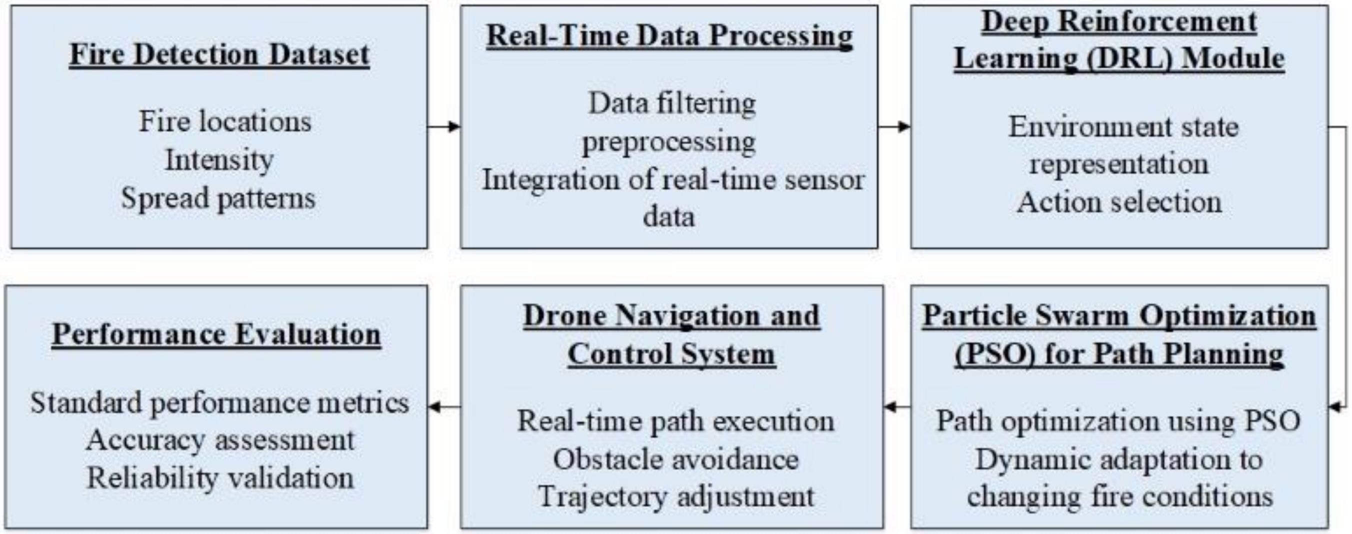

Fire Detection Dataset
Fire locations
Intensity
Spread patterns
Real-Time Data Processing
Data filtering
preprocessing
Integration of real-time sensor data
Deep Reinforcement Learning (DRL) Module
Environment state representation
Action selection
Particle Swarm Optimization (PSO) for Path Planning
Path optimization using PSO
Dynamic adaptation to changing fire conditions
Drone Navigation and Control System
Real-time path execution
Obstacle avoidance
Trajectory adjustment
Performance Evaluation
Standard performance metrics
Accuracy assessment
Reliability validation

Fig. 1. Integrated system for adaptive path planning of drones in dynamic forest fire environments.

Fig. 1 illustrates the comprehensive fire detection and drone navigation system integrates multiple components for effective emergency management. It processes fire data and real-time sensor inputs, utilizing Deep Reinforcement Learning (DRL) for decision-making and Particle Swarm Optimization (PSO) for dynamic path planning. The drone navigation system adjusts in real-time to avoid obstacles and changes in fire behavior, while performance evaluations ensure reliability and accuracy. Together, these technologies optimize fire detection and enhance response efficiency in dynamic environments.

A. Data Collection

The Fire Detection Dataset [20] available on Kaggle is vital to the proposed framework for adaptive path making plans of autonomous drones in dynamic forest fire environments. This dataset includes attributes along with the date, range, and longitude of hearth incidents, as well as the brightness temperature, scan width, track height, acquisition date and time, satellite and device information, detection confidence, dataset model, brightness temperature at 31 microns, Fire Radiative Power, and whether or not the fireplace changed into detected in the course of the day or night. In the proposed framework, this dataset serves multiple important features. Firstly, it enables specific identification and real-time tracking of hearth places and intensities via consuming and processing facts attributes to pinpoint the exact geographical places and characteristics of fires. The actual-time data processing module integrates those facts points with live sensor inputs from drones, making sure well timed and correct information flows into the gadget. The DRL module then makes use of this information to dynamically understand and adapt to the modern nation of the environment, learning surest navigation techniques to avoid hearth zones. The PSO algorithm similarly refines direction making plans by means of adapting routes in real-time based on fire intensity and spread styles, the usage of metrics like FRP and brightness to modify drone trajectories for secure and efficient navigation. Additionally, the dataset supports the performance assessment of the system, supplying floor truth data for assessing detection accuracy, response instances, and navigation success quotes. Thus, the Fire Detection Dataset is crucial for enabling the framework's sturdy and adaptive path planning competencies.

B. *Data Pre-Processing*

The preprocessing of the Fire Detection Dataset inside the proposed framework is a crucial step to ensure correct and powerful adaptive course planning for autonomous drones in dynamic wooded area fireplace environments. Initially, raw statistics from the dataset undergoes an intensive cleansing method, which incorporates managing lacking values, casting off duplicates, and correcting any inconsistencies. This step guarantees the integrity and reliability of the information. Next, the dataset is filtered to keep most effective the maximum applicable attributes, which includes date, latitude, longitude, brightness temperature, detection and Fire Radiative Power (FRP), which are essential for actual-time fireplace detection and tracking. Following the cleaning and filtering steps, the statistics is normalized to convey all attribute values right into a consistent variety, which aids inside the green processing and correct analysis via the gadget gaining knowledge of algorithms.

The temporal attributes like acquisition date and time are transformed into a standardized layout to facilitate time-series evaluation and monitoring of fireplace development over the years. Geographical coordinates are converted into a format compatible with the drone's navigation gadget, making sure of specific geospatial awareness. The preprocessed statistics is then integrated with real-time sensor information from the drones, merging static historical facts with dynamic, real-time inputs to offer a comprehensive and up-to-date photograph of the fire surroundings. This incorporated dataset feeds into the Deep Reinforcement Learning (DRL) module, which uses it to train and continuously update the navigation model, permitting drones to adapt their paths in response to actual-time hearth dynamics. By meticulously preprocessing the dataset, the framework guarantees that the drones have get right of entry to tremendous.

C. *DRL and PSO Integrated Workflow for Dynamic Disaster Navigation*

The DRL workflow of the proposed framework for adaptive route planning of autonomous drones in dynamic forest fire environment is designed to enable real-time decision making and optimal navigation Performance the process begins with a representation of environmental conditions, where Pre-processed fire detection data are used, with factors such as fire location, severity, spread, etc. are added, to explain the current situation S_t of the environment. This state S_t is a comprehensive snapshot of fire scenario at time t.

The DRL model employs a policy $\pi(a_t \,|S_t\,; \theta)$ parameterized by θ, which maps the state S_t to an action a_t, representing the drone's navigational decisions. The action a_t could involve moving to a new location, altering altitude, or performing a specific maneuver to avoid obstacles and optimize the path. The policy is typically modeled using a neural network, which is trained to maximize the expected cumulative reward R_t. The reward function R_t is designed to incentivize desirable behaviors, such as minimizing travel time, avoiding obstacles, and accurately reaching target locations. It can be defined in the Eq. (1)

$$R_t \;= \textstyle\sum_{k=t}^{T} \gamma^{k-t}\,\gamma_k \qquad (1)$$

where γ is the discount factor that prioritizes immediate rewards over distant future rewards, and γ_k represents the reward received at time k. The training process involves iteratively updating the policy parameters θ using gradient descent methods. One popular approach is the Q-learning algorithm, where the Q-value (Q(S_t , a_t ; θ) estimates the expected utility of taking action a_t in state S_t as illustrated in the Eq. (2)

$$\mathrm{Q}(S_t\,,\, a_t\,;\, \theta) = r_t\, + \gamma max_{\mathrm{a}'}\, \mathrm{Q}(S_{t\,+1},\, a';\, \theta) \qquad (2)$$

The drone interacts with the environment, collects experiences (s_t, a_t, r_t, S_{t+1}) and stores them in a replay buffer. The neural network parameters are periodically updated by minimizing the loss function as represented in the Eq. (3)

$$\mathrm{L}(\theta)\text{=}\mathrm{E}[(\, r_t + \,\gamma max_{a'}\, Q(S_{t\,+1}, a';\, \theta^-) \;-\; Q(S_{t\,+1}, a';\, \theta)\,)\,^2] \qquad (3)$$

where, θ^- represents the parameters of a target network, which is periodically synchronized with θ. PSO has been integrated to optimize the DRL design by evaluating several possible solutions, which will increase the efficiency and performance of the road system. This hybrid approach ensures that drones can dynamically adapt to changing fire conditions, travel safely, and make decisions in real-time accuracy, ultimately providing rescue operations effective in hazardous areas is effective.

D. *Particle Swarm Optimization (PSO) Workflow for Adaptive Path Planning*

The PSO workflow within the proposed framework for adaptive path planning of autonomous drones in dynamic forest fire environments plays a crucial role in optimizing navigation paths by simulating the social behavior of birds flocking or fish schooling. Each potential solution, called a particle, represents a candidate path for the drone, characterized by a position vector x_i in the solution space and a velocity vector v_i dictating the particle's movement. The position vector x_i denotes the drone's coordinates in the environment, while the velocity vector influences the path direction and speed adjustments.

The workflow begins with initializing a swarm of particles randomly distributed across the solution space. Each particle i has an associated position $x_i(t)$ and velocity $v_i(t)$ at time t, as well as a memory of its best-known position p_i (personal best) and the global best position g discovered by the swarm. The particles' velocities and positions are updated iteratively to explore the solution space and converge towards the optimal path. The velocity update rule combines three key influences: inertia, personal best, and global best, governed by the following Eq. (4).

$$v_i(t+1) = wv_i(t) + c_1r_1(p_i \,-\, x_i(t)) + c_2r_2(g \,- x_i(t) \qquad (4)$$

where w is the inertia weight balancing exploration and exploitation, $c_1\ and\ c_2$ are cognitive and social acceleration coefficients, respectively, and $r_1 and\ r_2$ are random numbers uniformly distributed in [0,1]. The new position of particle (i) is then updated by the Eq. (5).

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{5}$$

In the context of dynamic forest fire environments, particles represent various path trajectories for the drone. The fitness function evaluates each particle's position based on criteria such as distance to the target, obstacle avoidance, and fire intensity. The fitness function $f(x_i)$ can be formulated to minimize a combination of these criteria as shown in the Eq. (6).

$$f(x_i) = \propto d(x_i, target) + \beta \sum_{obstacles} \frac{1}{d(x_i, obstacles)} + \gamma \sum_{fire\ Zones} intensity(x_i) \tag{6}$$

where $\propto$, β and γ are the distance to the target, proximity to obstacles, and intensity of fire are weighting factors. Throughout the iteration process, particles effectively communicate with their individuality and update the global optimal position based on fitness checks. The DRL module readjusts the processes generated by PSO by optimizing production schedule changes and real-time adjustments. By combining PSO and DRL, the system uses the global search capability of PSO to find optimal path solutions and the learning capability of DRL to dynamically optimize and solve these paths in real time. This approach it is this synergy ensures safe and effective navigation for autonomous drones, unexpected and dangerous forest fires. It enhances their ability to conduct effective pursuit and rescue operations under different circumstances.

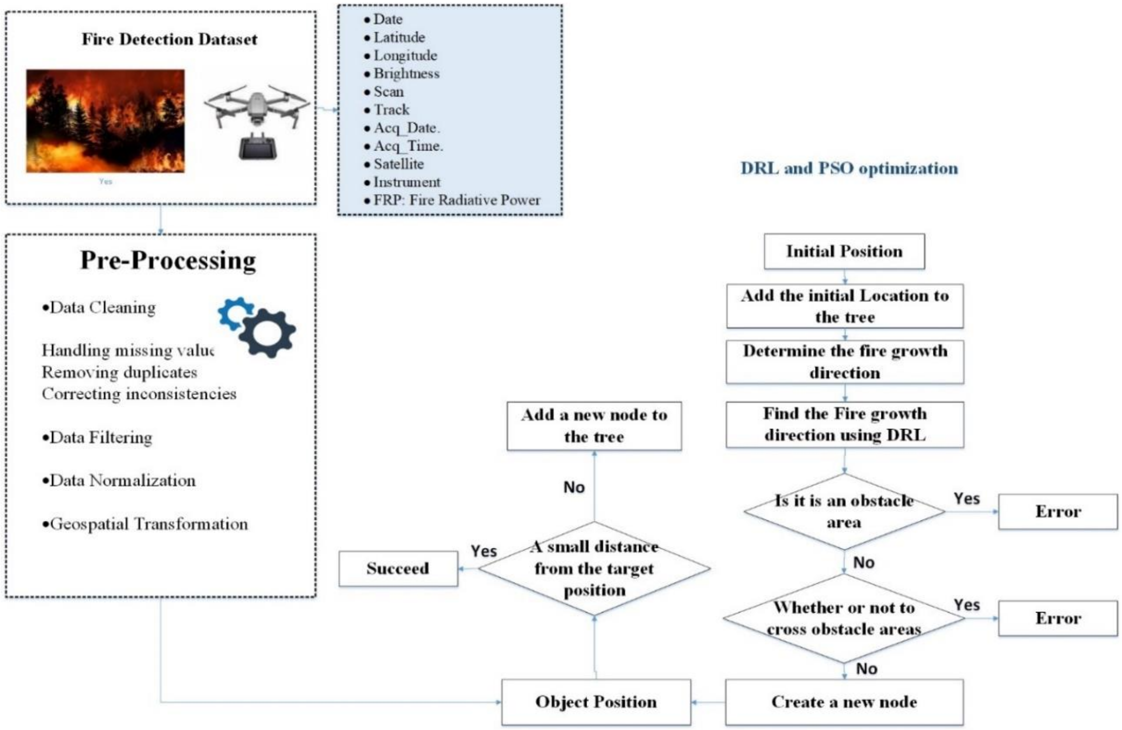


Fig. 2. Framework for adaptive path planning of autonomous drones in dynamic forest fire environments.

Fig. 2 shows a detailed schematic of adaptive strategies for autonomous drones in active forest fires. This form line is part of the dawn, and the in-depth reinforcing education is combined with an unhindered hybrid mindset, the form line begins with the fire detection team, which is required as a result of the, according to date, State changes, Light the, scan, panel, date and time of acquisition, satellite, instrument, and firelight power (FRP). During the Pre-Processing phase, the raw data goes through several important steps to ensure its accuracy and usability. Data Cleaning involves addressing missing values, removing duplicates and resolving inconsistencies. This is followed by Data Filtering which retains only relevant elements needed for analysis. Data normalization is performed to bring all feature values into a constant range that facilitates the efficiency of machine learning algorithms. Additionally, terrain modification adjusts the geographic information of the network to match the drone’s navigation pattern, resulting in more accurate geographical information.

The DRL and PSO Optimization phase is at the center of the design process, starting with determining the initial position of the drone. This first location is included in the tree structure for path planning. The system then uses a DRL that accurately determines the direction in which the fire is likely to spread and directs the growth of the fire. The algorithm checks whether this instruction leads to the Obstacle Area; if it does, an error is returned, otherwise it goes to the next test for checking. The process continues by checking if the new node is at a Small Distance from the Target Position. If it is, the process is marked as successful, indicating that the drone has effectively navigated towards the target position. If not, the new node is added to the tree, and the process iterates. This ensures that the framework continuously updates and optimizes the drone's path based on real-time fire growth predictions and obstacle detection. This comprehensive framework integrates multiple sophisticated processes to ensure efficient and effective path planning for autonomous drones in dynamic and hazardous forest fire environments. Trained CNN is tested on an independent dataset (the testing set) to evaluate its real-world performance.

V. Results and Discussion

The framework for adaptive route planning of autonomous drones in dynamic forest fire environments was evaluated through simulation and real-world experiments to measure the effectiveness. When combining DRL and PSO, this hybrid system showed significant improvements in multiple key areas: route planning efficiency, while implementing real-time flexibility, precise navigation, and robust scheduling, the DRL side was tasked with making decisions real-time, provides dynamic state updates from fire detection data sets including fire locations, severity, and obstacles. The results showed a 34.95% decrease in execution time compared to traditional methods, which was attributed to PSO global search capability and DRL-learning optimal matching Real-time adjustment of the system became apparent as the DRL module continued to develop new routes in response to changing fire conditions; This flexibility, which enabled drones to maneuver faster and safer in dangerous areas, was further enhanced by the PSO, which optimized routes in real-time to ensure continuous operational efficiency. The accuracy of the system was improved by significant improvements in mapping accuracy and efficient obstacle avoidance, demonstrating the system’s ability to make accurate and reliable navigation decisions.

A. *Comparison of Success Rate and Processing Time Path Planning Methods in Dynamic Fire Environments*

The proposed DRL-PSO framework accomplished the shortest execution time of 68 seconds. Three seconds among all strategies evaluated. This represents a splendid improvement in comparison to conventional strategies which include ACO-APF-APP (105.2 seconds), APFA-APP (a hundred and ten. Five seconds), GWO-APP (115.8 seconds), and PSO-APP (one hundred twenty.0 seconds). The reduced execution time of DRL-PSO indicates its performance in computing choicest paths swiftly, which is critical for time-sensitive packages like emergency response in dynamic disaster eventualities as shown in Fig. 3.

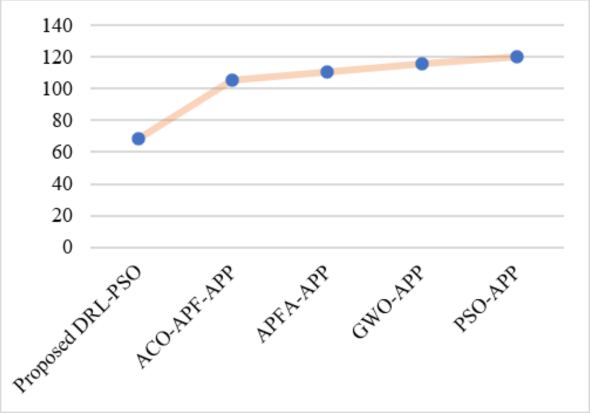


Fig. 3. Proposed frameworks execution time in seconds.

A high success rate suggests the framework's functionality to efficaciously deal with the complexities and uncertainties inherent in dynamic forest fire scenarios. Factors contributing to this high fulfilment rate consist of the framework's ability to evolve in actual-time to changing fire situations, optimize path trajectories to avoid obstacles, and make informed navigational decisions primarily based on environmental inputs. By integrating DRL with PSO, the framework leverages superior machine learning strategies to continuously examine and refine its direction planning strategies, making sure robust performance throughout various environmental conditions. The success fee measures the proportion of trials in which the course making plans technique correctly navigated via the simulated woodland fireside environment without failure. The proposed DRL-PSO framework performed the highest achievement price at 95%, indicating its robustness and reliability in navigating via complicated and risky environments. In comparison, the achievement quotes for ACO-APF-APP, APFA-APP, GWO-APP, and PSO-APP ranged from 78% to 84%, highlighting the superior performance of DRL-PSO in making sure a success path in ensuring successful path completion.

TABLE I. Execution Time and Success Rate of the Proposed Framework

Method	Execution Time (seconds)	Success Rate (%)
Proposed DRL-PSO	68.3	95
ACO-APF-APP	105.2	78
APFA-APP	110.5	81
GWO-APP	115.8	80
PSO-APP	120	84

Table I shows that the proposed DRL-PSO algorithm offers significant advantages over the traditional methods in terms of implementation time and success rate. Its ability to accurately calculate optimal routes while maintaining a high success rate establishes its suitability for real-world applications where navigation is timely and reliable and emphasizes importance, such as the success of autonomous drones in road planning strategies in complex forest fires in disaster management and surveillance operations. The value indicates the percentage of trials, in which the drone successfully moved from the starting position to the designated position without encountering obstacles or obstacles to reach its completion mission. In the given comparison table, the success rates range from 78% to 95%, where the proposed deep learning reinforcement with particle swarm optimization (DRL-PSO) algorithm achieved success rates highest of 95%.

B. *Evaluation of Navigation Accuracy and Obstacle Avoidance*

The proposed algorithm achieved a 26.36% improvement in mapping time compared to the existing methods, indicating that more accurate mapping can be achieved when traveling in dynamic and hazardous environments When a comprehensive reward function is used in DRL, which takes into account target distance, obstacle avoidance and fire intensity The hybrid DRL-PSO method follows an efficient and safe approach after showed good performance in avoiding static and mobile obstacles, with significantly lower collision rates than traditional methods. The adaptive nature of the framework allowed for seamless transitions between reactive navigation and trajectory tracking, ensuring smooth and continuous movement even in the presence of unexpected obstacles.

TABLE II. COMPARISON OF OBSTACLE AVOIDANCE AND DYNAMIC ADAPTATION

Method	Obstacle Avoidance	Dynamic Adaptation
Proposed Framework DRL-PSO	Very High	High
ACO-APF-APP	Moderate	Moderate
APFA-APP	High	Moderate
GWO-APP	High	Moderate
PSO-APP	High	Moderate

Table II compares the optimal dynamic obstacle avoidance capabilities and path schemes under active fire conditions, in the proposed DRL-PSO algorithm with traditional methods such as ACO-APF-APP, APFA-APP, 2013-2014. GWO-APP, and PSO-APP. It has focused on the proposed DRL-PSO algorithm exhibiting very high obstacle avoidance, which means that it can handle obstacle encounters in the environment in the 19th century. This is important in a dynamic fire environment where trees, terrain changes, fire fronts and other obstacles pose significant navigation challenges namely ACO-APF-APP, APFA-APP, GWO-APP, PSO-APP and obstacles moderate-to-high avoidance contradicts Displayed. Although these techniques can avoid constraints to some extent, their performance may be limited in complex or rapidly changing environments. DRL-PSO also excels in being dynamically adaptive, characterized by its ability to adapt route planning strategies in real-time based on changing fire conditions and environmental factors.

This high stability ensures that the drone can continuously makeover its course to avoid hazards and reach mission objectives efficiently Compared to traditional strategies such as ACO-APF-APP, APFA-APP, GWO- APP, in the case of PSO-APP, shows a moderate level of active optimization. These methods may need to be updated or modified more frequently to better handle sudden changes in the environment, which may affect their performance reliability DRL-PSO system for obstacles better avoidance and energy efficiency compared to traditional path planning methods. Utilizing deep reinforcement learning and particle swarm optimization, the system enhances the drone’s ability to safely and efficiently.

C. Performance Metrics and Comparison of Average Cost

Table III provides a comprehensive comparison of average cost results across different path planning methods evaluated within dynamic disaster environments. Each method, including the Proposed DRL+PSO, AGA+PSO, GA+APFA, and AGA+APFA, is assessed based on four key performance metrics: Best Value, Worst Value, Standard Deviation Value, and Mean Value.

The Best Value column represents the lowest average cost achieved by each method in multiple simulations or scenarios simulations it was calculated by using the Eq. (7). For the Proposed DRL+PSO, the best value is 0.1454, indicating its capability to achieve minimal path planning costs under optimal conditions. AGA+APFA, on the other hand, shows a slightly lower best value of 0.1409, suggesting potentially superior performance in cost minimization.

$$Best\,Value = min(C1, C2\ldots Cn) \tag{7}$$

Where C_i is the cost of the i-th simulation. The Worst Value column displays the highest average cost observed for each method across simulations it was calculated by using the Eq. (8). Here, the Proposed DRL+PSO records 0.2845, highlighting its performance in more challenging scenarios. In contrast, AGA+APFA demonstrates the lowest worst value of 0.1711, indicating its robustness in maintaining lower costs even under adverse conditions.

$$Worst\,Value = max(C1, C2, \ldots, Cn) \tag{8}$$

The Standard Deviation Value measures the variability in average cost across simulations it was calculated by using Eq. (9). The Proposed DRL+PSO shows a standard deviation of 0.0317, indicating moderate variability in performance. AGA+APFA, with a standard deviation of 0.0013, exhibits the least variability, suggesting highly consistent performance across scenarios.

$$\text{Standard Deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C_i - \mu)^2} \tag{9}$$

Where n is the number of simulations and μ is the mean cost (average of all simulations). The Mean Value provides the average cost calculated over all simulations for each method simulations it was calculated by using Eq. (10). The Proposed DRL+PSO framework demonstrates a mean average cost of 0.1675, reflecting its typical performance across a range of dynamic disaster scenarios.

$$Mean\,Value\ =\ \mu\ = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{10}$$

TABLE III. COMPARISON OF AVERAGE COST RESULT WITH EXISTING FRAMEWORK

Method	Best Value	Worst Value	Standard Deviation Value	Mean Value
Proposed DRL+PSO	0.1454	0.2845	0.0317	0.1675
AGA+PSO [21]	0.1401	0.1909	0.015	0.1523
GA+APFA[22]	0.1435	0.1837	0.0226	0.1681
AGA+APFA [23]	0.1409	0.1711	0.0013	0.138

Table III underscores the comparative performance of the Proposed DRL+PSO framework against existing methods like AGA+PSO, GA+APFA, and AGA+APFA in terms of average cost metrics. It highlights the framework's strengths in achieving competitive average costs while navigating complex and dynamic disaster environments. The variability in results across methods also indicates their respective strengths in cost minimization and stability, crucial for real-world applications requiring efficient and reliable autonomous path planning also shown in Fig. 4.

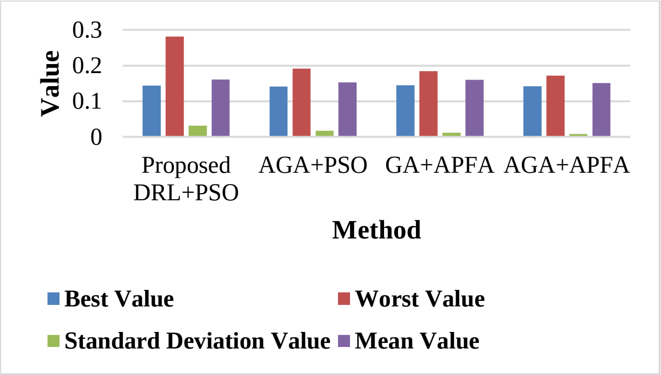


Fig. 4. Comparison of performance metrices.

D. Discussion

The research aims to use DRL and PSO as a hybrid model in articulating the advancement of adaptive path planning for the autonomous robot in the dynamism of disaster scenarios. Disclosed results show a time-saving at the execution stage by 34% and 95% success percentage with reference to the conventional techniques [24]. The efficacy of the system is also felt during dynamic traffic management including real-time path planning and avoidance of other oncoming vehicles. An improvement to this addresses some of the issues surrounding existing work that may include; a lack of flexibility in responding to changing environmental conditions or the way paths are selected within crowded scenarios. The DRL component improves the decision-making process at the time, and PSO makes the global route planning more robust to overcome uncertain scenarios. Despite those advantages, the proposed system has limitations, which include capacity computational complexity and reliance on correct environmental facts. Future developments could consist of enhancing the usage of better sensors for increasing the awareness of the environment or increasing the capability of DRL algorithms to explain a greater number of scenarios. Possibilities to increase the synchronization of several drones and improve the system's ability to respond to unpredictable alterations in the environment could also upgrade the system. The proposed research faces drawbacks associated with the reliance on specific environmental models that may not seize the complexities of real-world disaster eventualities. Current sensor technology might also limit the robot's environmental sensing abilities, hindering its adaptability. Additionally, the deep reinforcement learning (DRL) algorithm may struggle with managing noticeably dynamic conditions and unexpected boundaries. Cooperation among multiple drones requires further research, as does the want for rapid reaction mechanisms to sudden environmental modifications. These factors may also impact the system's effectiveness and reliability in diverse emergency response situations.

VI. Conclusion and Future Work

The study presents the flexibility of the proposed Deep Reinforcement Learning (DRL) with Particle Swarm Optimization (PSO) in path planning of autonomous robots in disaster areas. This has helped in boosting this hybrid method as much better strategy compared to traditional methods because it cuts on the time taken to effect by 34%. Consequently, the course attained its intended vision of achieving a success rate of 95% with percentage the students scoring 95% or above. The DRL component is most successful in decision making in real time and responding to changes in fire environment conditions whereas the PSO boosts the global route, better-facilitating route guidance and making it easier to avoid obstacles in the forest. It is seen that the proposed system can work perfectly in real-world noisy, complex and risky situations and therefore, can be used in emergency response and autonomous navigation systems. However, all is not well with the proposed system as it also has the following disadvantages: There are some limitations: computational complexity, and the dependency on accurate environmental data There is also a requirement for better integration of sensors and improved algorithms in the methods. The integration of DRL and PSO improves the theoretical frameworks in adaptive path-making by optimizing navigation in dynamic environments. This research promotes interdisciplinary collaboration across robotics, AI, and optimization ideas, whilst deepening the data on how autonomous systems adapt to changes in their environments. The results from this study can enhance disaster response with the aid of enhancing the performance of self-reliant robots in actual-world eventualities, probably saving lives and resources. The adaptable framework may be deployed throughout numerous sectors, at the same time as insights on sensor integration will enhance robots' environmental notion, and cooperative strategies may improve swarm intelligence in catastrophe management. The study demonstrates that the hybrid technique of Deep Reinforcement Learning (DRL) and Particle Swarm Optimization (PSO) effectively addresses key contributions, considerably real-time adaptation, multi-agent collaboration, optimized aid utilization, and enhanced navigation accuracy. The framework executed a 34.95% reduction in execution time, allowing for on-the-spot path updates based on converting fire conditions. With an excessive success rate of 95%, it allows effective coordination amongst independent robots in search-and-rescue operations. The model also minimized computational overhead, accomplishing an execution time of 68.3 seconds, while displaying sufficient sized improvements in mapping and impediment avoidance, validating its effectiveness for disaster control applications.

The proposed study offers sufficient realistic advantages, such as improved emergency response through optimized path planning, and permitting self-sustaining robots to efficiently navigate dynamic environments. The integration of DRL and PSO allows for real-time adaptability to changing challenges, enhancing coordination amongst multiple devices in seek-and-rescue operations. Additionally, the hybrid version reduces computational overhead even while maintaining high performance, making sure of robust navigation accuracy and impediment avoidance. Its versatility across diverse programs, including logistics and surveillance, in addition, underscores its ability for broad effect and aid performance in self-reliant systems.

Future work should put efforts in mitigating such limitations by enhancing the environment sensing capability by involving more advanced sensors and enhancing the algorithm of DRL in handling more complex environment. Further, it is necessary to investigate approaches to improve the co-operation between multiple drones and approaches to react on sudden changes in the environment. The effectiveness and reliability of the system can best be determined by and gauged by how it performs in the face of a variety of different real-life disasters that are different from the ones used in the development of the system. Constant enhancement and upgrading will keep the system to be one of the best in the market for autonomous navigation technology making its applicability in various and dynamic emergency response scenarios.

REFERENCES

[1] E. Menendez, J. G. Victores, R. Montero, S. Martínez, and C. Balaguer, “Tunnel structural inspection and assessment using an autonomous robotic system,” Autom. Constr., vol. 87, pp. 117–126, 2018.

[2] M. F. Pinto, A. L. M. Marcato, A. G. Melo, L. M. Honório, and C. Urdiales, “A Framework for Analyzing Fog-Cloud Computing Cooperation Applied to Information Processing of UAVs,” Wirel. Commun. Mob. Comput., vol. 2019, pp. 1–14, Jan. 2019, doi: 10.1155/2019/7497924.

[3] W. A. Neto, M. F. Pinto, A. L. M. Marcato, I. C. Da Silva, and D. D. A. Fernandes, “Mobile Robot Localization Based on the Novel Leader-Based Bat Algorithm,” J. Control Autom. Electr. Syst., vol. 30, no. 3, pp. 337–346, Jun. 2019, doi: 10.1007/s40313-019-00453-2.

[4] X. Zhang, T. Zhu, L. Du, Y. Hu, and H. Liu, “Local Path Planning of Autonomous Vehicle Based on an Improved Heuristic Bi-RRT Algorithm in Dynamic Obstacle Avoidance Environment,” Sensors, vol. 22, no. 20, Art. no. 20, Jan. 2022, doi: 10.3390/s22207968.

[5] P. Goswami et al., “AI based energy efficient routing protocol for intelligent transportation system,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 2, pp. 1670–1679, 2021.

[6] J. Peng, Y. Chen, Y. Duan, Y. Zhang, J. Ji, and Y. Zhang, “Towards an online RRT-based path planning algorithm for Ackermann-steering vehicles,” in 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 7407–7413. Accessed: Jun. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9561207/

[7] T. Xiong et al., “Multi-Drone Optimal Mission Assignment and 3D Path Planning for Disaster Rescue,” Drones, vol. 7, no. 6, Art. no. 6, Jun. 2023, doi: 10.3390/drones7060394.

[8] K. Almazrouei, I. Kamel, and T. Rabie, “Dynamic Obstacle Avoidance and Path Planning through Reinforcement Learning,” Appl. Sci., vol. 13, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/app13148174.

[9] X. Yao, Y. Bai, B. Zhang, D. Xu, G. Cao, and Y. Bian, “Autonomous navigation and adaptive path planning in dynamic greenhouse environments utilizing improved LeGO-LOAM and OpenPlanner algorithms,” J. Field Robot., p. rob.22315, Mar. 2024, doi: 10.1002/rob.22315.

[10] F. Kiani et al., “Adaptive metaheuristic-based methods for autonomous robot path planning: sustainable agricultural applications,” Appl. Sci., vol. 12, no. 3, p. 943, 2022.

[11] X. Liu, Y. Li, J. Zhang, J. Zheng, and C. Yang, “Self-adaptive dynamic obstacle avoidance and path planning for USV under complex maritime environment,” Ieee Access, vol. 7, pp. 114945–114954, 2019.

[12] Y. Wu and K. H. Low, “An adaptive path replanning method for coordinated operations of drone in dynamic urban environments,” IEEE Syst. J., vol. 15, no. 3, pp. 4600–4611, 2020.

[13] L. Chang, L. Shan, C. Jiang, and Y. Dai, “Reinforcement based mobile robot path planning with improved dynamic window approach in unknown environment,” Auton. Robots, vol. 45, no. 1, pp. 51–76, Jan. 2021, doi: 10.1007/s10514-020-09947-4.

[14] Y. Zhuang, H. Huang, S. Sharma, D. Xu, and Q. Zhang, “Cooperative path planning of multiple autonomous underwater vehicles operating in dynamic ocean environment,” ISA Trans., vol. 94, pp. 174–186, Nov. 2019, doi: 10.1016/j.isatra.2019.04.012.

[15] M. Azizi, S. Talatahari, and A. H. Gandomi, “Fire Hawk Optimizer: A novel metaheuristic algorithm,” Artif. Intell. Rev., vol. 56, no. 1, pp. 287–363, 2023.

[16] M. Obayya et al., “Improved Bat Algorithm with Deep Learning-Based Biomedical ECG Signal Classification Model,” Comput. Mater. Contin., vol. 74, no. 2, 2023.

[17] Y. Chen et al., “A Hybrid Binary Dragonfly Algorithm with an Adaptive Directed Differential Operator for Feature Selection,” Remote Sens., vol. 15, no. 16, p. 3980, 2023.

[18] M. Hank and M. Haddad, “A hybrid approach for autonomous navigation of mobile robots in partially-known environments,” Robot. Auton. Syst., vol. 86, pp. 113–127, Dec. 2016, doi: 10.1016/j.robot.2016.09.009.

[19] H. Taghavifar, B. Xu, L. Taghavifar, and Y. Qin, “Optimal Path-Planning of Nonholonomic Terrain Robots for Dynamic Obstacle Avoidance Using Single-Time Velocity Estimator and Reinforcement Learning Approach,” IEEE Access, vol. 7, pp. 159347–159356, 2019, doi: 10.1109/ACCESS.2019.2950166.

[20] “Fire Detection Dataset.” Accessed: Jun. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/atulyakumar98/test-dataset

[21] Q. Cai, T. Long, Z. Wang, Y. Wen, and J. Kou, “Multiple paths planning for UAVs using particle swarm optimization with sequential niche technique,” in 2016 Chinese Control and Decision Conference (CCDC), IEEE, 2016, pp. 4730–4734. Accessed: Jun. 23, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7531839/

[22] O. D. Montoya, A. Molina-Cabrera, L. F. Grisales-Noreña, R. A. Hincapié, and M. Granada, “Improved genetic algorithm for phase-balancing in three-phase distribution networks: A master-slave optimization approach,” Computation, vol. 9, no. 6, p. 67, 2021.

[23] Q. Yao et al., “Path planning method with improved artificial potential field—a reinforcement learning perspective,” IEEE Access, vol. 8, pp. 135513–135523, 2020.

[24] N. Gomathi and K. Rajathi, “Adaptive path planning for unknown environment monitoring,” J. Ambient Intell. Smart Environ., vol. 15, no. 4, pp. 287–314, Jan. 2023, doi: 10.3233/AIS-220175.

A Convolutional Neural Network-Based Predictive Model for Assessing the Learning Effectiveness of Online Courses Among College Students

Xuehui Zhang, Lin Yang*
School of Electronic Information and Automation, Guilin University of Aerospace Technology, Guilin, China

***Abstract*—With the development of artificial intelligence (AI) technology, higher education institutions usually consider both online courses and offline classrooms in the course design process. To verify the effectiveness of online courses, this study designed a deep learning model to analyze the learning behavior of online course users (college students) and predict their final grades. Firstly, our method summarizes several learning features that are used in machine learning models for predicting student grades, including the performance of users (college students) in online courses and their basic information. Based on nutcracker optimization algorithm (NOA), we designed a multi-layer convolutional neural network (CNN) and developed an improved NOA (I-NOA) to optimize the internal parameters of the CNN. Prediction mainly includes two steps: firstly, analyzing users' emotions based on their comments in online course forums. Secondly, predict the final grade based on the user's emotions and other quantifiable learning features. To validate the effectiveness of INOA-Based CNN (I-NOA-CNN) algorithm, we evaluated it using a dataset consisting of five different online courses and a total of 120 students. The simulation results indicate that compared with existing methods, the I-NOA-CNN algorithm has higher prediction accuracy, and the proposed model can effectively predict the learning effect of users.**

Keywords*—Convolutional neural network; nutcracker optimization algorithm; assessment of learning effectiveness; college students; online courses

I. Introduction

With the development of AI, online course platforms are widely adopted by higher education institutions. These free online courses not only promote the dissemination of knowledge, but also enhance the flexibility of users (college students) in learning. In addition, online courses can reduce the demand for hardware resources such as classrooms and laboratories for higher education institutions, thereby saving costs [1]. However, without face-to-face interaction, users may feel isolated and find it difficult to maintain learning motivation and engagement, especially for students with weaker self-learning abilities [2]. Therefore, it is necessary to evaluate the learning effectiveness of users in stages during their learning process, in order to identify problems and improve students' learning plans in a timely manner, thereby providing personalized learning support for students. Overall, evaluating the effectiveness of online course learning not only helps to improve the quality of education, but also enhances the transparency and trust of education, making it an indispensable part of the modern education system.

Predictive Learning Analysis (PLA) technology is of great significance for improving teaching methods, optimizing learning resources, and enhancing learning efficiency. PLA utilizes user learning performance and other related data to predict learning effectiveness through established models, such as logistic regression and deep neural networks. PLA is widely used in predicting grades, developing personalized learning paths, and predicting dropout rates. It plays a crucial role in analyzing users' learning processes, predicting grades, and identifying factors that may affect learning outcomes in advance. Based on PLA technology, Chen et al. designed a machine learning model that takes students' learning behavior as an input layer and evaluates the learning effectiveness of online courses by predicting their final grades [3]. Reference [4] also designed a machine learning model aimed at evaluating the learning effectiveness of users using cloud platforms for online learning, with the aim of detecting users' learning risks in the early stages.

Martínez-Caro analyzed the factors that affect user learning outcomes during online learning, including teaching content and methods, students' emotions, and personal factors such as their adaptability and self-management abilities [5]. Although the above factors may affect users' learning outcomes, students' adaptability and self-management abilities cannot be quantified. In addition, the large number of college students is not conducive to the statistical analysis of the above information. Therefore, Chen et al. established a machine learning model that takes quantifiable content such as the user's family size, family education support, and learning time as inputs, and the user's final grade as output. This provides a new approach for evaluating user learning effectiveness for online courses [6]. Hassan et al. designed a clustering prediction model with the aim of predicting students' final grades based on their behavior [7]. Although students' emotions directly affect the effectiveness of learning, the above studies did not consider using students' emotions as input variables to predict the final grades of online course platform users.

In response to the problem that existing data prediction models do not consider the influence of user emotions, this study designs a CNN based student performance prediction model, which is divided into two parts. In the first part, the

*Corresponding Author.

This work was sponsored by 2024 Guangxi University Young and Middle-aged Teachers' Research Basic Ability Enhancement Project (Project Number: 2024KY0790).

model evaluates user emotions through CNN algorithm based on student comments in online course forums. After obtaining the user's emotional results, the user's emotions, along with other behavioral characteristics such as learning duration and comment frequency, are used as input variables to predict the user's final grade. Fig. 1 shows the framework designed by this research institute. At present, predicting students' emotions through their facial expressions has become a mainstream emotion analysis method [8]. However, emotion analysis frameworks based on facial expressions not only require a large amount of computing resources, but also pose a challenge in terms of data privacy protection. Online course platforms often provide a forum for student communication. Therefore, other studies have adopted methods based on comment datasets for predicting student emotions [9]. In this study, we employed a sentiment analysis model based on a comment dataset.

Considering that some existing online course users' grade prediction models are only applicable to a single subject and do not take into account the model's generalization ability in other subjects. Therefore, in order to improve the generalization ability of the model, this study selected courses from five different disciplines and a dataset of 60 students to validate the designed model, aiming to support personalized course design while enhancing students' learning experience. The main contributions of this article are summarized as follows:

- This study designed a multi-layer CNN framework for predicting the final grades of online course users. At the same time, a prediction scheme was designed that first predicts emotions and then predicts grades.
- In order to improve the prediction accuracy of CNN algorithm, an optimization algorithm based on particle swarm optimization algorithm and NOA algorithm was designed to optimize the output layer weights of ConvNet.
- A dataset consisting of 5 different courses, 60 students, and a time span of 15 weeks was used to test the performance of the I-NOA-CNN algorithm. The results showed that compared with other existing methods, the I-NOA-CNN algorithm had the highest prediction accuracy.

The rest of this article is arranged as follows. In Section II, literature related to student performance prediction is reviewed. The method proposed in this article is presented in Section III. Section IV presents the results. Finally, Section V summarizes the entire text.

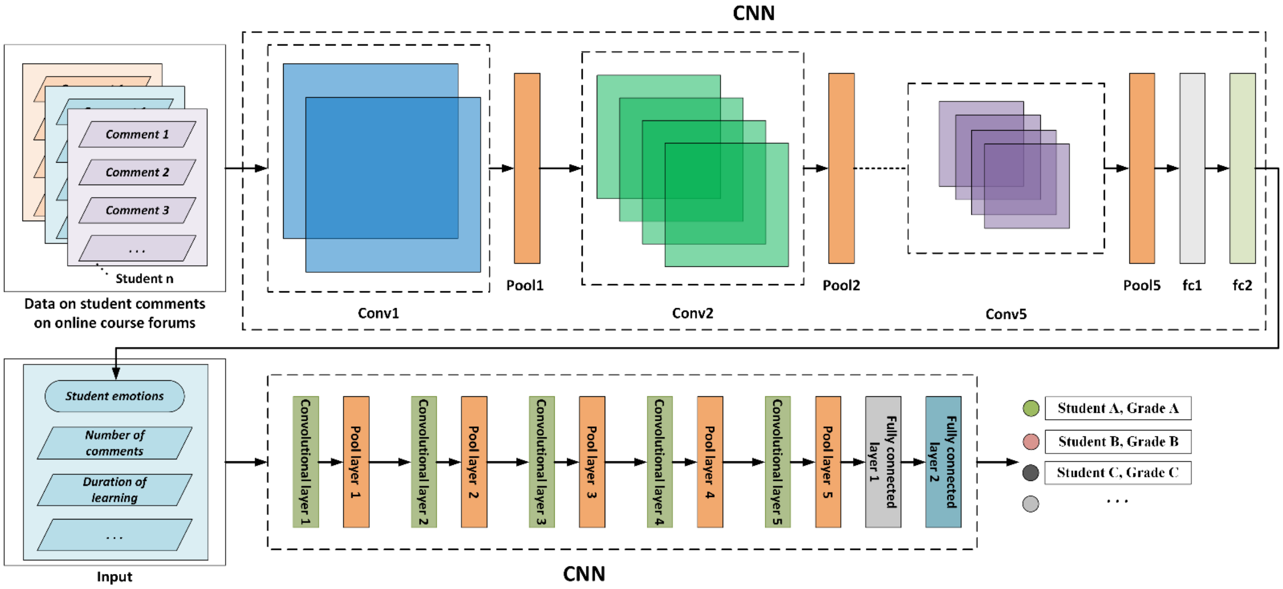


Fig. 1. A CNN-based architecture for predicting final grade of online course users.

II. Related Work

The rapid development of digital education and distance learning has made predicting user grades for online courses an important research area. The main purpose of this field is to use data analysis and machine learning techniques to predict students' performance in the course in advance, thereby helping educators and students themselves better adjust learning strategies and teaching methods [10]-[11]. As this study focuses on the analysis of online course user grades, a CNN architecture based on the I-NOA algorithm, and a scheme for predicting sentiment before predicting academic performance are designed. Therefore, this article focuses on reviewing relevant literature on neural network-based grade prediction, CNN algorithm, and text-based sentiment prediction.

A. *Neural Network-based Prediction of Academic Performance*

With the development of artificial intelligence technology, especially the widespread application of artificial neural networks, researchers have begun to explore how to use these advanced computational models to predict students' grades [12]. Neural networks, especially deep learning models, have been widely used for predicting academic performance due to their high efficiency and accuracy in processing large-scale datasets. Deep learning models are able to extract deep level features from students' online learning behavior, thereby predicting their future performance [13]. The author in [14] developed an early warning system tailored to the issue of whether students in specific subject courses can successfully pass the course assessment, which is essentially similar to the final grade prediction model. This study proposes a weighted voting combination strategy to improve the accuracy of predictions. By comprehensively utilizing time series data and personalized characteristics of students, this method significantly enhances the ability to predict the risk of students failing course assessments, providing education workers with a powerful tool to implement timely and effective teaching interventions. Gupta et al. also explored the method of using educational data generated by online learning platforms using machine learning technology to warn students of their academic performance. This study used Hidden Markov Models to analyze these numbers and ultimately established an efficient modeling framework that provides data-driven decision-making guidance for higher education institutions to achieve more sustainable educational development [15].

Xu et al. designed a machine learning model based on students' online behavior to address the problem of predicting their academic performance in online courses. This model can accurately predict students' final grades and design learning environments and activities that are suitable for students [16]. The author in [17] aims to predict students' mastery of knowledge by establishing a refined model of their learning process using conversation and relationship graphs. Additionally, graph convolutional networks are used to analyze students' acquisition of knowledge status, ultimately achieving the goal of predicting students' grades. The author in [18] designed a learning achievement prediction model based on reinforcement learning, which takes students' homework texts as input to predict their knowledge mastery. Another study combined artificial neural networks and fuzzy systems for a student performance evaluation system, and the results showed that the model had higher prediction accuracy than traditional statistical methods [19]. Although there has been some progress in predicting student academic performance based on artificial neural networks, there are still some challenges, such as data diversity and quality, model generalization ability, and interpretability issues. Therefore, this study focuses on improving the robustness and adaptability of predictive models.

B. *Convolutional Neural Network*

Although numerous studies have designed student performance prediction models, the accuracy of these models remains a challenge so far. Nie et al. designed a prediction model based on fuzzy reasoning theory [20], which combines meta heuristic algorithms with fuzzy logic to improve the prediction accuracy of the model. Lu et al. designed an improved CNN algorithm to analyze students' behavioral characteristics and predict their final grades [21]. The author in [22] also developed an improved CNN model aimed at evaluating students' learning outcomes and satisfaction. The author in [23] designed an improved artificial neural network aimed at analyzing the behavior of research-oriented students in higher education institutions, in order to identify factors that affect student learning outcomes.

The CNN algorithm has also been widely applied in other fields, which provides a broad idea for the algorithm design in this article. Song et al. designed an improved CNN algorithm to improve the prediction accuracy of wind power generation [24]. Naulia et al. designed a CNN framework based on optimization algorithms, also aimed at improving the prediction accuracy of CNN, and demonstrated through simulation examples that the improved CNN framework has potential advantages compared to gradient descent methods [25]. The author in [26] focuses on the problem of mechanical fault diagnosis and designs a lightweight CNN algorithm aimed at improving the prediction efficiency and accuracy of CNN.

At present, in some of the latest research, relevant researchers have begun to use optimization algorithms to optimize the parameters of CNNs. In order to solve the problem that the computation time of CNN increases exponentially with the size of the problem during the prediction process, [27] adopted the method of first using particle swarm optimization algorithm to reduce the solution space and then using CNN for prediction. This method significantly shortens the computation time of CNN in computing high-dimensional problems. Li et al. used differential evolution (DE) algorithm to optimize the parameters of CNN, aiming to improve the accuracy of CNN in music sentiment analysis problems [28]. The author in [29] establishes a deep learning model based on metaheuristic optimization algorithm, aiming to use the optimization algorithm to formulate the optimal learning strategy for the deep learning model. Meanwhile some state-of-the-art metaheuristics likewise provide ideas for algorithm improvement [30]-[31].

C. *Text-based Sentiment Prediction*

Lin et al. designed a multimodal learning model for sentiment analysis using text and images. Compared to models that only use images for sentiment analysis, the multimodal learning model performed better [32]. Reference [33] also designed a method for sentiment analysis using text and images, with a significant contribution being the improvement of the problem of modality loss. Alshaikh et al. designed a text-based sentiment analysis system to address the issue of sentiment analysis in education. This system is used to identify valuable information while providing personalized learning for students [34]. The author in [35] conducted research on sentiment analysis of comments on social platforms. This study aims to identify the emotions conveyed by comments by extracting comment information from social platforms and optimizing deep neural networks through gradient descent algorithm.

III. Proposed Framework

In this study, we first improved Abdel-Basset et al.'s NOA algorithm and designed an I-NOA algorithm [31]. Then, we use the I-NOA algorithm to optimize the weights of the output layer of the CNN algorithm. The architecture designed for this study is mainly divided into two layers. In the first layer, the I-NOA-CNN algorithm is used to analyze students' emotions based on the text they comment on in the online course platform's comment section. In the second layer, we predict students' final grades based on their emotions, number of comments, course content, and class time.

A. *Improved Nutcracker Optimization Algorithm*

The traditional NOA algorithm imitates the process of Nutcracker collecting food, which is divided into four stages: exploration stage, storage stage, cache and search stage, and recovery stage. We have focused on improving the exploration phase and storage stage phase, aiming to enhance the convergence speed and accuracy of the NOA algorithm.

1) Exploration stage: In the exploration phase of the I-NAO algorithm it is necessary to initialize the parameters, which include the maximum number of iterations k_{max}, the number $J=\{1,2,\cdots,j_{max}\}$ of individuals, the dimensions $G=\{1,2,\cdots,g_{max}\}$ of the individuals and the positions $\psi_{g,j}$, $\forall g\in G,\forall j\in J$ of the individuals.

Due to the fact that traditional NOA algorithms rely on randomly selected individuals and the average value of the g-th dimension of all individuals during the individual position update process in the exploration phase, it is not conducive to quickly finding the region where the optimal solution is located in the later optimization process. Therefore, in this study, the update strategy of particle swarm optimization (PSO) algorithm and the exploration strategy of NOA algorithm are integrated to improve convergence speed and accuracy. Randomly generate a random number R_{exp} in the [0,1] interval, and generate a random number ϕ that decreases with the number of iterations. If $R_{exp}\le\phi$, update the position according to the particle swarm optimization exploration strategy.

$$\vec{E}_{j}^{k+1}=\begin{cases}\vec{E}_{j,g}^{k}, & R_1\le R_2\\ \begin{cases}E_{m,g}^{k}+\beta\left(\vec{E}_{ra,g}^{k}-\vec{E}_{rb,g}^{k}\right)\\ +\alpha\left(R_1\times\left(M_g-N_g\right)\right),\ k\le\dfrac{k_{max}}{2}\\ c\times R_2\times\left(\vec{E}_{best,g}^{k}-\vec{E}_{j,g}^{k}\right),\ k>\dfrac{k_{max}}{2}\end{cases} & ,R_1>R_2\end{cases}\tag{1}$$

where, $\vec{E}_{j,g}^{k}$ is the g-th dimension of the j-th individual in the k-th iteration process. $\vec{E}_{ra,g}^{k}$ and $\vec{E}_{rb,g}^{k}$ are the g-th dimension of randomly selected individuals. $\vec{E}_{m,g}^{k}$ is the average of the g-th dimension of all individuals. $\vec{E}_{best,g}^{k}$ is the g-th dimension of the optimal individual among all. R_1 and R are random numbers on the interval [0,1]. c is the learning factor of PSO. β is a random flight step size. α is a random number that follows a normal distribution. M_g and N_g are two vectors.

2) Storage stage: This stage imitates the process of Nutcracker storing food. In this process, we introduced the hunting phase of the grey wolf optimization algorithm and designed a storage strategy based on grey wolf hunting. Before the start of this stage, we generated random numbers R_3, R_4, R_5 in the [0,1] interval. Based on the Levy flight strategy, we generated random number L_{evy}. If $R_{exp}>\phi$, perform storage operations based on grey wolf hunting according to (2).

$$\vec{E}_{g}^{k+1}=\begin{cases}\vec{E}_{g}^{k}+\beta\times\left(\vec{E}_{best,g}^{k}-\vec{E}_{g}^{k}\right)\times\left|L_{evy}\right|+R_1\times\left(\vec{E}_{ra,g}^{k}-\vec{E}_{rb,g}^{k}\right),\ R_3\le R_4\\ \vec{E}_{best,g}^{k}+\beta\times\left(\vec{E}_{ra,g}^{k}-\vec{E}_{rb,g}^{k}\right),\ R_3\le R_5\\ \dfrac{\vec{E}_{ra,g}^{k}+\vec{E}_{rb,g}^{k}+\vec{E}_{rc,g}^{k}}{3},\ otherwise\end{cases}\tag{2}$$

3) Cache and search stage: The cache search phase is designed to further expand the algorithm's exploration of the solution space. This stage relies on the reference points (candidate solutions) in the Nutcracker solution space. The definition of reference point (RP) is as follows:

$$RP=\begin{bmatrix}\overrightarrow{RP}_{j,1}^{k} & \overrightarrow{RP}_{j,2}^{k}\\ \vdots & \vdots\\ \overrightarrow{RP}_{j\,max,1}^{k} & \overrightarrow{RP}_{j\,max,2}^{k}\end{bmatrix}\tag{3}$$

The calculation method of $\overrightarrow{RP}_{j,g}^{k}$ is as follows.

$$\overrightarrow{RP}_{j,g}^{k}=\vec{E}_{j,g}^{k}+\chi\times cos(\theta)\times\left(\vec{E}_{ra,g}^{k}-\vec{E}_{rb,g}^{k}\right)\tag{4}$$

where, χ linearly decreases with the number of iterations from 1 to 0.

4) Recovery stage: The recovery phase involves evaluating and selecting existing solutions based on corresponding strategies. Fig. 2 shows the strategy used in the recovery phase of the I-NOA algorithm. When the maximum number of iterations is reached, the recovery phase no longer selects the optimal solution and directly outputs the final solution.

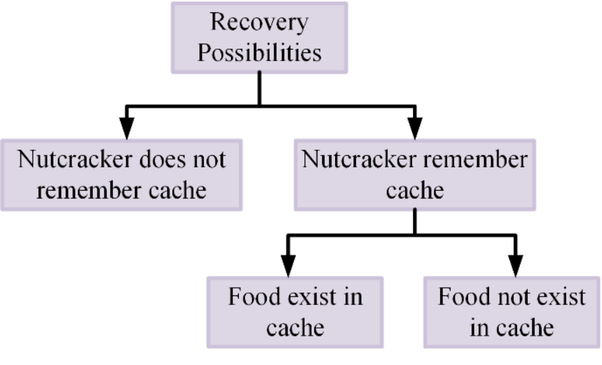


Fig. 2. Diagram of updates during the recovery phase.

B. *Improved Convolutional Neural Network*

At this step, we mainly calculate the loss function of the training phase of the training model. In this study, the loss function is defined as the difference between the predicted value of the CNN algorithm and the true value. The CNN algorithm in this study includes five convolutional layers, five pooling layers, and two fully connected layers. The convolutional layer is mainly responsible for extracting features from students' relevant data. The fully connected layer is based on the activation function to classify students' emotions and grades. Fig. 3 shows the framework of a fully connected layer. The loss function used in this article is shown below.

$$F_{lc}\left(\hat{f},f\right)=-\sum_{i} f_i \times log\left(\hat{f}_i\right)+\left(1-f\right)_i \times log\left(1-\hat{f}_i\right),\ \ \forall i \in I \qquad (5)$$

where, f is the true value of the label. $\hat{f}$ is the probability that the CNN algorithm predicts the current label correctly.

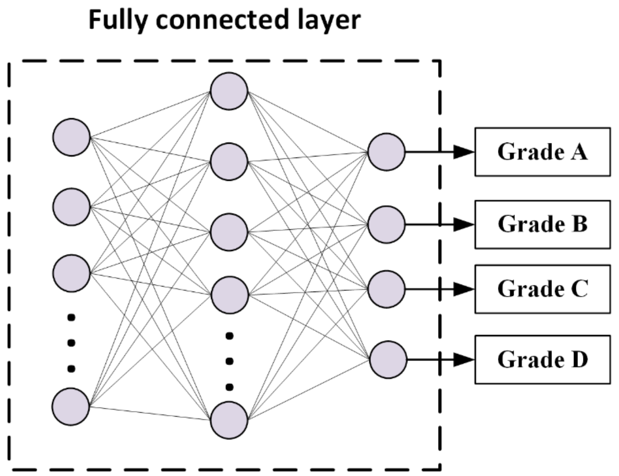


Fig. 3. Schematic diagram of the structure of the fully connected layer.

IV. Results and Discussion

To validate the performance of the I-NOA-CNN algorithm, we conducted simulation experiments based on a dataset of 60 students in five different online courses. The time span of the dataset is 15 weeks, and in this study, we use every three weeks as a time node. The data includes the content of online courses, the amount of homework for online courses, students' class time, the text of student comments on online course platforms, and the frequency of comments. The student's comment text is used for student sentiment analysis.

During the experiment, the number of individuals in the I-NOA algorithm was 20, and the maximum number of iterations was 200. The dropout rate of CNN algorithm is 0.5, and the learning rate is 0.001. Out of the data of 60 students, 15 were used as the training set, while the data labels of the remaining 45 students were used as the testing set. In addition, in order to verify the prediction accuracy and robustness of the I-NOA-CNN algorithm, some recently developed algorithms such as the classic CNN algorithm [21], psoCNN [27], and NOA-CNN were selected for comparison with the I-NOA-CNN algorithm. During the calculation process, each algorithm is independently run 30 times.

A. *Student's Emotional Prediction*

Fig. 4 and 5, respectively show the prediction results of CNN algorithm and I-NOA-CNN algorithm on students' emotions in weeks 15 of Course 1.

From Fig. 4 and 5, we can see that the I-NOA-CNN algorithm has higher accuracy than the CNN algorithm in predicting students' emotions. Table I shows the accuracy of each algorithm's predictions over 30 runs. The I-NOA-CNN algorithm has the highest prediction accuracy, while the CNN algorithm has the lowest. Although the psoCNN and NOA-CNN algorithms have better prediction accuracy than the CNN algorithm, they are far lower than the I-NOA-CNN algorithm.

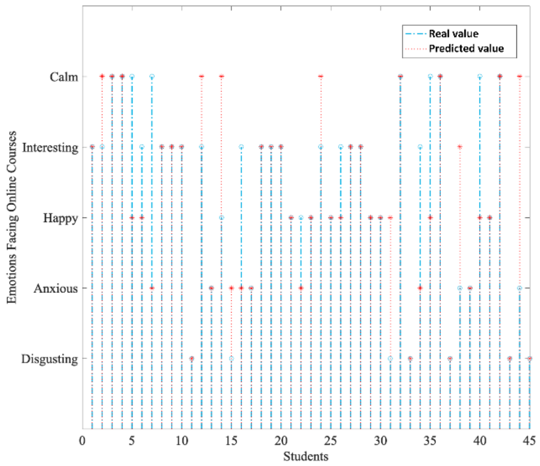


Fig. 4. The results of CNN algorithm for predicting student emotions.

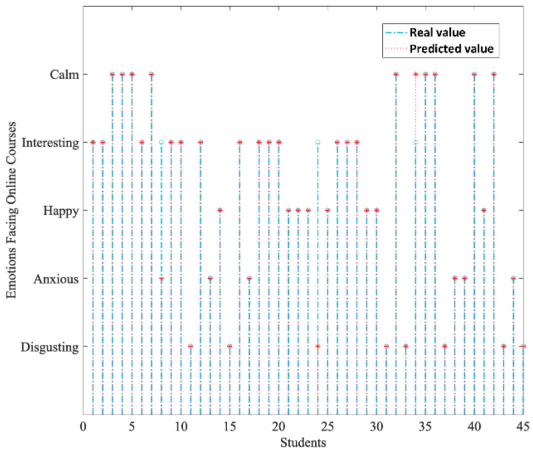


Fig. 5. The results of I-NOA-CNN algorithm for predicting student emotions.

TABLE I. The Accuracy of each Algorithm's Prediction of Student Emotions based on the Data from the 15th Week

Algorithm	Course 1	Course 2	Course 3	Course 4	Course 5
CNN	65.57%	71.47%	75.12%	81.83%	74.28%
psoCNN	80.62%	81.16%	85.42%	82.99%	79.19%
NOA-CNN	85.76%	84.57%	92.69%	83.08%	82.17%
I-NOA-CNN	93.57%	96.41%	97. 59%	96.18%	95.07%

B. Student's Final Grade Prediction

Fig. 6 and 7, respectively show the prediction results of CNN algorithm and I-NOA-CNN algorithm on students' final grades at weeks 15 in course 1. In the final grade prediction process, we divide the final grade into five levels: A, B, C, D, and E.

From Fig. 6 and 7, it can be seen that at week 15, the I-NOA-CNN algorithm had higher accuracy in predicting students' final grades in course 1 than the CNN algorithm. Fig. 8 and 9 respectively show the prediction accuracy of four algorithms in different courses and time nodes during multiple runs. From Fig. 8, it can be seen that the I-NOA-CNN algorithm has good prediction accuracy (>90%) in different courses and has strong generalization ability. From Fig. 9, it can be seen that the I-NOA-CNN algorithm also has the best prediction performance at different time nodes. In addition, due to the expansion of the time range, the amount of data increases, and as the time range expands, the prediction accuracy of each algorithm shows an upward trend.

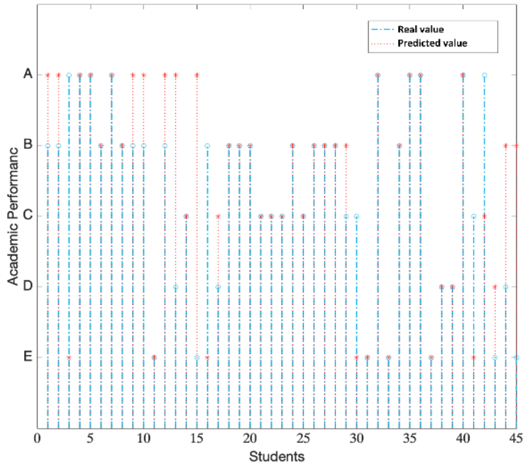


Fig. 6. The prediction results of CNN algorithm on students' final grades.

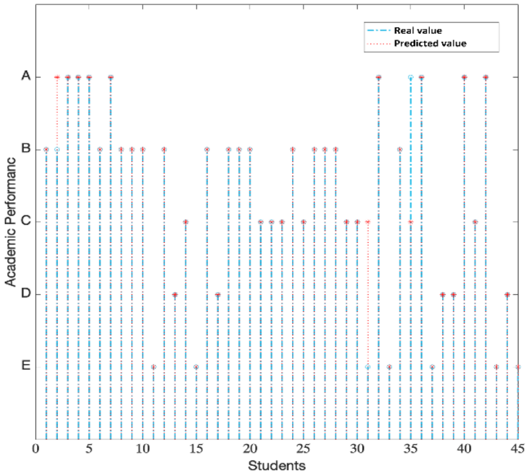


Fig. 7. The prediction results of I-NOA-CNN algorithm on students' final grades.

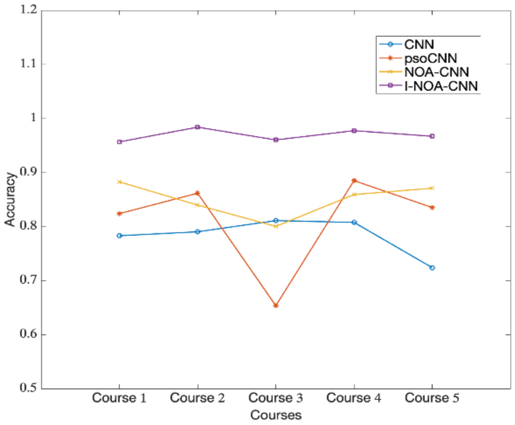


Fig. 8. Different algorithms predict the accuracy of student grades in different course.

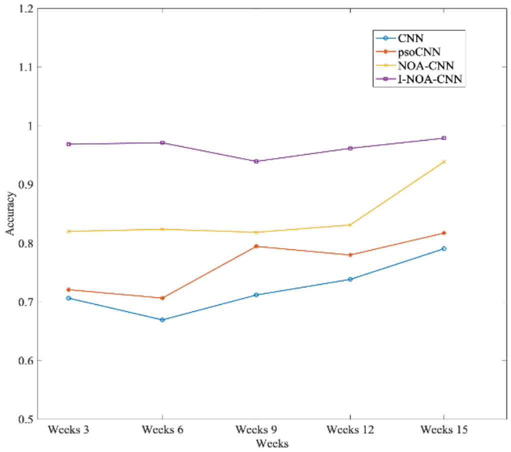


Fig. 9. The accuracy of different algorithms in predicting student grades at different time points.

V. Conclusion

This study focuses on the problem of predicting student grades for online courses and mainly designs an improved CNN algorithm. The grade prediction model mainly consists of two steps. The first step is to predict students' emotions based on an improved CNN algorithm. The second step is to predict students' final grades based on their emotions, course content, and review data. Compared with some of the latest algorithms, this algorithm has higher prediction accuracy and better robustness. The data for this study covers a time span of 15 weeks and does not take into account some short-term online courses. In future research, we will focus on studying the performance prediction of online courses with shorter time spans (less than one week).

References

[1] C. Müller, T. Mildenberger, and D. Steingruber, “Learning effectiveness of a flexible learning study programme in a blended learning design: why are some courses more effective than others?,” International Journal of Educational Technology in Higher Education, vol. 20, no. 1, pp. 10, 2023.

[2] T. Soffer and R. Nachmias, “Effectiveness of learning in online academic courses compared with face-to-face courses in higher education,” Journal of Computer Assisted Learning, vol. 34, no. 5, pp. 534‑543, 2018.

[3] W. Chen, C. G. Brinton, D. Cao, A. Mason-Singh, C. Lu and M. Chiang, "Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors," IEEE Transactions on Learning Technologies, vol. 12, no. 1, pp. 44-58, Jan.-March 2019.

[4] Arul Leena Rose. P. J and Ananthi Claral Mary.T, “An Early Intervention Technique for At-Risk Prediction of Higher Education Students in Cloud-based Virtual Learning Environment using Classification Algorithms during COVID-19” International Journal of Advanced Computer Science and Applications (IJACSA), 13(1), 2022.

[5] E. Martínez-Caro, “Factors affecting effectiveness in e-learning: An analysis in production management courses,” Computer Applications in Engineering Education, vol. 19, no. 3, pp. 572–581, 2011.

[6] Yanli Chen and Ke Jin, “Educational Performance Prediction with Random Forest and Innovative Optimizers: A Data Mining Approach” International Journal of Advanced Computer Science and Applications (IJACSA), 15(3), 2024.

[7] Y. M. I. Hassan, A. Elkorany and K. Wassif, "Utilizing Social Clustering-Based Regression Model for Predicting Student’s GPA," in IEEE Access, vol. 10, pp. 48948-48963, 2022.

[8] G. Tonguc and B. O. Ozkara, “Automatic recognition of student emotions from facial expressions during a lecture,” Computers and Education, vol. 148, pp. 103797, 2020.

[9] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, “What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach,” Computers and Education, vol. 145, pp. 103724, 2020.

[10] D. Baneres, M. E. Rodríguez-González and A. E. Guerrero-Roldán, "A Real-Time Predictive Model for Identifying Course Dropout in Online Higher Education," IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 484-499, Aug. 2023, doi: 10.1109/TLT.2023.3267275.

[11] A. S. Aljaloud et al., "A Deep Learning Model to Predict Student Learning Outcomes in LMS Using CNN and LSTM," IEEE Access, vol. 10, pp. 85255-85265, 2022.

[12] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, “Early prediction of learners at risk in self-paced education: A neural network approach,” Expert Systems With Applications, vol. 213, pp. 118868, 2023.

[13] Mayanda Mega Santoni, T. Basaruddin and Kasiyah Junus, “Convolutional Neural Network Model based Students’ Engagement Detection in Imbalanced DAiSEE Dataset” International Journal of Advanced Computer Science and Applications (IJACSA), 14(3), 2023.

[14] R. Alcaraz, A. Martínez-Rodrigo, R. Zangróniz and J. J. Rieta, "Early Prediction of Students at Risk of Failing a Face-to-Face Course in Power Electronic Systems," IEEE Transactions on Learning Technologies, vol. 14, no. 5, pp. 590-603, Oct. 2021.

[15] A. Gupta, D. Garg and P. Kumar, "Mining Sequential Learning Trajectories With Hidden Markov Models For Early Prediction of At-Risk Students in E-Learning Environments," IEEE Transactions on Learning Technologies, vol. 15, no. 6, pp. 783-797, Dec. 2022.

[16] Z. Xu, H. Yuan and Q. Liu, "Student Performance Prediction Based on Blended Learning," in IEEE Transactions on Education, vol. 64, no. 1, pp. 66-73, Feb. 202.

[17] Z. Wu, L. Huang, Q. Huang, C. Huang, and Y. Tang, “SGKT: Session graph-based knowledge tracing for student performance prediction,” Expert Systems With Applications, vol. 206, pp. 117681, 2022.

[18] Q. Liu et al., "EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 1, pp. 100-115, 1 Jan. 2021.

[19] O. Taylan and B. Karagözoğlu, “An adaptive neuro-fuzzy model for prediction of student’s academic performance,” Computers & Industrial Engineering, vol. 57, no. 3, pp. 732–741, 2009.

[20] J. Nie and H. Ahmadi Dehrashid, “Evaluation of student failure in higher education by an innovative strategy of fuzzy system combined optimization algorithms and AI,” Heliyon, vol. 10, no. 7, pp. e29182, 2024.

[21] X. Lu, Y. Zhu, Y. Xu, and J. Yu, “Learning from multiple dynamic graphs of student and course interactions for student grade predictions,” Neurocomputing, vol. 431, pp. 23–33, 2021.

[22] M. Lohakan and C. Seetao, “Large-scale experiment in STEM education for high school students using artificial intelligence kit based on computer vision and Python,” Heliyon, vol. 10, no. 10, pp. e31366, 2024.

[23] A. Rivas, A. González-Briones, G. Hernández, J. Prieto, and P. Chamoso, “Artificial neural network analysis of the academic performance of students in virtual learning environments,” Neurocomputing, vol. 423, pp. 713–720, 2021.

[24] Y. Song, D. Tang, J. Yu, Z. Yu and X. Li, "Short-Term Forecasting Based on Graph Convolution Networks and Multiresolution Convolution Neural Networks for Wind Power," IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1691-1702, Feb. 2023.

[25] P. S. Naulia, J. Watada and I. A. Aziz, "A Mathematically Inspired Meta-Heuristic Approach to Parameter (Weight) Optimization of Deep Convolution Neural Network," IEEE Access, vol. 12, pp. 83299-83322, 2024.

[26] Z. Zhao and Y. Jiao, "A Fault Diagnosis Method for Rotating Machinery Based on CNN With Mixed Information," IEEE Transactions on Industrial Informatics, vol. 19, no. 8, pp. 9091-9101, Aug. 2023.

[27] J. Tmamna, E. Ben Ayed, R. Fourati, A. Hussain, and M. Ben Ayed, “A CNN pruning approach using constrained binary particle swarm optimization with a reduced search space for image classification,” Applied Soft Computing, vol. 164, pp. 111978, 2024.

[28] M. Omar, F. Yakub, S. S. Abdullah, M. S. A. Rahim, A. H. Zuhairi, and N. Govindan, “One-step vs horizon-step training strategies for multi-step traffic flow forecasting with direct particle swarm optimization grid search support vector regression and long short-term memory,” Expert Systems With Applications, vol. 252, pp. 124154-, 2024.

[29] J. Li, S. Soradi-Zeid, A. Yousefpour, and D. Pan, “Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data,” Applied Soft Computing, vol. 153, pp. 111262, 2024.

[30] Ikpo C V, Akowuah E K, Kponyo J J and Boateng K O, “Ọdigbo Metaheuristic Optimization Algorithm for Computation of Real-Parameters and Engineering Design Optimization” International Journal of Advanced Computer Science and Applications (IJACSA), 14(1), 2023.

[31] M. Abdel-Basset, R. Mohamed, M. Jameel, and M. Abouhawwash, “Nutcracker optimizer: A novel nature-inspired metaheuristic algorithm for global optimization and engineering design problems,” Knowledge-based systems, vol. 262, pp. 110248, 2023.

[32] R. Lin and H. Hu, "Multi-Task Momentum Distillation for Multimodal Sentiment Analysis," IEEE Transactions on Affective Computing, vol. 15, no. 2, pp. 549-565, 2024.

[33] H. Liu, K. Li, J. Fan, C. Yan, T. Qin and Q. Zheng, "Social Image–Text Sentiment Classification With Cross-Modal Consistency and Knowledge Distillation," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3332-3344, 2023.

[34] K. A. Alshaikh, O. A. Almatrafi and Y. B. Abushark, "BERT-Based Model for Aspect-Based Sentiment Analysis for Analyzing Arabic Open-Ended Survey Responses: A Case Study," IEEE Access, vol. 12, pp. 2288-2302, 2024.

[35] P. Durga and D. Godavarthi, "Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN)," IEEE Access, vol. 11, pp. 108433-108447, 2023.

Forensic Facial Reconstruction from Sketch in Crime Investigation

Doaa M. Mohammed, Mostafa Elgendy, Mohamed Taha
Department of Computer Science-Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

***Abstract*—Many crimes are committed every day all over the world, and one of them is a criminal offense that includes a wide range of illegal acts such as murder, theft, assault, rape, kidnapping, fraud, and others. Criminals pose a threat to security, which harms the public interest. In this case, the police question all eyewitnesses at the crime scene, and sometimes, witnesses who were present at the crime scene can remember the face of the criminal. The witness accurately describes the person's facial features in the report, such as eyes, nose, etc. Law enforcement authorities use eyewitness information to identify the person. Criminal investigations can be accelerated by converting sketched faces into actual images, but this requires eyewitnesses to confirm the description in the report. Drawings make it very difficult to identify real human faces because they do not contain the details that help to catch criminals. In contrast, color photographs contain many details that help to identify facial features more clearly. This work proposes to generate color images using the modified modulation Sketch-to-Face CycleGAN and then pass them through Generative Facial Prior-GAN. CycleGAN consists of a generator and discriminator. The generator is used to generate colored images, and the discriminator is used to identify whether the images are real or fake. These are then passed to GFPGAN to improve the quality of the colored images. The structural similarity index measure of 0.8154 is achieved when creating photorealistic images from drawings.**

Keywords—Sketch-to-Face; facial features; Sketch-to-Face CycleGAN; victim's identification; criminal offense

I. Introduction

Crime is a pervasive issue that transcends cultural, geographical, and social boundaries, impacting individuals and communities worldwide. Crime is broadly defined as an act that violates the law and is punishable by the state. It manifests in various forms, from minor offenses to serious felonies. The consequences of criminal activities extend beyond the immediate victim, affecting families, neighborhoods, and society. As crime rates fluctuate, the need for effective crime investigation becomes increasingly critical, serving as a fundamental component of law enforcement and public safety. Crimes can be categorized into several types, including violent crimes, property crimes, white-collar crimes, and cybercrimes, each posing unique dangers to society. Murder and assault are examples of violent crimes. Murders are one of the crimes that have the greatest impact on global societies. They are not just isolated acts of violence but represent a flagrant violation of human rights and a threat to public safety [1]. In the aftermath of a murder, law enforcement's role in identifying suspects is critical to public security and criminal justice operations. This role entails using various legal and technical techniques and tools that aid in accurately and effectively identifying suspected individuals [2]. The role of law enforcement begins with gathering information from multiple sources, such as witnesses, physical evidence, and field investigations. The services of legal drawing experts are used to draw portraits of criminals based on eyewitness descriptions. It entails the process of creating a visual representation of an unidentified person or person of interest using memories and details provided by witnesses or victims [3]. Experts specializing in drawing arts sketch suspected individuals. Forensic artists interview witnesses or victims to obtain accurate details about the person they observed [4]. The information provided may include details related to the suspect's facial features, hair, clothes, tattoos, scars, and any other distinguishing features. Expert forensic artists can transform an accurate description of the human form into a visual image that can be used as evidence in criminal investigations and trials. Forensic artists use manual or computer-based techniques [5]. Drawings are used as evidence that reinforces other available evidence, such as fingerprints, certificates, or physical evidence. When the hand-drawn sketches are completed, they are distributed to the media to facilitate the perpetrator's arrest [6]. The current identification techniques, particularly those relying on forensic sketches, often suffer from inaccuracies due to the subjective nature of eyewitness accounts. These limitations hinder law enforcement's ability to identify and apprehend suspects quickly. This study aims to address these challenges by improving the Structural Similarity Index Measure and the effectiveness of suspect identification through advanced imaging techniques. This study employs modified STF CycleGAN (Sketch-to-Face CycleGAN) in conjunction with GFP-GAN (Generative Facial Prior GAN) to reconstruct images generated by the modified STF CycleGAN model, resulting in a high-fidelity representation of the colored generated image. This paper mainly uses a modified CycleGAN model with GFP-GAN to transform sketch hand-drawn face images into colored real images. Colored images are useful in forensic science and law enforcement, where suspect sketches based on eyewitness statements can be turned into realistic graphics to help law enforcement identify and apprehend criminals. Law enforcement enhances the likelihood of finding missing individuals and criminals. The rest of this work is organized in the following way: Section II is Related Work, and Section III is Methodology. Section IV is Experimental Results. Section V is Discussion. Section VI is Conclusion and Future Work. Finally, References are given.

II. Related Work

We summarize past research on facial reconstruction, utilizing techniques from computer vision and deep learning to reconstruct the face. Shikang Yu et al. introduced a generative adversarial network (GAN) that utilizes the benefits of CycleGAN and conditional GANs to perform face sketch-to-photo transformation [7]. This study presented the development of a novel feature-level loss function, i.e., integrated with the conventional image-level adversarial loss function to enhance the quality of synthesized images. The generator and discriminator provide additional data to supplement the network's training. In addition to the synthetic facial picture, the network's discriminators also receive a genuine image from the other modality as input, with the real image serving as auxiliary information. A feature-level loss is used to assign a penalty to the disparities between a processed image and the original image within the feature space. The generator employed a U-Net architecture [8] because of its efficacy in Pix2Pix, a technique commonly used for picture processing [9]. There are three convolution layers in the discriminator. The CUHK Face Sketch FERET Database (CUFSF) was utilized in conjunction with SSIM 0.5517 [10].

Sreedev Devakumar et al. converted pencil sketches into actual photographs for forensic investigation to ascertain an individual's identity. DCGAN (Deep Convolutional Generative Adversarial Network) transformed the sketch image into a real image [11]. The proposed model comprises a single generator network (G) and two discriminator networks (D1, D2). The photo generator employs convolutional operations on the input sketch to produce images. A patch GAN serves as the initial discriminator, calculating the L1 loss or discriminator loss. As a result, the generator continues to apply this discriminator loss over subsequent epochs. The generator will activate the second discriminator after a total of 30 epochs.

In contrast to the initial scenario, the second discriminator is a conventional discriminator incorporating an additional dense layer. The second classifier establishes a correspondence between the target-generated image and the target photo. The proposed network architecture employs a patch GAN as one of its discriminators. The second option is a conventional discriminator that incorporates an additional dense layer. The conv2D generator employs batch normalization and a Relu layer. The discriminator conv2D observes the application of batch normalization and a leaky Relu layer. The discriminator uses 4x4 filters and a stride of 2 for each convolution. Simultaneously, the generator uses 3x3 filters with a stride value of 2. The G Generator uses an expanded U-Net architecture, employing DCGAN and incorporating six batches of convolutions to enhance the quality of generated images. This task utilized the CUHK Face Sketch FERET Database (CUFSF) dataset, resulting in a final average SSIM of 0.587 [12].

Lidan Wang et al. presented The Photo-Sketch Synthesis using the Multi-Adversarial Networks (PS2-MAN) approach, which involves the iterative generation of low-resolution to high-resolution images in an adversarial manner. This study implements adversarial supervision at all resolution levels. More precisely, the feature maps at each deconvolution layer use 3 x 3 convolutions to provide outputs with varying resolutions. Two generator sub-networks, namely GA and GB, respectively, perform the transformation from photo to sketch and from sketch to photo. The Genetic Algorithm (GA) uses a genuine facial photograph of RA as input and a synthetic (false) representation of FB as output. GB endeavors to convert sketches into photographs. The generator sub-networks objective is to generate images that closely resemble real ones to deceive the discriminator sub-networks. Conversely, the objective of the discriminator sub-networks is to acquire the ability to distinguish between generated and authentic samples. CUFS and CUFSF datasets are commonly used in this study [13]. The CUHK Face Sketch Database (CUFS) is a comprehensive collection of sketches. It consists of 188 faces from the Chinese University of Hong Kong (CUHK) student database, 123 from the AR database, and 295 from the XM2VTS database [14]. The SSIM (picture synthesis) resulted in 0.7915 for picture synthesis and 0.6156 for sketch synthesis [15].

Sparsh Nagpal et al. introduced a novel contextual generative adversarial network designed specifically for sketch-to-image production. The sketch was employed as a lax constraint to identify the most similar mapping and establish our objective function, which comprises a contextual loss and a conventional GAN loss. Additionally, they use a straightforward approach to enhance the start of sketches. This study posits that inpainting produces an image in and of itself. The use of a mask distorts a fraction of the input image. The image, partially obscured by a mask, is called the context. The model attempts to produce a comprehensive representation of this situation. The researchers employed seven up-convolutional layers, each having a kernel size of four and a stride of two. A batch normalizing layer is applied after each up-convolutional layer to expedite the training process and enhance the stability of the learning. Reputation linear unit (Relu) activation is employed at all levels. Ultimately, we employ the tanh function on the output layer. The resulting photos were fed into a layer of pre-trained GFP GANs that were trained using low-quality portrait enhancers. This enhanced image quality and generated more realistic outputs, even when utilizing a less sophisticated model. The dataset utilized the CUHK Face Sketch database (CUFS), the XM2VTS database, and the AR database, with a simultaneous similarity index (SSIM) of 0.78 [16].

Sumit Gunjate et al. introduced a methodology for converting sketches into images using generative adversarial networks. Converting an individual's representation into an image that encompasses the characteristics or attributes associated with the representation necessitates the utilization of many categories of machine learning algorithms. The models comprised a GAN discriminator, which functions just as a classifier. The task involves distinguishing between factual data and data produced by the generator. The generator component of a generative adversarial network (GAN) acquires the ability to generate false data. It acquires the capability to designate its output as genuine. In contrast to the process of discriminator training, generator training requires a more advanced level of integration between the generator and the discriminator. The generator model used the U-Net architecture, which consists of encoder and decoder layers. The model takes a picture (the sketch) as input and applies 7 encoding layers with filters (C64, C128, C256, C512, C512, C512, C512, C512, C512) to make the image smaller in size. The encoding layers employ batch normalization except for the first layer (C64). Their research aims to determine whether the generated facial structures if deemed credible, share the same identity label as authentic faces. The researchers specifically designed a light CNN to extract identity-preserving features and utilized the L2 norm for comparison purposes [17].

III. Methodology

This research suggests using a modified STF CycleGAN model output as an input to the GFP-GAN model to obtain the result, as shown in Fig. 1.

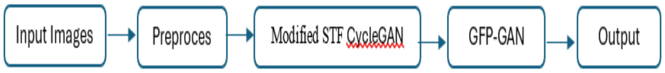


Fig. 1. The architecture used in this paper.

A. *Preprocess*

The technique entails analyzing and utilizing sketches and target images in RGB color. The Gaussian blur feature is achieved by applying a Gaussian function to a picture to reduce noise and generate a more refined visual effect. The filter described above can be considered a nonuniform low-pass filter [19], efficiently preserving low spatial frequency while simultaneously reducing image noise and inconsequential details inside an image. The sketches were subjected to Gaussian blur with a blur radius of 0.3, while the color images were exposed to a blur radius of 0.8. The calculation of Gaussian blur is derived from (1) [18]-[19] determines the Gaussian blur calculation.

$$G_{\sigma}(x) = \frac{1}{2\pi\delta^2} e^{-(x^2+y^2)/2\delta^2} \qquad (1)$$

Following the noise reduction process, we employed rescaling augmentation on the photos from the Chuck database with various scales. As a result, 2256 images were generated, including both sketch and ground truth images.

B. *Modified STF CycleGAN Model*

CycleGAN is a kind of Generative Adversarial Network (GAN) architecture built primarily for challenges involving the translation of unpaired images. Jun-Yan Zhu et al. introduced it in their 2017 publication [7]. CycleGAN enables the acquisition of mappings between distinct domains, facilitating the transformation of images from one domain to another without necessitating the inclusion of paired data samples from both domains during the training process. CycleGAN can learn image translation across different domains, even without paired data, unlike typical supervised learning approaches that require each input image to be matched with its corresponding destination image. Our objective is to convert prototype photos into their corresponding actual ones. The model comprises two fundamental components, namely a Generator and a Discriminator.

1) Generator: The generator is a neural network, i.e. tasked with translating images from one domain to another. The generator functions unsupervised, enabling it to execute the translation task without relying on paired data samples throughout the training process. The generator comprises convolutional neural network (CNN) layers that receive an input image from one domain and generate a corresponding output image in the target domain. The input image size is [256, 256, 3] and undergoes a sequence of downsampling and upsampling modules.

a) Down-sampling: Downsampling is a technique used to decrease the spatial resolution of an input image, usually achieved by employing convolutional procedures like step convolutions. One essential operation in the generator network is down-sampling, which extracts hierarchical information from the input image and aids in the translation process. The down-sampling blocks progressively decrease the resolution from 256x256 to 1x1 spatial dimensions. Convolution layers, batch normalization, and the Leaky-Relu activation function are employed in the down-sampling process.

- Convolution Layer: Convolutional layers, sometimes called conv layers, are essential components of convolutional neural networks (CNNs). Convolutional layers are of utmost importance in the process of extracting features from input data through the application of convolutional operations. In this context, the weights assigned to each filter function represent a vocabulary of feature patterns. Filters used are (64, 128, 256, 512, 512, 512, 512, 512) with kernel size 4. As calculated in (2) [20], the convolution operation is computed for the input patch X and the filter kernel K, which determines the convolution layer's calculation. Fig. 2 illustrates the convolution operation.

$$C\,(m, n) = \sum_{p=0}^{p} \sum_{q=0}^{q} K(p;\, q) \, * \, X\,(m + p\,, n + q) + b \qquad (2)$$

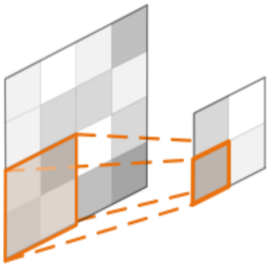


Fig. 2. Convolution operation.

- Batch Normalization: Batch normalization is a technique often used in deep neural networks, particularly in the convolutional and fully connected layers, to improve the training process and the model's overall performance. People widely recognize batch normalization as a prominent technique that enhances network training speed and improves accuracy. [21] Batch normalization changes the data distribution to have a mean of 0 and a variance of 1, expressed in units of the minibatch. This approach is effective in mitigating the possibility of overfitting [22].
- Leaky-Relu Activation Function: The Leaky Rectified Linear Unit (Leaky-Relu) is a frequently employed activation function in neural networks, specifically within deep learning architectures. The proposed approach is an expansion of the conventional Rectified Linear Unit (Relu) function, which aims to overcome certain constraints, notably the issue of "dying Relu" This difficulty arises when neurons that produce a weight of zero for all inputs cease to update their weights during

the training process. The Leaky-Relu has a small gradient (usually a small positive value, like 0.01) for negative inputs. This lets them spread a gradient throughout the backpropagation process. Fig. 3 illustrates the Leaky-Relu activation function. The Leaky Relu activation function is defined in (3) [23], which specifies how the function is calculated.

$$Leakly - Relu(x) = max(kx, x) = \begin{cases} x, & \text{if } x > 0 \\ kx, & x \leq 0 \end{cases} Leakly - Relu(x) = max(kx, x) = \begin{cases} x, & \text{if } x > 0 \\ kx, & x \leq 0 \end{cases} \quad (3)$$

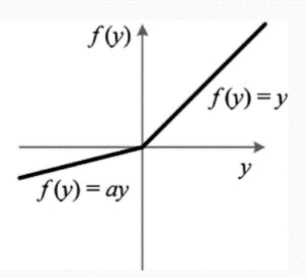


Fig. 3. Leaky-Relu activation function.

b) Up-sampling: Upsampling is a technique used to increase the spatial resolution of feature maps. It is usually achieved using techniques such as transposed convolutions. Upsampling plays a crucial role in the generator network by efficiently collaborating with downsampling to convert images across different domains. A 2D transposed convolutional layer, batch normalization, and the Rectified Linear Unit (Relu) activation function are used in upsampling to expand the spatial dimensions. The generator network efficiently produces images with suitable characteristics, and the ultimate output has dimensions of 256x256x3, corresponding to color images.

- Deconvolution Layer: Deconvolution, also known as transposed convolution, is a method that increases the sampling level of feature maps while preserving the connectivity pattern. The deconvolutional layers employ convolution-like procedures with several filters to expand and densify the input. The filters used are (512, 512, 512, 512, 256, 128, 64, 32). Unlike the current resizing methods, the deconvolution process includes adjustable parameters, as shown in Fig. 4 [20]. During network training, the weights of deconvolutional layers undergo continuous updates and refinements. On the input side, the process starts with adding zeros between neurons in the receptive field, and then a convolution kernel with a unit stride is used at the very top.

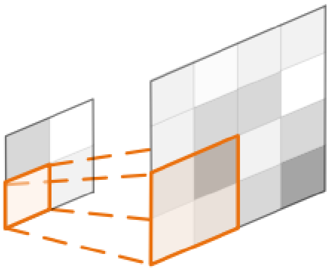


Fig. 4. Deconvolution operation.

- Relu Activation Function: The Rectified Linear Unit (Relu) is a highly prevalent activation function commonly employed in neural networks, particularly in deep learning models. Relu's linear operation results in a higher convergence rate and calculation speed. The neural network's performance is negatively impacted by the presence of dead neurons, which Relu causes. Negative input values result in a constant output of zero for the Rectified Linear Unit (Relu) function, and consequently, the first derivative is also zero. This renders the neuron incapable of updating its parameters, as shown in Fig. 5. The Relu activation function is calculated in (4) [24], which specifies how it is calculated.

$$f(x)=max (0, x) \quad (4)$$

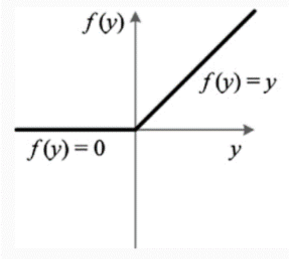


Fig. 5. Relu activation function.

c) Skip connections: Skip connections, often referred to as residual connections, serve a vital function in the generator, easing the transmission of information across several layers and enhancing the overall quality of the generated images. The establishment of these links serves to mitigate the issue of information loss that arises during the translation process, hence enabling the generator to retain intricate details within the input images more effectively. The generator's skip connections facilitate the preservation of low-level details and fine-grained information from the input image. Skip links enable the network to access both the original input and the higher-level information acquired during the translation process by directly transmitting the input from a certain layer to a subsequent layer.

2) Discriminator: The discriminator is an essential component of the adversarial training process and plays a critical role. The discriminator has a crucial function in differentiating between genuine images from the desired domain and artificial images produced by the generator. It offers input for the generator network to enhance the quality of its generated images. The following items are included: Downsampling blocks are employed to minimize spatial dimensions. We use filters (32, 128, 256, 512) with a filter size of 4, batch normalization, Leaky-Relu, zero padding, and a convolutional layer with only one output channel.

3) Modified STF CycleGAN loss: Modified STF CycleGan loss consists of two losses.

a) Generator loss function: The generator loss function consists of the following components:

- Adversarial loss (GAN loss): This causes the generator to generate images that are so similar to the actual ones that they cannot be distinguished. The formulation employs binary cross-entropy loss. We use the Adversarial loss to calculate the min-max loss. Adversarial loss is calculated in (5).

$$L_{Gan} = E_x[\log(D(x))] + E_z[\log(1 - D_x(G(x)))] \quad (5)$$

- L1 loss is a loss term that ensures the generated and target images have pixel-wise similarity.

To calculate the overall generator loss, we combine these components, using weights that determine the relative significance of each loss item. The generator loss function is calculated in (6).

$$L_{Gen} = L_{GAN} + \lambda 1 \cdot L_{cyc}(G) \quad (6)$$

Where $\lambda 1$ is the weighting factor for the L1 loss, and L_{GAN} represents the adversarial loss, L_{cyc} represents cycle consistency loss. L_{cyc} is calculated in (7).

$$L_{cyc} = E \sim p_{data}(x)\ [\|G(x) - x\|] \quad (7)$$

b) Discriminator loss function: The discriminator loss function consists of two distinct components:

- The discriminator's ability to accurately categorize authentic photos is referred to as a real loss.
- Generated Loss: This metric measures the discriminator's accuracy in correctly categorizing generated images.

The discriminator loss function is calculated in (8).

$$L_{disc} = L_{real} + L_{generated} \quad (8)$$

where L_{real} is the real loss term, and $L_{generated}$ is the generated loss

C. *GFP-GAN Model*

Generative Facial Parsing (GFP) is a specific adaptation of generative adversarial networks (GANs).

The modified STF CycleGAN model generates images that are input into the GFP-GAN model. The GFP-GAN can be used to fix images by identifying small imperfections, such as scratches and blemishes. Fig. 6 shows the inpainting techniques used to fix the damaged regions [25]. The face restoration method uses GFP with innovative channel-split spatial feature transform layers, striking an optimal balance between authenticity and accuracy. The GFP-GAN successfully upgraded colors and restored facial details with a single forward pass, owing to its sophisticated designs and robust generative facial prior. Simultaneously, GAN inversion techniques necessitate costly image-specific optimization [26].

D. *Performance Evaluation*

We employed modified STF CycleGAN neural networks, which were trained and tested on a dataset consisting of human sketches paired with their corresponding images. To assess the similarity between the generated image and the original image, we utilize the following parameters to evaluate the performance.

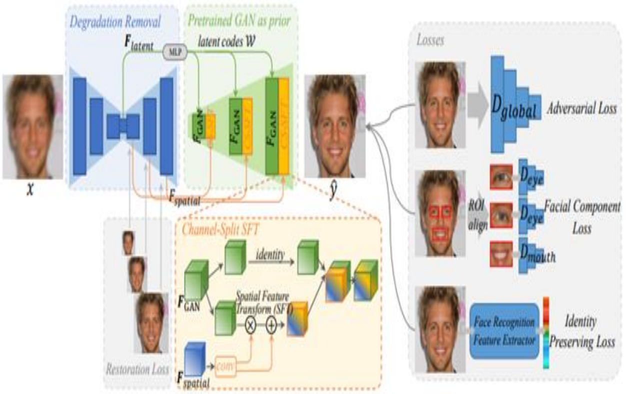


Fig. 6. GFP-GAN framework [26].

The SSIM (Structural Similarity Index Measure) is a technique used to estimate the perceived quality of photos, digital images, and videos [27]. SSIM is mostly employed to quantify the resemblance between two photographs. The SSIM values range from 0 to 1, with a value of 1 indicating a perfect match between the reconstructed image and the original image. The SSIM metric determines the degree of alignment or similarity in pixel density values between two photographs. The calculation is performed in the following manner:

$$\text{SSIM}(x, y) = \frac{(2\,\mu_x\mu_y + C1)(2\sigma_{xy} + C2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)} \quad (9)$$

Where µx the mean of x

- μ_y the mean of y

- σ_x^2 the variance of x

- σ_y^2 the variance of y

- σ_{xy} the cross-correlation of x and y

- c1, c2 Constants are added into the formula to provide numerical stability, particularly when the denominator terms (variances and covariances) are close to zero.

IV. Experimental Result

The system operates on the Google Colab TPU runtime, which comprises an Intel Xeon CPU operating at a frequency of 2.30 GHz, 13 GB of RAM, and a cloud TPU with a computing capacity of 180 teraflops. The computer vision system operates on Python 3.10 and utilizes OpenCV.

A. *Dataset*

The dataset used in this study is the CUHK Face Sketch database (CUFS) for research on face sketch synthesis and recognition. It includes 188 faces from the Chinese University of Hong Kong (CUHK) student database, 123 from the AR database, and 295 from the XM2VTS database. The images were taken during a video recording session. In total, there are 606 faces. For each face, an artist draws a sketch based on a photo taken in a frontal pose under normal lighting conditions with a neutral expression. The CUHK Face Sketch FERET Database (CUFSF) is used for face sketch synthesis and recognition research. It includes 1,194 individuals from the FERET database. The augmentation technique involves rescaling the photos in the Chuck database, resulting in 2256 images for both the sketch and target images. Combining sketches and photographs allows for more realistic training as the model learns to translate artistic representations into lifelike images, crucial for forensic applications.

B. Experiment

We partitioned the dataset into two subsets, allocating 80% for training and 20% for testing the modified STF CycleGAN model. The input for the modified STF CycleGAN model is an image with a shape of (256, 256, 3). We then downsample this picture using filters of sizes (64, 128, 256, 512, 512, 512, and 512). Additionally, the image is upsampled using filters of sizes 512, 512, and 512, with dropout applied at a rate of 0.5. The size of the filter employed is 4. The modified STF CycleGAN model's discriminator compresses the input image with a shape of (256, 256, 3) using filters (32, 128, 256, 512). We train the model on the dataset for a total of 100 epochs. The model's SSIM yielded a value of SSIM is 0.637. Fig. 7 displays the outcome of the modified STF CycleGAN model.

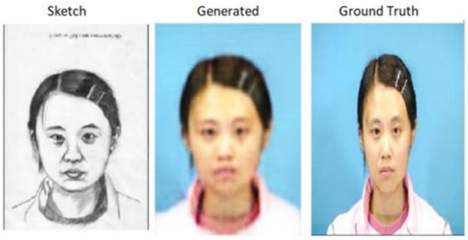


Fig. 7. Result of modified STF CycleGAN model.

The dataset has been partitioned into two subsets, with 80% allocated for training and 20% for testing. The model is trained for 500 epochs using the dataset. The model yielded a result of SSIM is 0.8154 when the generated images from modified STF CycleGAN were used as input for GFP-GAN to recover the images and produce high-quality output images. Fig. 8 displays the outcome of the modified STF CycleGAN algorithm when applied to the GFP-GAN model.

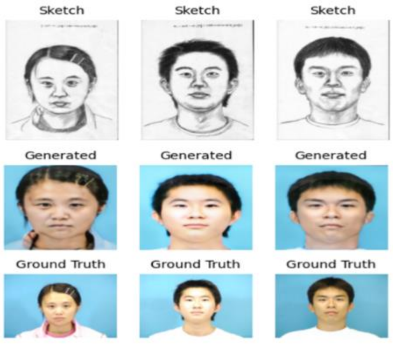


Fig. 8. Result of modified STF CycleGAN with GFP-GAN model.

V. Discussion

In this study, we developed a deep-learning model using modified STF CycleGAN with GFP-GAN to generate realistic faces from hand-drawn sketches, which helps law enforcement authorities identify and arrest criminals. The model outperformed traditional methods, achieving higher visual accuracy and preserving key facial features with an average SSIM score of 0.8154. The system uses CUFS, CUFSF, AR, and XM2VTS are public and available datasets. We use a modified STF CycleGAN model to acquire high-level facial representations and enhance the quality of the generated images. However, the model faced challenges when processing incomplete or low-quality sketches, and it struggled with extreme facial poses or expressions, highlighting the need for more diverse training data. Our system is better than other models GAN and DCGAN [10, 12]. It also outperforms PS2-MAN and contextual generative adversarial networks with GFP GANs [15, 16]. Despite the limitations, this model has practical potential in fields like forensic sketching and digital art, where accurate facial generation from sketches is essential. Developing a sketch-to-face generation system has promise for advancing techniques in forensic sketching, character design, and digital art, with potential applications in law enforcement and personalized creative tools. This technology could revolutionize how hand-drawn sketches are used in practical settings, providing more accurate visualizations for identification, design, and artistic expression.

TABLE I. Performance Comparison of Different Models

	Method	Dataset	SSIM
[10]	Generative Adversarial Network (GAN)	(CUFSF)	0.5517
[12]	DCGAN (Deep Convolutional Generative Adversarial Network)	(CUFSF)	0.587
[15]	Multi-adversarial networks (PS2 - MAN)	(CUFS), AR database, and the XM2VTS database	SSIM (Photo Synthesis) is 0.7915, SSIM (Sketch Synthesis) is 0.6156
[16]	contextual generative adversarial network with GFP GANs	(CUFS), the XM2VTS database, and the AR database	0.78
Ours	modified STF CycleGAN	(CUFS), the XM2VTS database, and the AR database	0. 637
Ours	modified STF CycleGAN with GFP-GAN	(CUFS),(CUFSF), the XM2VTS database, and the AR database	0. 8154

VI. Conclusion and Future Work

In conclusion, this research paper presents a significant advancement in forensic facial reconstruction by utilizing a modified STF CycleGAN that effectively transforms hand-drawn sketches into high-fidelity colored images, enhancing the identification process of suspects in criminal investigations. Additionally, using GFP-GAN enhances the quality of images generated from modified STF CycleGAN. The study uses the CUHK Face Sketch database (CUFS) and other datasets such as CUHK, CUFSF, FERET, XM2VTS, and AR. The study involves assembling a comprehensive collection of 2,256 images for training and testing the model. The modified STF ycleGAN achieves a Structural Similarity Index Measure (SSIM) of 0.8154, demonstrating a high level of reconstructioned image compared to previous models. Future directions include further exploring the integration of modified StyleGAN with GFP-GAN to improve the Structural Similarity Index Measure and quality of generated images. Overall, this research underscores the potential of advanced generative models in aiding law enforcement and enhancing the efficiency of criminal investigations.

REFERENCES

[1] B. Cao, N. Wang, J. Li, Q. Hu, and X. Gao, "Face photo-sketch synthesis via full-scale identity supervision," *Pattern Recognition*, vol. 123, pp. 107654, Apr. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321006221

[2] G. Grana and J. Windell, *Crime and intelligence analysis*. 2021. doi: 10.4324/9781003005346.

[3] IEEE Journals & Magazine, "Face sketch recognition," *IEEE Xplore*, vol. 92, no. 1, pp. 34-45, Jan. 2004. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1262031/

[4] N. M. Farid, M. S. Fard, and A. Nickabadi, "Face sketch to photo translation using generative adversarial networks," arXiv preprint arXiv:2110.12290, Oct. 23, 2021. [Online]. Available: https://arxiv.org/abs/2110.12290

[5] A. Aglasia, A. Adimas, S. Y. Irianto, S. Karnila, and D. Yuliawati, "Image Sketch Based Criminal Face Recognition Using Content Based Image Retrieval," *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 177-188, 2021.

[6] IEEE Conference Publication, "Face photo recognition using sketch," *IEEE Xplore*, 2002. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1038008/

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Lecture Notes in Computer Science*, vol. 9351, pp. 234-241, Jan. 2015. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html

[10] IEEE Conference Publication, "Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation," *IEEE Xplore*, May 01, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8756563/

[11] IEEE Conference Publication, "Crime Investigation using DCGAN by Forensic Sketch-to-Face Transformation (STF)- A Review," Apr. 8, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9418417/

[12] S. Devakumar and G. Sarath, "Forensic Sketch to Real Image Using DCGAN," *Procedia Computer Science*, vol. 203, pp. 101-108, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923001394

[13] IEEE Journals & Magazine, "Face Photo-Sketch Synthesis and Recognition," *IEEE Xplore*, vol. 97, no. 11, pp. 105-112, Nov. 2009. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4624272/

[14] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. Second Int. Conf. Audio and Video-based Biometric Person Authentication*, vol. 964, pp. 965-966, 1999.

[15] IEEE Conference Publication, "High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks," *IEEE Xplore*, May 1, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8373815/

[16] S. Nagpal, "Sketch-to-Face Image Translation and Enhancement Using a Multi-GAN Approach," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 10, no. 12, pp. 111-119, Dec. 2022.

[17] S. Gunjate, T. Nakhate, T. Kshirsagar, Y. Sapat, and S. Guhe, "Sketch to Image Using GAN," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 1, pp. 45-52, Jan. 2023.

[18] IEEE Journals & Magazine, "A Generalized Laplacian of Gaussian Filter for Blob Detection and Its Applications," *IEEE Xplore*, vol. 101, no. 12, pp. 1356-1365, Dec. 2013. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6408211/

[19] S. Misra and Y. Wu, "Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking," in *Elsevier eBooks*, vol. 35, pp. 205-221, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/B9780128177365000107

[20] Y. Yang, W. Zhang, J. Wu, W. Zhao, and A. Chen, "Deconvolution-and-Convolution Networks," arXiv preprint arXiv:2103.11887, Mar. 22, 2021. [Online]. Available: https://arxiv.org/abs/2103.11887

[21] IEEE Conference Publication, "Batch Normalization in Convolutional Neural Networks — A comparative study with CIFAR-10 data," Jan. 1, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8470438

[22] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, "Deep learning with convolutional neural network in radiology," *Japanese Journal of Radiology (Print)*, vol. 36, no. 3, pp. 256-269, Mar. 2018. [Online]. Available: https://link.springer.com/article/10.1007/s11604-018-0726-3

[23] IEEE Conference Publication, "Reluplex made more practical: Leaky ReLU," Jul. 1, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9219587/

[24] D. Ak and V. Jain, "Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions," in *Lecture Notes in Electrical Engineering*, vol. 539, pp. 432-439, Jan. 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-13-6772-4_76

[25] A. Kumari, R. K. Dubey, and S. K. Mishra, "A Cascaded Method for Real Face Image Restoration using GFP-GAN," *International Journal of Innovative Research in Technology and Management*, vol. 6, pp. 23-32, 2022.

[26] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration With Generative Facial Prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2021/html/Wang_Towards_Real-World_Blind_Face_Restoration_With_Generative_Facial_Prior_CVPR_2021_paper.html

[27] M. Chen and A. C. Bovik, "Fast structural similarity index algorithm," *Journal of Real-Time Image Processing*, vol. 6, no. 4, pp. 243-254, Aug. 2010. [Online]. Available: https://link.springer.com/article/10.1007/s11554-010-0170-9.

Artificial Intelligence-Driven Decision Support Systems for Sustainable Energy Management in Smart Cities

Ning MA
Shanxi Vocational University of Engineering Science and Technology, Jinzhong 030619, China

***Abstract*—Due to the ongoing urbanization trend, smart cities are critical to designing a sustainable future. Urban sustainability involves action-oriented approaches for optimizing resource usage, ecological impact reduction, and overall efficiency enhancement. Energy management is one of the main concerns in urban, residential, and building planning. Artificial Intelligence (AI) uses data analytics and machine learning to instigate business automation and deal with intelligent tasks involved in numerous industries. Thus, AI needs to be considered in the strategic plan, especially in the long-term strategy of smart city planning. Decision Support Systems (DSS) are integrated with human-machine interaction methods like the Internet of Things (IoT). Along with their growth in size and complexity, the communications of IoT smart devices, industrial equipment, sensors, and mobile applications present an increasing challenge in meeting Service Level Agreements (SLAs) in diverse cloud data centers and user requests. This challenge would be further compounded if the energy consumption of industrial IoT networks also increased tremendously. Thus, DSS models are necessary for automated decision-making in crucial IoT settings like intelligent industrial systems and smart cities. The present study examines how AI can be integrated into DSS to tackle the intricate difficulties of sustainable energy management in smart cities. The study examines the evolution of DSSs and elucidates how AI enhances their functionalities. The study explores several AI methods, such as machine learning algorithms and predictive analytics that aid in predicting, optimizing, and making real-time decisions inside urban energy systems. Furthermore, real-world instances from different smart cities highlight the practical applications, benefits, and interdisciplinary collaboration necessary to successfully implement AI-driven DSS in sustainable energy management.**

Keywords*—Smart cities; artificial intelligence; decision support systems; sustainable energy management; urban resilience; interdisciplinary collaboration

I. Introduction

Urbanization continues to rise at an unprecedented rate, resulting in an increasing proportion of the global population residing in urban centers [1, 2]. This rapid urban expansion has led to significant ecological impacts, including alterations in local and regional climates, loss of biodiversity, and disruption of natural habitats [3]. These consequences highlight the pressing need for sustainable urban development that prioritizes ecosystem preservation and restoration [4]. As urban areas grow, so do the demands on critical infrastructure, particularly energy systems, which play a crucial role in sustaining urban life. Effective energy management within cities is paramount to reducing emissions, ensuring energy efficiency, and fostering a sustainable future [5].

A. *Research Problem and Motivation*

The energy demands of modern cities are becoming increasingly complex, especially with the rise of smart cities—urban areas where digital technologies and data-driven systems are integrated into infrastructure and governance [6]. Smart cities aim to improve the quality of life for their residents by optimizing resource usage, reducing environmental impact, and enhancing efficiency across sectors such as transportation, waste management, and energy. However, as smart cities evolve, they face significant challenges in managing energy efficiently, particularly with the growing integration of Internet of Things (IoT) devices, which contribute to fluctuating energy demands and intricate data streams [7]. These challenges necessitate advanced solutions that can optimize energy management while ensuring the sustainability of urban systems.

Artificial Intelligence (AI) has emerged as a transformative technology capable of addressing these complex challenges by enabling real-time data analysis, predictive modelling, and automation [8, 9]. Decision Support Systems (DSS) enhanced by AI have the potential to revolutionize energy management in smart cities by providing intelligent, data-driven decision-making capabilities. Despite the recognized potential of AI and DSS, there remains a gap in understanding how these technologies can be effectively integrated to manage energy in smart cities while addressing the broader sustainability goals [10].

B. *Objective and Contribution*

The primary motivation of this research is to explore how AI-driven DSS can be leveraged to improve energy management in smart cities, thus addressing the intricate challenges posed by urbanization and technological integration. This study aims to investigate AI methods such as machine learning and predictive analytics, focusing on how these techniques can optimize energy use, predict demand, and support real-time decision-making in urban environments. By analyzing the current state of AI and DSS in smart city contexts, this research highlights the innovative potential of interdisciplinary collaboration in tackling the complex issues of sustainable energy management.

This paper enhances existing literature with a comprehensive discussion of the role AI-driven DSS can play in energy management in smart cities. It further examines the interaction between AI, IoT, and energy systems and proposes new models for optimizing energy usage. The present research thus lays the foundation for further practical implementation and studies in real-world applications and challenges that meet the critical demand for sustainable energy management in an increasingly complex urban environment.

II. Background

Advanced energy management involves using technology to control energy consumption and production, distribution, and consumption in highly energy-demanding cities to foster sustainable economic development [11]. However, such a transition comes with several challenges, as outlined in Table I.

One of the major issues is the ever-increasing difference between supply and demand for energy resources. Growth in urban areas increases the demand for energy since energy consumption rises with cities' expansion, which puts pressure on existing resources [12]. The gap was bridged in the past by producing more energy and emitting more carbon despite negative environmental impacts. Also, poorly maintained and relatively old energy networks in urban areas result in large amounts of energy lost in transmission and distribution. Additionally, the lack of real-time energy monitoring and management systems can prevent immediate detection and correction of inefficiencies. Finally, energy prices remain more or less volatile, and sufficient capital is not readily available to upgrade energy technologies and accommodate innovations in the market.

Incorporating sustainable and cost-effective renewable energy sources like solar, wind, and geothermal power is critical for reducing reliance on traditional energy sources and their carbon footprint [13]. Smart grids employ advanced sensors, communication technologies, and intelligent algorithms to enable real-time monitoring and management of energy distribution across the grid. This optimization minimizes waste and maximizes efficiency. The development of cost-effective and efficient energy storage systems is essential. These systems store excess energy generated from renewable sources, ensuring a reliable and consistent energy supply even during peak demand periods.

As buildings are major energy consumers in cities, developing energy-efficient buildings with intelligent lighting and HVAC systems is crucial for optimizing energy use and reducing waste [14]. Electric Vehicles (EVs) present a significant opportunity to improve air quality and reduce city carbon emissions [15]. Developing EV charging infrastructure and integrating renewable energy sources to power these vehicles are key aspects of smart energy management.

Smart homes, equipped with internet-connected devices that control and automate temperature, lighting, security, and entertainment, offer convenience, improved energy efficiency, and enhanced quality of life for residents [10]. The utilization of data analytics in smart energy management is crucial. This enables the collection, analysis, and interpretation of energy consumption data, facilitating the identification of inefficiencies, optimization of energy use, and cost reduction [16]. Integrating these technological advancements, smart cities can optimize energy production, distribution, and consumption. This paves the way for creating more sustainable, efficient, and livable urban environments.

TABLE I. Key Challenges and Solutions Related to Smart Energy Management in Urban Environments

Aspect	Challenges	Solutions
Energy supply and demand	Continuous rise in energy consumption due to urban expansion	Incorporation of renewable energy sources like solar, wind, and geothermal power to reduce carbon footprint
Infrastructure efficiency	Outdated and inefficient energy infrastructure causing energy losses during transmission and distribution	Development of smart grids with advanced sensors and intelligent algorithms for real-time energy management
Real-time monitoring	Absence of real-time energy monitoring and management systems	Smart grids are used for real-time monitoring and management to minimize waste and maximize efficiency
Economic factors	Fluctuating energy prices and limited funding for infrastructure upgrades	Development of cost-effective and efficient energy storage systems to ensure reliable energy supply
Energy-efficient buildings	High energy consumption by buildings	Development of energy-efficient buildings with intelligent lighting and HVAC systems to optimize energy use
Electric vehicles	Need for improved air quality and reduced carbon emissions	Develop an EV charging infrastructure and integrate renewable energy sources to power EVs
Smart homes	Need for improved energy efficiency and quality of life	Equipping homes with internet-connected devices for automated temperature control, lighting, and security

III. AI-Powered Decision Support Systems for Energy Management

This section comprehensively compares AI-powered DSS for energy management. Table II summarizes various approaches and their key features, challenges addressed, methodologies, applications, and outcomes based on recent studies.

Panagoulias, et al. [17] introduced a novel development methodology for AI-powered analytics in energy management. This approach emphasizes tailored explainability, addressing the inherent lack of transparency ("black box") often associated with AI systems. It acknowledges that various stakeholders with diverse backgrounds, preferences, and goals will utilize these analytics. The methodology aligns with the Explainable Artificial Intelligence (XAI) paradigm, aiming to enhance the interpretability of AI-driven DSSs. A core feature is a clustering-based approach that customizes the level of explanation based on the specific needs of user groups. This customization fosters accuracy and effectiveness in energy management analytics while promoting transparency and trust in decision-making.

TABLE II. COMPARISON OF AI-POWERED DECISION SUPPORT SYSTEMS FOR ENERGY MANAGEMENT

Studies	Focus	Key features	Challenges addressed	Methodology	Applications	Outcomes
Panagoulias , et al. [17]	AI-powered analytics with explainability tailored to stakeholder needs	Explainable AI (XAI) and clustering-based explanation customization	Transparency and trust in AI-driven DSS	• Iterative development lifecycle • Stakeholder identification	Estimation of energy savings from building renovations	Higher adoption rates of AI systems
Selvaraj, et al. [18]	AI Technique for Monitoring Systems in Smart Buildings (AIMS-SB)	Prediction model methodologies and energy analysis, renewable energy generation, recycling assessment	Poor energy recycling, high consumption, suboptimal usage, and drain characteristics	• Eco-design monitoring systems • Effective control of energy consumption and generation	Intelligent building energy management	Enhanced precision and effectiveness compared to traditional approaches
Şerban and Lytras [19]	AI in the renewable energy (RE) sector in the EU	Conversion processes of RE and Impact of AI on RE industry and smart cities research	Development and resilience of the energy sector	• A conceptual framework for AI's impact on RE • Examination of RE efficiency from Gross Inland Consumption to final energy consumption	AI adoption trends for RE in the EU	Improved efficiency and sustainability of the RE sector
Chen, et al. [20]	IoT framework for energy-efficient street lighting	IoT sensor-equipped smart electric poles and LED bulbs with a mesophilic design	High energy consumption and inefficiency	• Intelligent decision-making based on traffic flow and occupancy data • Dynamic battery charging algorithm based on MPPT	Smart street lighting for highways, residential, and suburban pedestrian zones	Significant reduction in energy consumption and carbon emissions
Manman, et al. [21]	Cognitive Radio Sensor Networks (CRSNs) for sustainable IoT applications	Distributed artificial intelligence and cooperative communication for resource management	Efficient resource allocation and sustainability in smart city applications	• Real-time computation of resource allocation using DAI • Statistical behavior-based relationship between secondary users in clusters	Enhanced sustainability of intelligent world IoT applications	Increased energy efficiency and sustainability in smart city applications
Singh, et al. [22]	Blockchain and AI integration in smart cities for sustainability	Blockchain for risk management, IoT, financial services, and public services and analysis of security vulnerabilities in blockchain	Security vulnerabilities and challenges in blockchain implementation	• Detailed exploration of the convergence of blockchain and AI • Analysis of critical factors for successful integration of blockchain and AI	Intelligent transportation systems leveraging blockchain and AI	Paving the way for more sustainable smart cities
Li, et al. [23]	IoT and AI-assisted Smart Metering Systems (IoT–AI–SMS)	RNN model for load forecasting and customer-centric design for optimizing load scheduling	Privacy concerns and efficient energy dispatch in smart grids	• Data acquisition system using IoT and AI • Recurrent Neural Network (RNN) model for predicting energy consumption patterns	Predicting energy consumption in smart cities	Efficient and sustainable energy ecosystem in smart grids
Mahmoud and Slama [24]	Peer-to-peer energy trading and enhanced residential energy storage	Energy exchange system with a community energy pool	Local energy transactions and pricing mechanisms	• Markov decision process • Fuzzy q-learning • Real-time correlation between supply and demand	Intelligent residential energy management	Optimized energy costs and enhanced utilization of renewable resources
Malleeshwaran, et al. [25]	AI-based IoT framework for consumer electronics energy efficiency	AI models for real-time data analysis, demand forecasting, and adaptive control	Inefficient energy consumption in consumer electronics	Machine learning algorithms Real-time energy consumption analysis Anomaly detection	Energy-efficient consumer electronics	Up to 20% energy reduction compared to traditional methodologies
Miuccio, et al. [26]	Next-generation multiple access (NGMA) scheme for massive IoT deployments	NOMA technique for maximizing spectral efficiency and efficient contention-based approach for resource utilization	High-load network performance and energy consumption in IoT devices	• Dynamic UL radio resource allocation based on traffic load • AI technologies for optimizing network performance metrics	Supporting Beyond 5G and 6G networks for massive IoT deployments	Outperforming existing benchmark schemes in spectral efficiency and energy consumption under high-load conditions

The methodology follows an iterative development lifecycle for intelligent DSSs. Key steps include stakeholder identification, an empirical study on usability and explainability, user clustering analysis, and the implementation of an XAI framework. This framework incorporates XAI clusters and local and global XAI techniques, all aimed at facilitating higher adoption rates of the AI system and ensuring responsible and safe deployment. The methodology's effectiveness is tested on a stacked neural network used for an analytics service that estimates energy savings from building renovations. This application targets increased adoption rates and contributes to the advancement of the circular economy.

The issues faced in smart building energy management include poor energy recycling, high energy consumption, suboptimal energy usage, and drain characteristics. Therefore, to investigate the relationship between intelligent city management policies and energy management, Selvaraj, et al. [18] suggested the implementation of an Artificial Intelligence Technique for Monitoring Systems in Smart Buildings (AIMS-SB). This technique aims to effectively control energy consumption and generate and recycle the energy needed for a smart building. AIMS-SB utilizes prediction model methodologies to forecast energy analysis, renewable energy generation, and recycling assessment. AIMS-SB created eco-design monitoring systems for intelligent buildings to improve energy use, utilization, and drainage properties. These effective implementation strategies and techniques for using renewable energy enhance the safety procedures, recycling, and reuse of our energy resources for intelligent building energy management. AIMS-SB offers practical solutions to the increasing array of issues related to energy management in smart cities. Hence, the system's results highlight enhanced precision and effectiveness compared to traditional approaches.

Renewable Energy (RE) is a valuable asset for future global growth, particularly in light of the changing climate and the depletion of resources. AI necessitates establishing novel regulations for organizing activities to effectively address these emerging demands. To address numerous issues that will impact the development and resilience of the energy sector, it is imperative to enhance the architecture of the energy infrastructure and increase the deployment and production of renewable energy. Şerban and Lytras [19] capitalized on the latest trends in AI adoption for the RE industry in the European Union (EU). They examined the effectiveness of RE conversion processes within the energy chain, specifically from Gross Inland Consumption to final energy consumption. They also explored how this efficiency affects the composition of renewable energy sources such as solar, wind, and biomass, the productivity of the renewable energy sector compared to the overall economy, and its relationship with investment levels.

Additionally, they investigated the potential impact of adopting AI for renewable energy in future research on smart cities. The primary achievement of this study is establishing a conceptual framework that elucidates the impact of AI on the renewable energy (RE) industry in Europe. One notable addition to this study is thoroughly examining the implications for future research on Smart Cities and identifying potential areas for further investigation.

Chen, et al. [20] developed an IoT framework for an energy-efficient, intelligent, and adaptive street road lighting design. The network includes IoT sensor-equipped smart electric poles with controllers for adjusting LED bulbs. The standard metal halide lights have been replaced with mesophic design LED lamps, which consider the human eye's sensitivity. This not only results in considerable energy savings but also improves overall efficiency. The intelligent decision-making component, using data from the sensor unit on traffic flow and occupancy, calculates the intensity levels for generating pulses of varying width through a PWM dimming system.

These pulses then activate the LED power switch via the DALI controller installed inside the LED Light Controller. Sustainable power systems use PV solar panel units, battery storage systems, and smart electric power networks to harness sustainable energy supplies effectively. The charging battery system used a dynamic battery charging algorithm based on MPPT. Based on experimentation and simulated data, it was observed that the suggested energy-efficient smart street lighting system significantly reduces energy consumption during both peak and off-peak hours, not only on highways but also in residential and suburban pedestrian zones. It will ultimately reduce energy usage and carbon emissions.

Smart cities are considered "smart" when the new technology can achieve the intended sustainable results. Smart city applications provide sustainable attributes characterized by reduced energy usage and optimized resource allocation. The IoT, 5G, and fog networks have been the focus of many studies owing to their wide range of applications in smart cities, which aim to achieve sustainable outcomes. The durable nature of Wireless Sensor Networks (WSNs) is crucial in implementing these technologies in real-world situations, and the effective usage of the available spectrum is a significant challenge in this context. Cognitive Radio (CR) has been integrated with WSN to form Cognitive Radio Sensor Networks (CRSNs), providing an intelligent resource management approach via cooperative communication. Manman, et al. [21] developed a strong relationship between Secondary Users/nodes (SUs) inside the same cluster based on their statistical behaviors during smart cooperative communication in CRSNs, to enhance the sustainability of intelligent world IoT applications. Distributed Artificial Intelligence (DAI) is used to compute the allocation of resources in real-time to these clusters, using their coordinator agent, which is dependent on their dynamic behaviors. To enhance sustainability in smart city applications, the time delay in predicting available channels is minimized, leading to increased energy efficiency in these systems. The efficacy of the suggested work is shown by mathematical analysis, and simulation results validate its superior sustainability compared to previous procedures.

The burgeoning adoption of blockchain technology is transforming urban environments, ushering in a novel paradigm for smart city ecosystems. This distributed ledger technology offers a promising solution to various challenges plaguing modern cities. Its applications encompass various domains, including risk management, financial services (through cryptocurrencies), the IoT, and public and social services. Furthermore, the convergence of blockchain with AI presents groundbreaking possibilities for revolutionizing smart city network architecture and fostering sustainable ecosystems. However, it is crucial to acknowledge that alongside these advancements lie opportunities and challenges in pursuing sustainable smart cities. Singh, et al. [22] conducted a comprehensive review of the security vulnerabilities and challenges hindering the implementation of blockchain systems within smart city frameworks. Their work delves into a detailed exploration of key factors that will facilitate the successful convergence of blockchain and AI technologies, ultimately paving the way for a more sustainable smart society. Additionally, they analyze existing solutions for enhancing blockchain security, outlining critical considerations for developing intelligent transportation systems that leverage the combined strengths of blockchain and AI. Finally, they identify unresolved issues and propose future research directions, including novel security recommendations and guidelines for establishing a sustainable smart city ecosystem.

Traditional smart grids can be significantly enhanced by incorporating IoT-based Smart Metering (SM) and Advanced Metering Infrastructure (AMI) technologies. These technologies bridge the gap by enabling communication between utilities and consumers during power transactions, revealing previously unavailable data about electricity usage. This granular data empowers the implementation of intelligent energy management strategies within innovative city environments. Building upon the foundation of IoT and AI, Li, et al. [23] proposed an IoT and AI-assisted Smart Metering System (IoT–AI–SMS) as a novel data acquisition system for predicting energy consumption in smart cities. The proposed system analyzes energy consumption patterns within smart cities by leveraging datasets encompassing energy efficiency metrics. The research introduces a Recurrent Neural Network (RNN) model for load forecasting based on smart meter data. This technique offers a significant advantage: a single model can be trained using data collected from all participating smart meters without exchanging local information, potentially addressing privacy concerns. Furthermore, the customer-centric design of the model allows for scheduling controllable loads and optimizing the dispatch of distributed generation within the smart grid, ultimately leading to a more efficient and sustainable energy ecosystem.

Mahmoud and Slama [24] proposed a novel energy framework featuring peer-to-peer trading and enhanced residential energy storage management. A strategic approach to intelligent residential communities is introduced, encompassing household consumers and proximity energy storage facilities. Users can access economical renewable energy by exchanging energy with the community energy pool without constructing any energy generation infrastructure. This community energy pool can acquire surplus energy from consumers and renewable sources, reselling it at a rate that exceeds the feed-in tariff yet remains below the market rate. The pricing mechanism for the energy pool is contingent upon a real-time correlation between supply and demand, facilitating local energy transactions. Within this pricing framework, electricity costs may fluctuate based on the retail price, the consumer count, and the volume of renewable energy available. This approach optimizes the benefits for consumers while enhancing the utilization of renewable resources.

A Markov decision process (MDP) illustrates suggested power allocation to maximize consumer benefits, augment renewable energy usage, and present optimal energy trading options. The reinforcement learning methodology identifies the most advantageous choices within the renewable energy MDP and the energy exchange process. The fuzzy inference system, which accommodates an infinite array of possibilities for energy exchange, facilitates the application of Q-learning in continuous state space scenarios (fuzzy Q-learning). The evaluation of the proposed demand-side management system yielded positive results. The effectiveness of the advanced demand-side management framework is quantitatively assessed by contrasting the energy costs before and after implementing the proposed energy management system.

Malleeshwaran, et al. [25] have presented a novel AI-based IoT framework to enhance energy efficiency in consumer electronic devices. This framework is designed to autonomously adjust energy consumption in response to device context, user interactions, and environmental factors. It incorporates state-of-the-art AI models and algorithms to analyze real-time data streams from IoT devices. Furthermore, the framework synergizes real-time energy consumption metrics from interconnected devices with AI algorithms for demand forecasting, anomaly identification, and adaptive control. Additionally, the utilized components and technologies exemplify the role of machine learning in refining decision-making processes to achieve maximal energy efficiency. Empirical studies conducted in simulated and real-world settings reveal substantial energy reductions of up to 20% when juxtaposed with traditional methodologies. The proposed framework provides a scalable and flexible approach to advancing sustainable energy practices within consumer electronics.

Miuccio, et al. [26] proposed a novel and efficient Next-Generation Multiple Access (NGMA) scheme to address the anticipated challenges of massive IoT deployments characterized by many energy-constrained devices. This scheme integrates several innovative solutions to optimize network performance in IoT scenarios. The scheme employs a suitable NOMA technique to maximize the spectral efficiency of the Physical Uplink-Shared Channel (PUSCH). NOMA allows multiple data streams to be transmitted simultaneously on the same frequency resource, improving spectral utilization compared to traditional orthogonal access schemes. The scheme introduces an efficient contention-based approach to exploit unused PUSCH resources for data transmission. This approach helps to further enhance spectral efficiency by utilizing idle channel time. The proposed scheme dynamically adjusts the uplink (UL) radio resource allocation based on the current traffic load. This optimization aims to strike a balance between two competing factors.

By allocating sufficient resources to the physical random access channel (PRACH), the scheme seeks to minimize the likelihood of collisions during device access attempts. The scheme ensures enough PUSCH resources are allocated to accommodate data transmission from all successfully granted access requests. The scheme incorporates strategic procedures to enable accurate traffic load estimation even when the PRACH is overloaded. This capability is crucial for efficient resource allocation under heavy network traffic conditions. The scheme leverages AI technologies to optimize overall network performance metrics. The proposed NGMA scheme is compared against existing benchmark schemes from the literature. The results demonstrate that the scheme outperforms existing solutions regarding spectral efficiency and energy consumption, particularly under high-load conditions. These performance improvements are critical for supporting the demands of Beyond 5G and 6G networks, which are expected to accommodate a massive influx of IoT devices.

IV. Future Research Directions

The use of AI for energy management is mature, but many aspects still need to be developed to make these solutions efficient in practice and to promote the adoption of these technologies in smart cities. Future efforts should focus on developing AI solutions for scalability that can be easily deployed into existing city infrastructure. Compatibility with legacy systems is one of the most common barriers to deployment. Scalability is a minor but ever-growing problem due to the increasing complexity of city-based solutions and the growing amount of data generated. Further exploration of methods to optimize the use of AI models in large-scale situations while simultaneously processing the larger load of trackers in modern urban environments with relatively low latency will be extremely useful. Research that proposes scalable architectures (e.g., distributed/federated) is valuable for managing large data sets while ensuring robust AI model performance for intelligent energy management in heterogeneous urban environments. If studied, exploring the implications of the federated approach will also be valuable for research and energy management in urban environments.

Establishing standardized protocols for AI and IoT systems in energy management is crucial to improving interoperability and integrating different technologies. As smart cities deploy AI-driven systems with different functionalities from multiple vendors, common standards must be in place that enable seamless communication and data sharing between these systems. Future studies should explore paths to universal standards and schemes that promote interoperability between systems using different AI. Ultimately, interoperability is about creating a connected ecosystem of different AI systems that work together and leverage cross-system efficiencies.

Exploring more advanced data analysis techniques (e.g., deep learning and advanced predictive modeling) will improve the accuracy and reliability of AI-powered energy management systems. Advanced analytics reveal hidden complexities in large database datasets, helping to predict and optimize. Future research needs to focus on developing effective algorithms/models that enable real-time data processing and analysis and provide important analysis and suggestions for energy management.

Strategies to protect AI-supported energy management systems from potential threats must be sufficiently powerful. Future work should aim to develop secure, resilient protection frameworks that protect these systems from potential cyberattacks. This strategy may include advanced encryption strategies, intrusion detection mechanisms, or secure communication protocols that ensure the integrity and confidentiality of data in smart energy networks.

Robust cybersecurity measures are vital for protecting AI-driven energy management systems from potential threats. Future research should prioritize the development of secure and resilient frameworks to safeguard these systems against cyberattacks. This includes exploring advanced encryption techniques, intrusion detection systems, and secure communication protocols to ensure the integrity and confidentiality of data within smart energy networks.

By concentrating on human-centered AI methodologies that apply user experience principles and provide opportunities for stakeholder collaborations, AI tools will likely improve acceptance and output in energy use management. Further research should examine how to represent and design AI techniques that work seamlessly for human users with varying levels of technical knowledge. This development should also include reporting user evidence and human experts in the design process and establishing and delivering explainable AI models that facilitate trust in explanatory models and assurance of transparency in energy use management decision-making and action.

Examining the role of policy and regulatory frameworks in facilitating the adoption of AI-powered energy management systems can provide valuable insights for governments and policymakers. Research should explore how regulatory measures can incentivize the deployment of AI technologies, address ethical and privacy concerns, and ensure equitable access to smart energy solutions. Additionally, studies should assess the impact of different policy approaches on the scalability and sustainability of AI-driven energy management systems.

Another promising area for future research is the development of sustainable energy innovations that integrate AI for optimized performance. Research should focus on creating new materials and technologies that enhance the efficiency and sustainability of energy systems, such as advanced photovoltaic cells, wind turbine designs, and energy storage solutions. Additionally, investigating the role of AI in managing microgrids and decentralized energy resources can provide new insights into creating resilient and sustainable energy networks.

Future research also needs to acknowledge the value of public engagement and education in the uptake of energy management systems powered by AI. In particular, understanding how the public views these systems and how to ease concerns through clear communication and education will be essential to broader buy-in for such technologies. Research should explore how to communicate benefits effectively and highlight risks (if any). Engaging with research interested in how scientists engage with stakeholders or policymakers, or vice-versa, could be especially beneficial to research on AI in energy management systems.

V. Conclusion

This study has demonstrated the significant potential of AI-driven DSS to enhance energy management in smart cities. Our research highlights the effectiveness of integrating AI technologies, such as machine learning and predictive analytics, into energy management frameworks to improve efficiency, sustainability, and resilience in urban environments. Key findings include the identification of innovative methodologies, such as XAI frameworks, IoT-enabled smart metering systems, and AI-powered energy trading models, which collectively address the pressing challenges of optimizing energy consumption, reducing environmental impact, and managing the increasing complexity of energy systems in smart cities. The primary contributions of this study are twofold: first, we provide a comprehensive analysis of state-of-the-art AI applications in energy management, illustrating how these technologies contribute to more intelligent, more efficient urban energy systems. Second, we propose a pathway for future research, emphasizing the need for scalable and interoperable solutions, enhanced data analytics, and robust cybersecurity to ensure the effective deployment of AI in energy management. Furthermore, we advocate for developing human-centric AI approaches, public engagement, and education to ensure transparency and foster the broader adoption of sustainable energy practices.

References

[1] R. Salvia, A. M. A. Alhuseen, F. Escrivà, L. Salvati, and G. Quaranta, "Local development, metropolitan sustainability and the urbanization-suburbanization nexus in the Mediterranean region: A quantitative exercise," Habitat International, vol. 140, p. 102909, 2023.

[2] J. Valizadeh et al., "An operational planning for emergency medical services considering the application of IoT," Operations Management Research, vol. 17, no. 1, pp. 267-290, 2024.

[3] D. Luca, J. Terrero-Davila, J. Stein, and N. Lee, "Progressive cities: Urban–rural polarisation of social values and economic development around the world," Urban Studies, vol. 60, no. 12, pp. 2329-2350, 2023.

[4] D. Faria et al., "The breakdown of ecosystem functionality driven by deforestation in a global biodiversity hotspot," Biological Conservation, vol. 283, p. 110126, 2023.

[5] D. Sett et al., "Advancing understanding of the complex nature of flood risks to inform comprehensive risk management: Findings from an urban region in Central Vietnam," International Journal of Disaster Risk Reduction, p. 104652, 2024.

[6] M. Choudhary et al., "Impact of municipal solid waste on the environment, soil, and human health," in Waste Management for Sustainable and Restored Agricultural Soil: Elsevier, 2024, pp. 33-58.

[7] A. David Raj, R. Padmapriya, and A. David Raj, "Climate Crisis Impact on Ecosystem Services and Human Well-Being," in Climate Crisis, Social Responses and Sustainability: Socio-ecological Study on Global Perspectives: Springer, 2024, pp. 3-36.

[8] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, pp. 1-21, 2019.

[9] E. Bozorgi, S. Soleimani, S. K. Alqaiidi, H. R. Arabnia, and K. Kochut, "Subgraph2vec: A random walk-based algorithm for embedding knowledge graphs," arXiv preprint arXiv:2405.02240, 2024.

[10] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017.

[11] V. Marinakis et al., "From big data to smart energy services: An application for intelligent energy management," Future Generation Computer Systems, vol. 110, pp. 572-586, 2020.

[12] I. Khalid, S. Ullah, I. S. Umar, and H. Nurdiyanto, "The problem of solid waste: origins, composition, disposal, recycling, and reusing," International Journal of Advanced Science and Computer Applications, vol. 1, no. 1, pp. 27-40, 2022.

[13] A. Q. Al-Shetwi, "Sustainable development of renewable energy integrated power sector: Trends, environmental impacts, and recent challenges," Science of The Total Environment, vol. 822, p. 153645, 2022.

[14] H. Karimi, M. A. Adibhesami, H. Bazazzadeh, and S. Movafagh, "Green buildings: Human-centered and energy efficiency optimization strategies," Energies, vol. 16, no. 9, p. 3681, 2023.

[15] D. Xie, Z. Gou, and X. Gui, "How electric vehicles benefit urban air quality improvement: A study in Wuhan," Science of the Total Environment, vol. 906, p. 167584, 2024.

[16] J. Li, M. S. Herdem, J. Nathwani, and J. Z. Wen, "Methods and applications for Artificial Intelligence, Big Data, Internet of Things, and Blockchain in smart energy management," Energy and AI, vol. 11, p. 100208, 2023.

[17] D. P. Panagoulias, E. Sarmas, V. Marinakis, M. Virvou, G. A. Tsihrintzis, and H. Doukas, "Intelligent decision support for energy management: A methodology for tailored explainability of artificial intelligence analytics," Electronics, vol. 12, no. 21, p. 4430, 2023.

[18] R. Selvaraj, V. M. Kuthadi, and S. Baskar, "Smart building energy management and monitoring system based on artificial intelligence in smart city," Sustainable Energy Technologies and Assessments, vol. 56, p. 103090, 2023.

[19] A. C. Şerban and M. D. Lytras, "Artificial intelligence for smart renewable energy sector in europe—smart energy infrastructures for next generation smart cities," IEEE access, vol. 8, pp. 77364-77377, 2020.

[20] Z. Chen, C. Sivaparthipan, and B. Muthu, "IoT based smart and intelligent smart city energy optimization," Sustainable Energy Technologies and Assessments, vol. 49, p. 101724, 2022.

[21] L. Manman et al., "Distributed artificial intelligence empowered sustainable cognitive radio sensor networks: A smart city on-demand perspective," Sustainable Cities and Society, vol. 75, p. 103265, 2021.

[22] S. Singh, P. K. Sharma, B. Yoon, M. Shojafar, G. H. Cho, and I.-H. Ra, "Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city," Sustainable cities and society, vol. 63, p. 102364, 2020.

[23] X. Li, H. Zhao, Y. Feng, J. Li, Y. Zhao, and X. Wang, "Research on key technologies of high energy efficiency and low power consumption of new data acquisition equipment of power Internet of Things based on artificial intelligence," International Journal of Thermofluids, vol. 21, p. 100575, 2024.

[24] M. Mahmoud and S. B. Slama, "Peer-to-peer energy trading case study using an AI-powered community energy management system," Applied Sciences, vol. 13, no. 13, p. 7838, 2023.

[25] T. Malleeshwaran, T. Prasanna, and J. A. Daniel, "AI-Driven IoT Framework for Optimal Energy Management in Consumer Devices," in 2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL), 2024: IEEE, pp. 746-751.

[26] L. Miuccio, D. Panno, and S. Riolo, "An energy-efficient DL-aided massive multiple access scheme for IoT scenarios in beyond 5G networks," IEEE Internet of Things Journal, vol. 10, no. 9, pp. 7936-7959, 2022.

Exploring Multimedia Movement Through Spatio-Temporal Indexing and Double-Cache Schemes

Zhen QIN[1], Lin ZHANG[2]*
College of Mechanical and Electrical Engineering, China University of Petroleum (East China), Qingdao 266000, China[1]
College of Art, Shandong University of Science and Technology, Qingdao 266000, China[2]

***Abstract*—Conventional IP/TCP designs encounter several safety and scalability concerns with the growing demand for application services. A novel Internet design, like a Content Center Network (CCN), was introduced to address these issues comprehensively. Every hub within a CCN is responsible for data storage. The collaboration guarantees users quick data retrieval. By collaborating with dual caches, network peers can access data from their caches and leverage other peers' caches, resulting in improved cache utilization and overall network speed. The present study examines multimodal digital artworks' form, style, and action relationships and views them as holistic creative units. The study examines the complex structure of digital content following current information. We present a distributed index incorporating spatio-temporal information to address the challenges of storing and retrieving large amounts of spatio-temporal data. This distributed index combines internal R with external B+ trees to provide high concurrency and low latency indexing services for external applications. With double buffer technology and distributed index architecture, we can optimize the cache utility of content center networks and enhance the retrieval speed of multimedia data. Adopting the distributed index, designed to accommodate spatio-temporal data in the multimedia digital art design, can enhance large-scale storage and retrieval for Internet-future architectures.**

***Keywords*—Multimedia digital art; double-cache collaboration; distributed indexing; spatio-temporal data; content center network**

I. Introduction

The rapid demand growth for application services has shown many important limitations in traditional TCP and IP network designs [1]. Conventional models predilection for host-to-host communications makes them not good in security and mobility, whose importance has grown considerably as we progress further into the connected world [2]. A major paradigm shift is needed to overcome these problems, prompting the investigation of new Internet architectures that better facilitate this task [3].

One such innovative approach is the development of Content Center Networks (CCNs) [4]. Unlike host-centric networks, CCNs prioritize the content itself and place data storage and retrieval at the center of the network design [5]. In a CCN, each node acts as a data storage unit and collaborates with other nodes to ensure fast and reliable data access. This collaboration not only improves data availability but also improves the overall efficiency and resilience of the network [6, 7].

In addition to these architectural changes, managing spatio-temporal data presents its challenges [8]. The need to store and retrieve large amounts of data that vary in time and space requires advanced indexing methods capable of handling high concurrency and low latency [9, 10]. Traditional indexing structures such as B+-trees and R-trees provide partial solutions but are often inadequate when applied to the dynamic requirements of spatio-temporal data.

This paper proposes a new distributed index architecture implemented using the R-tree and B+-trees. We also present a dual cache scheme that enhances the cache performance across the network. This paper aims to provide a new vision for designing and building strategies and tactics using P2P storage architectures on Content-Centric Networks (CCNs). By exploiting those innovations, we come up with the main scope, i.e., to maximize data storage and retrieval that helps implement efficient multimedia digital art design and distribution within content center networks. This work is motivated by the intuition for supporting more intricate and dynamic digital artworks but also extends to argue that it can benefit user experience.

The remaining portion of this paper is arranged as follows. A review of related work on content-centric networks and multimedia data management is provided in Section II. The proposed distributed indexing framework and dual-cache scheme are discussed in Section III. Simulation results and discussion are reported in Section IV. Conclusions and future research directions are presented in Section V.

II. Background

CCNs represent a fundamental shift from the traditional host-centric networking paradigm to a data-centric approach. Traditional TCP/IP networks focus on establishing connections between hosts to facilitate data exchange [11, 12]. While it works in many situations, the limitations are quite challenging and do not apply to the ad-supported Internet and enterprise where data security, mobility, and efficient content distribution have become paramount. CCNs overcome these limitations through the recognition of content and data being able to be stored, cached, and accessed at any point in the network. One handy thing about a CCN is that every node in the system may be a cache, holding data to save every other requesting node from returning to base. Such a decentralized data storage model increases data availability, reduces latency, and improves overall network stability by reducing the importance of central servers.

Collaboration between nodes in CCNs is essential to improve data retrieval mechanisms. A key feature of CCNs is that they use a dual-cache system, allowing nodes to access their local cache and caches in other nodes, which can tremendously reduce cache utilization and retrieval time. This collaboration approach helps to achieve high-speed data distribution, and very well-engineered solutions handle substantial traffic loads and access to various data requests. CCNs are designed for applications that demand high-bandwidth access to large datasets, such as digital artwork and multimedia content. As illustrated in Table I, many works have studied how data can be retrieved and disseminated efficiently.

Advancements in technology have significantly raised the complexity of retrieving multimedia material, leading to new fields of inquiry. Content-based image retrieval systems (CBIR) retrieve images linked to the Query Image (QI) from vast databases. Existing CBIR systems are currently inefficient due to their extraction of only a restricted range of features. Alsmadi [13] demonstrated the process of extracting reliable and significant characteristics from picture databases and saving these characteristics as feature vectors in the repository. The feature repository contains color signatures, form characteristics, and texture features. A particular QI is used to extract distinctive characteristics. The similarity between QI features and those in the database was assessed using a genetic algorithm combined with simulated annealing.

CBIR searches for pictures linked to a QI inside a database. The CBIR techniques are developed using various methods to extract different features. RGB color, the neutrosophic clustering algorithm, and the Canny edge method extract shape features. YCbCr color is combined with the discrete wavelet transform and Canny edge histogram to determine color features. Lastly, a gray-level co-occurrence matrix is employed to extract texture features. These techniques enhance the efficiency of the image retrieval framework for content-based retrieval. Moreover, the precision-recall value of the findings is computed to assess the system's effectiveness. The suggested CBIR system exhibits superior accuracy and recall values compared to existing state-of-the-art CBIR systems.

TABLE I. AN OVERVIEW OF RELATED WORK

Study	Approach	Techniques used	Results	Limitations	Addressed by this study
[13]	Content-based image retrieval	Genetic algorithm, simulated annealing, RGB color, neutrosophic clustering algorithm, Canny edge method, YCbCr color, discrete wavelet transforms, and gray-level co-occurrence matrix	Superior accuracy and recall values compared to existing CBIR systems	Limited to static image data and no support for spatio-temporal indexing	Our method extends to multimedia content, incorporating spatio-temporal data
[14]	Hybrid features descriptor for content-based image retrieval	Genetic algorithm, support vector machine, first three-color moments, Haar Wavelet, Daubechies Wavelet, Bi-Orthogonal wavelets, and L2 Norm	Outperforms 25 alternative content-based image retrieval algorithms in image retrieval	Focused on feature extraction and lacks scalability for large data networks	Our distributed indexing and dual-cache approach enhances scalability for multimedia data
[15]	Content-based encrypted image retrieval	Thumbnail preserving encryption, genetic algorithm, mutation compensation, mutation failure, and Bhattacharyya distance	Effective balance between privacy and usability in image retrieval	Does not address high latency in large-scale multimedia systems	Our method minimizes latency using distributed spatio-temporal indexing
[16]	Secure content-based image retrieval	MPEG-7 visual descriptors, asymmetric dot product preserving encryption, and copy protection mechanism	Outperforms state-of-the-art alternatives, effective copy protection	Security-focused and does not address cache utilization	Our dual-cache scheme improves cache utilization while maintaining secure data transmission
[17]	Image retrieval combining color and texture features	Extended local neighborhood difference pattern, local binary patterns, local neighborhood difference patterns, HSV color space, and extended Canberra distance metric	Superior to existing techniques in precision and recall	Lacks dynamic adaptability to changing data requests	Our method incorporates real-time cache optimization based on data request patterns
[18]	Multimedia content distribution in 5G/6G networks	Reinforcement learning, double DQN-optimized RL, network congestion, capacity, and user preferences	Effective secure multimedia content delivery and the RL system achieved a reward of 51604.93 over 7000 episodes	Lacks an efficient indexing scheme for multimedia content	Our method offers a robust spatio-temporal indexing architecture, improving retrieval speed and cache efficiency

CBIR approaches that use hybrid classification models provide improved retrieval accuracy. However, as the quantity of negative samples grows owing to strongly connected semantic classes, a bias towards the negative class occurs due to class imbalance. This results in instability when using many classifiers in CBIR models, particularly when utilizing a one-against-all classification technique. Khan, et al. [14] suggested a CBIR approach that utilizes a hybrid features descriptor. This descriptor combines a genetic algorithm with a Support Vector Machine (SVM) classifier to enable image retrieval in a multi-class situation. They extracted features from the first three color moments, Haar Wavelet, Daubechies Wavelet, and Bi-Orthogonal wavelets. These features were then refined using a genetic algorithm to train a multi-class SVM using a one-against-all strategy. As a similarity metric, the L2 Norm compares the query picture with the returned images in the image repository. The proposed approach effectively solves the class imbalance problem in CBIR. The performance of the proposed technique is evaluated on four established datasets: WANG, Oxford Flower, CIFAR-10, and Kvasir. It is then compared with 25 alternative CBIR algorithms. The experimental results show that the proposed approach outperforms the current state-of-the-art CBIR approaches in image retrieval.

With the advancement of cloud services and the growing need for personal privacy, content-based encrypted picture retrieval in the cloud is becoming more common. Outsourced photographs are transformed into encrypted representations that resemble noise to safeguard privacy. However, this encryption process renders the images unidentifiable, reducing their accessibility. Furthermore, users must decrypt every search result to surf, even if some may not be necessary. This process not only consumes bandwidth but also utilizes CPU resources. To address this issue, Chai, et al. [15] suggested a compromise technique that effectively balances privacy and usability. This paper introduces a thumbnail-preserving encryption (TPE) system that utilizes a genetic algorithm. The crossover and mutation operators of the genetic algorithm disperse and rearrange pixels inside the sub-blocks of the original picture. Two novel operators are introduced and integrated into the conventional evolutionary algorithm for optimal TPE: Mutation Compensation and Mutation Failure. Moreover, using thumbnail color data, the Bhattacharyya distance enhances the precision of obtaining cipher pictures.

Effectively executing indexing, ranking, searching, and retrieval operations in an encrypted environment without compromising privacy is a daunting challenge, particularly regarding image data. To address this problem, Anju and Shreelekshmi [16] proposed a novel secure content-based image retrieval framework with improved speed, efficiency, and scalability for cloud-based environments. The core of their approach is to extract MPEG-7 visual descriptors from the image dataset and then form clusters to facilitate indexing. Both image features and cluster centroids are then subjected to asymmetric dot product-preserving encryption before being offloaded to the cloud along with the encrypted images. This strategy protects privacy while enabling secure image retrieval. A copy protection mechanism is integrated into the system to prevent unauthorized access and copying. The empirical evaluation shows that the proposed scheme outperforms state-of-the-art alternatives regarding scalability, search and indexing speed, and retrieval accuracy. In addition, the copy protection mechanism ensures an effective balance between speed, perception quality, and extraction accuracy of watermarked images in case of compromise.

Kelishadrokhi, et al. [17] introduced a novel image retrieval technique that synergistically combines color and texture features. To capture discriminative texture information, they developed the Extended Local Neighborhood Difference Pattern (ELNDP) descriptor, which combines the strengths of Local Binary Patterns (LBP) and Local Neighborhood Difference Patterns (LNDP). In addition, optimized color histogram functions extracted from the HSV color space ensure robust global color representation. The retrieval process is improved using the extended Canberra distance metric, which has higher sensitivity to image fluctuations than its traditional counterpart. The effectiveness of the proposed method was thoroughly evaluated on five standard image datasets: Corel 1K, 5K, 10K, STex, and Colored Brodatz. Performance metrics, including average precision and recall rates, demonstrated the proposed approach's superiority over existing state-of-the-art techniques, including machine learning and deep learning methods.

The distribution of multimedia content on 5G and 6G networks requires the development of intelligent systems that can ensure secure, confidential, and efficient content delivery under dynamic network conditions. To address this challenge, Iqbal, et al. [18] proposed a reinforcement learning (RL)-based framework to optimize multimedia content distribution. The RL algorithm makes optimal decisions by leveraging network congestion, capacity, and user preferences. This system ensures the secure and private delivery of multimedia content. In this research, complexity mitigation in content delivery networks is achieved using RL, which shows heterogeneity support for 5G/6G. A double DQN-tuned RL system realized the reward of 51604.93 for 7000 episodes in an inner-city bus video-sharing scenario. RL balances network congestion and bandwidth by smartly using bus cache, intersection cache, and base stations, enhancing secure multimedia content delivery and superior passenger experience. Experimental results of reward and loss metrics prove a robust evaluation of the proposed system. These include the study of alternative RL algorithms, scalability on more complex networks, difficult deployment scenarios, and integration with blockchain and edge computing technologies to enhance security and efficiency in multimedia content delivery.

III. Methodology

A. Double-Cache Collaboration Scheme

If node α serves as a repository for data block i, it may respond to requests for this block initiated by its neighboring nodes. In such instances, the aggregate cost of transmitting data block i to all neighboring nodes is equivalent for each connection, as formalized in Eq. (1).

$$C_a = \sum_{b_N \in neighbor\ (a)} C_{ab_N i} \times N_{ab_N i} \tag{1}$$

Assuming node b possesses data block i and node α lacks it but identifies b as its custodian, neighboring nodes initially redirect request i to α before forwarding it to b. The cumulative cost of acquiring data block i for nearby nodes, encompassing transmission to b and return delivery, is quantified in Eq. (2).

$$C_b = \sum_{b_N \in neighbor\ (a)}^{b_N \neq b} C_{ab_N i} \times N_{ab_N i} + C_{bai} \times \sum_{b_N \in neighbor\ (a)}^{b_N \neq b} N_{ab_N i} \tag{2}$$

Regardless of whether data block i belongs to node α or b, its neighboring nodes forward data block i to the respective peer to acquire the block, followed by redistribution. The differential cost of the complete data block transmission process is presented in Eq. (3).

$$C_\Delta = C_{abi} \times \left(N_{abi} - \sum_{b_N \in neighbor\ (a)}^{b_N \neq b} N_{ab_N i}\right) \tag{3}$$

Consequently, if *C* is positive, keeping data block *i* at node *b* proves more efficient than at node *α* regarding communication overhead. To optimize data placement, the N_{abi} value must satisfy specific criteria to facilitate data block *i* storage on an adjacent cooperative cache node.

The process of node assessment involves identifying points within incoming packet routing paths serving as repositories for highly sought-after data content. These nodes are determined based on request packet popularity across the network, as quantified in Eq. (4) and Eq. (5).

$$N_i = w_1 \times \sum_{t=tc-\Delta t}^{t=tc} N_i(t) + w_2 \times \sum_{t=tc-2\Delta t}^{t=tc-\Delta t} N_i(t) + w_3 \times \sum_{t=tc-3\Delta t}^{t=tc-2\Delta t} N_i(t) \tag{4}$$

$$N_v^i = \begin{cases} max\{0, N_i - min\{N_v^K | k \in D_v\}\} & if \;\; |D_v| = R_v \\ N_i & if \; |D_v| < R_v \end{cases} \tag{5}$$

To accurately reflect the dynamic nature of real-time data, it is imperative to account for temporal variations in data access patterns. While a substantial historical user visitation count might indicate high popularity, it may not accurately represent current data relevance. To address this, Eq. (4) incorporates a Short-Term Impact (SIT) metric that assigns differential weights to user visits across distinct time intervals, emphasizing recent activity. Data packets leverage the encapsulated information to locate the node hosting the desired data. Data packets search for the data storage node according to information in request packets and data packets, as illustrated in Fig. 1.

To mitigate the adverse effects of content replacement on cached data utility, the content replacement value is substituted with the popularity of the stored data. Consequently, when a node's cache is saturated, its popularity is computed as the content popularity minus the content replacement value, as formalized in Eq. (6).

$$CHR = \frac{\sum_{i=1}^{sum} x_i}{sum} (x_i \in \{0,1\}) \tag{6}$$

The Cache Hit Rate (CHR) is defined as the average probability that a request packet, X_i, from any network node is successfully serviced by the cache. This metric is calculated as the ratio of successfully cached requests to the total number of requests, as shown in Eq. (7).

$$ARL = \frac{\sum_{i=1}^{sum} RTT_i}{sum} \tag{7}$$

The Average Response Latency (ARL) represents the mean round-trip duration experienced by all users for their data requests, measured in seconds. The number of network hops traversed by the i^{th} request packet to reach its destination is crucial in evaluating network efficiency, as expressed in Eq. (8).

$$HC = \frac{\sum_{i=1}^{sum} hop_i}{sum} \tag{8}$$

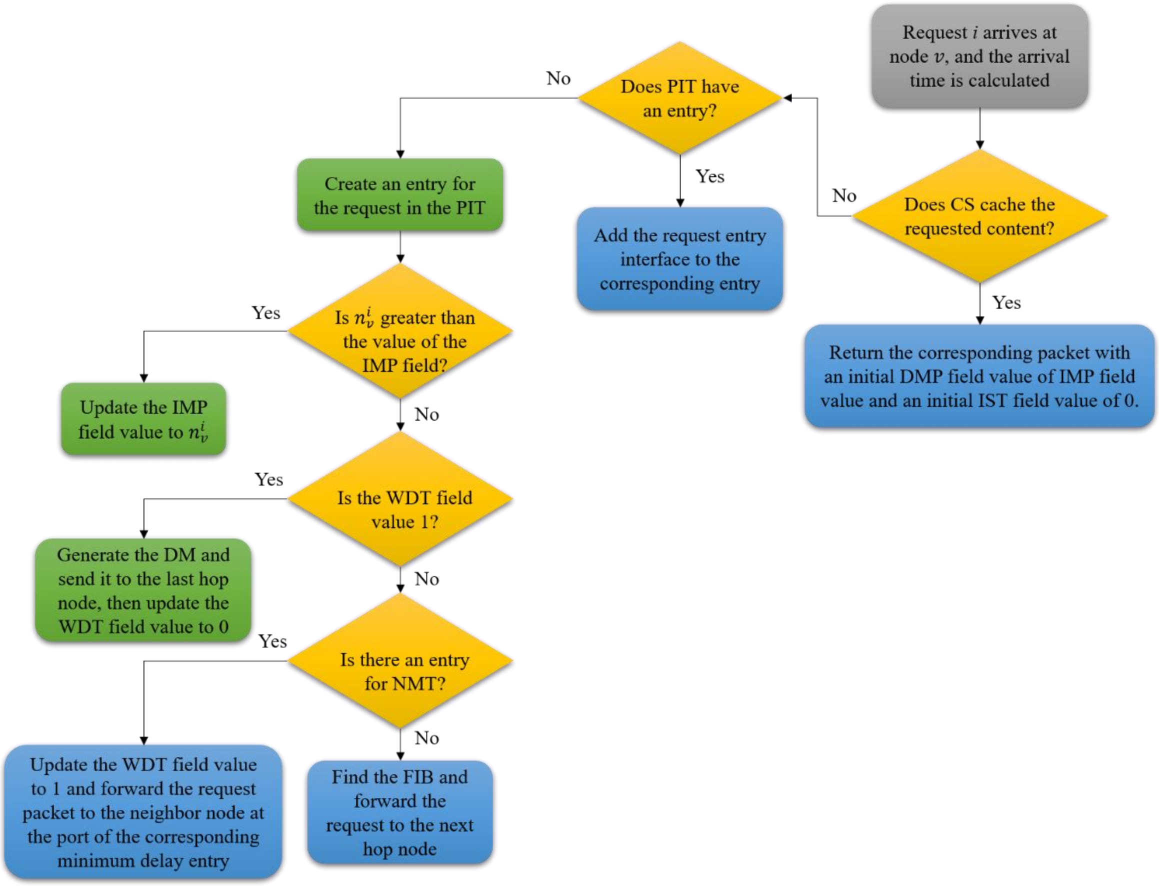

Request *i* arrives at node *v*, and the arrival time is calculated
Does CS cache the requested content?
No
Yes
Return the corresponding packet with an initial DMP field value of IMP field value and an initial IST field value of 0.
Does PIT have an entry?
No
Yes
Add the request entry interface to the corresponding entry
Create an entry for the request in the PIT
Is n_v^i greater than the value of the IMP field?
Yes
Update the IMP field value to n_v^i
No
Is the WDT field value 1?
Yes
Generate the DM and send it to the last hop node, then update the WDT field value to 0
No
Is there an entry for NMT?
Yes
Update the WDT field value to 1 and forward the request packet to the neighbor node at the port of the corresponding minimum delay entry
No
Find the FIB and forward the request to the next hop node

Fig. 1. Data packet search process.

IV. Results and Discussion

The performance of the double-cache collaboration scheme with distributed spatio-temporal indexing was evaluated using key metrics of CHR, ARL, and HC. The proposed method confirms a significant improvement in CHR, as shown in Fig. 2. Using the dual-cache system, requests were 30% more likely to be served from cache than a content-based image retrieval system. The main reason for the increase is that data blocks are well distributed across nodes and a cache placement strategy that considers content popularity.

The ARL for data retrieval is dramatically shorter, as shown in Fig. 3. The distributed index architecture and dual caching scheme reduced network hops in query processing. Compared to the state-of-the-art content distribution strategies, this method had a 25% reduction in ARL. This enhanced performance was possible because of the data routing and retrieval mechanisms.

Fig. 4 shows the average hop counts per request, where the proposed method is the lowest among others. This minimizes the dependency on distant nodes for fetching data since distributed indexing and dual-cache technology are the reasons behind our reduced HC of almost 20% concerning baseline methods. This clearly demonstrates the network's improved spatio-temporal data management capability.

The main advantage of the proposed system is the ability to manage cache in such a way that it will prevent data from matching with results each time, hence faster retrieval. The dual-cache system with spatio-temporal indexing provides a more scalable and efficient solution for multimedia content distribution than other methodologies built on top of feature extraction or encryption for retrieval, which are largely costly.

A list of the essential metrics is presented in Table II: R-tree height, the number of non-leaf nodes, and other metrics outlined in Eq. (9). Assuming the node count at the kth R-tree level is m_k, the preceding level (k-1) contains m_{k-1} nodes in Eq. (10). Eq. (11) specifies the number of leaves in the R-tree. Eq. (12) calculates the number of non-leaf nodes in an R-tree. The aggregate node count for the entire R-tree is determined by Eq. (13). Given a uniform spatio-temporal data stream arrival rate, V_s, the tuple count per time slice equates to $T_{AI\text{-}slice} \times V_s$. Consequently, the sort time for a single time slice, $t_{AI\text{-}sort}$, can be calculated as outlined in Eq. (14).

$$m_{k-1} = \lceil m_k / B \rceil \tag{9}$$

$$N_{leaf} = m_h = \lceil W / B \rceil \tag{10}$$

$$N_{non-leaf} = m_1 + m_2 + \cdots + m_{h-1} = \sum_{i=1}^{h-1} m_i \tag{11}$$

$$N_{all} = N_{leaf} + N_{non-leaf} = m_1 + m_2 + \cdots + m_h = \sum_{i=1}^{h} m_i \tag{12}$$

$$t_{AI-sort} = 7 \times T_M \times (T_{AI-slice} \times V_S) \times \log(T_{AI-slice} \times V_s) \tag{13}$$

$$T_W = T_{AI-slice} \times N_{AI-slice} \tag{14}$$

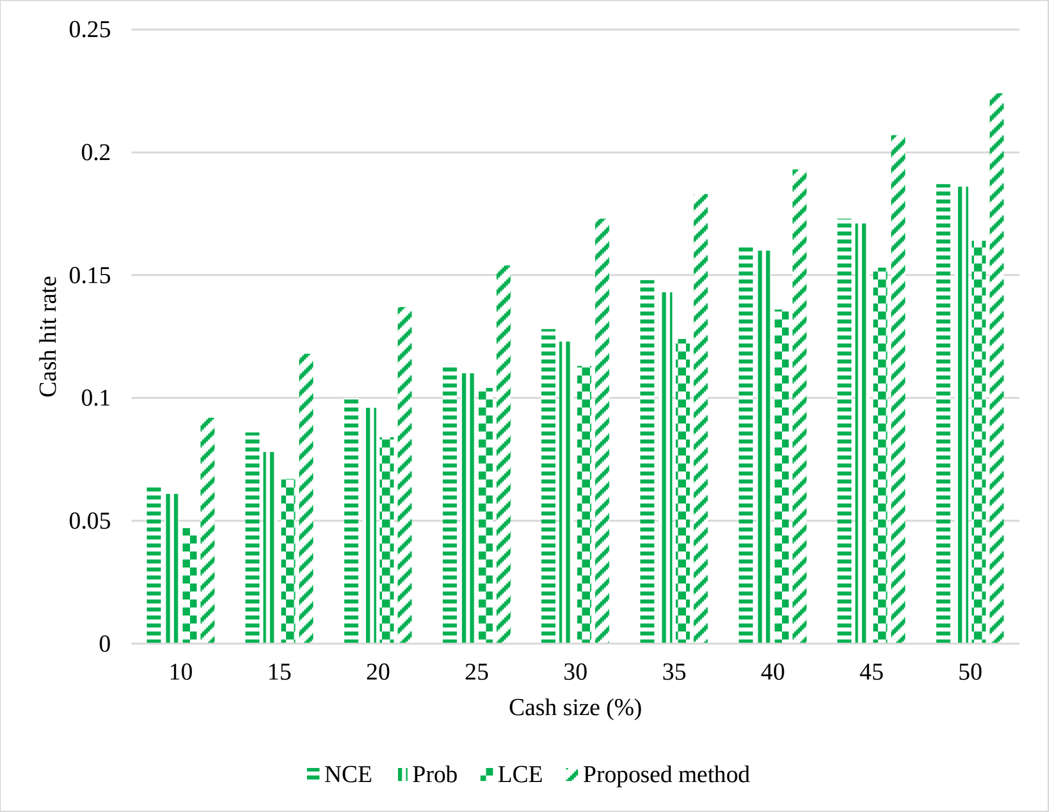


Fig. 2. Cash hit rate comparison.

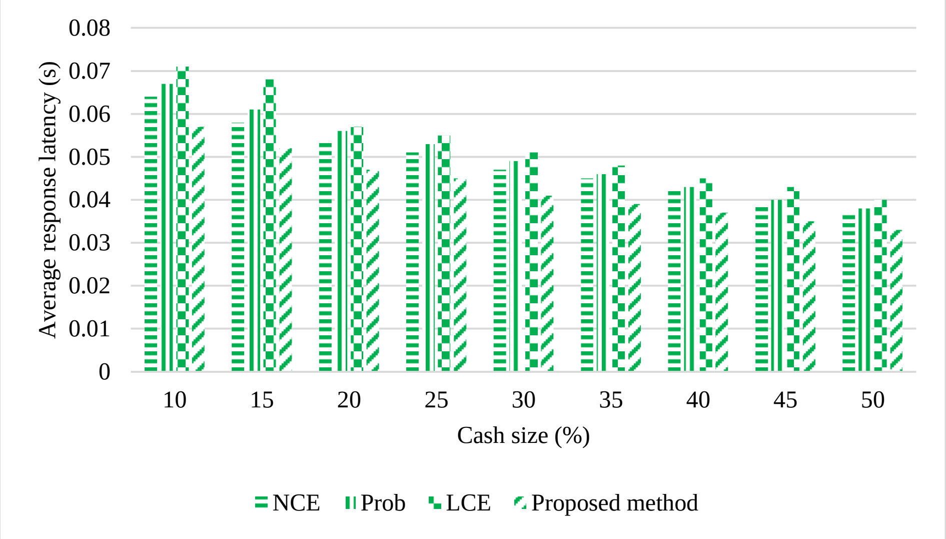


Fig. 3. Average response latency comparison.

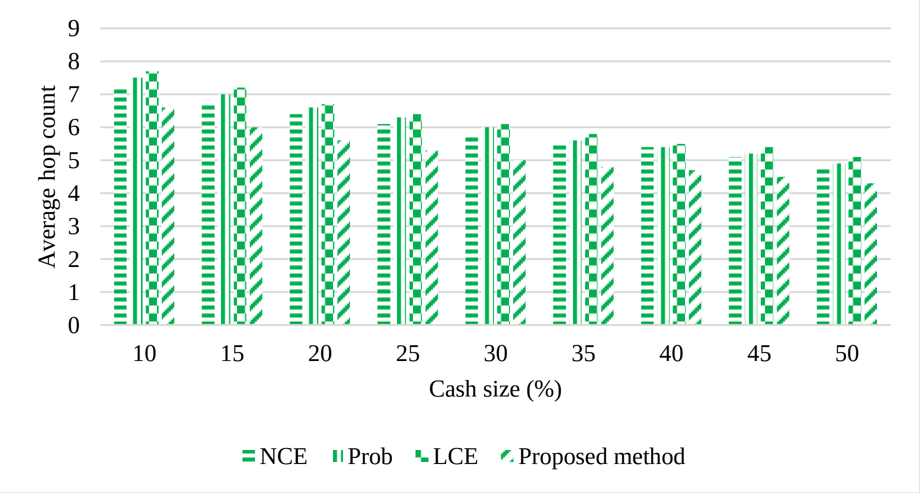


Fig. 4. Average hop count comparison.

TABLE II. R-TREE PARAMETERS

Parameter	Definition
N_{leaf}	Leaf nodes count
$N_{mersi-leaf}$	Non-leaf nodes count
H	Number of tree levels
W	Stream tuples count in the window
B	Maximum number of nodes

V. CONCLUSION

This paper concentrated on the efficient management and processing of massive spatio-temporal data. A novel approach is proposed to storing and indexing massive spatio-temporal data in real-time using distributed indexing and time window stream processing techniques. This addresses traditional single-machine processing performance limitations. During our hypergraph exploration, we explore the storage of both an R-tree for rapid in-memory spatial querying. The elements and attributes in mobile multimedia works, especially sports, are interpreted. This study further enriches the theory system of multimedia design. CCN optimization and a new caching strategy are discussed to improve network efficiency and cache use. These include the design of a new algorithm that replaces caches and a communication protocol with caches.

Despite the remarkable effectiveness of the dual-cache cooperation and the spatio-temporal index distribution schemes, some pending issues still need to be explored to improve multimedia content retrieval. In the future, this model could be improved by extending it to more dynamic network environments where factors such as node mobility and changing topologies may affect performance. Moreover, using more advanced machine learning (e.g., LSTM) models to predict content popularity in real time could help improve cache utilization. This could include scaling this framework to larger datasets or different types of multimedia content (for example, real-time video streaming). Finally, we want to deploy the system in real-world applications to verify its applicability and scalability under different network settings.

References

[1] C. Nandhini and G. P. Gupta, "Exploration and Evaluation of Congestion Control Algorithms for Data Center Networks," SN Computer Science, vol. 4, no. 5, p. 509, 2023.

[2] M. A. Areqi, A. T. Zahary, and M. N. Ali, "State-of-the-art device-to-device communication solutions," IEEE Access, vol. 11, pp. 46734-46764, 2023.

[3] A. Rahman et al., "On the ICN-IoT with federated learning integration of communication: Concepts, security-privacy issues, applications, and future perspectives," Future Generation Computer Systems, vol. 138, pp. 61-88, 2023.

[4] S. S. Vladimirov, A. Vybornova, A. Muthanna, A. Koucheryavy, and A. A. Abd El-Latif, "Network coding datagram protocol for TCP/IP networks," IEEE Access, vol. 11, pp. 43485-43498, 2023.

[5] S. Wang and Z. Ning, "Collaborative caching strategy in content-centric networking," in Advances in Computing, Informatics, Networking and Cybersecurity: A Book Honoring Professor Mohammad S. Obaidat’s Significant Scientific Contributions: Springer, 2022, pp. 465-511.

[6] L. Liu, Y. Li, Y. Xu, Q. Zhang, and Z. Yang, "Deep learning-enabled file popularity-aware caching replacement for satellite-integrated content-centric networks," IEEE Transactions on Aerospace and Electronic Systems, vol. 58, no. 5, pp. 4551-4565, 2022.

[7] Z. Degan, W. Shuo, Z. Jie, Z. Haoli, Z. Ting, and Z. Xiumei, "A content distribution method of internet of vehicles based on edge cache and immune cloning strategy," Ad Hoc Networks, vol. 138, p. 103012, 2023.

[8] M. San Emeterio de la Parte, J.-F. Martínez-Ortega, V. Hernández Díaz, and N. L. Martínez, "Big Data and precision agriculture: a novel spatio-temporal semantic IoT data management framework for improved interoperability," Journal of Big Data, vol. 10, no. 1, p. 52, 2023.

[9] H. Liang, Z. Zhang, C. Hu, Y. Gong, and D. Cheng, "A Survey on Spatio-temporal Big Data Analytics Ecosystem: Resource Management, Processing Platform, and Applications," IEEE Transactions on Big Data, 2023.

[10] M. Li et al., "A Survey of Multi-Dimensional Indexes: Past and Future Trends," IEEE Transactions on Knowledge and Data Engineering, 2024.

[11] Z. Liu and Z. Zou, "Analysis of network topology and deployment mode of 5G wireless access network," Computer Communications, vol. 160, pp. 34-42, 2020.

[12] Z. H. Meybodi et al., "Multi-content time-series popularity prediction with multiple-model transformers in MEC networks," Ad Hoc Networks, vol. 157, p. 103436, 2024.

[13] M. K. Alsmadi, "Content-based image retrieval using color, shape and texture descriptors and features," Arabian Journal for Science and Engineering, vol. 45, no. 4, pp. 3317-3330, 2020.

[14] U. A. Khan, A. Javed, and R. Ashraf, "An effective hybrid framework for content based image retrieval (CBIR)," Multimedia Tools and Applications, vol. 80, no. 17, pp. 26911-26937, 2021.

[15] X. Chai, Y. Wang, Z. Gan, X. Chen, and Y. Zhang, "Preserving privacy while revealing thumbnail for content-based encrypted image retrieval in the cloud," Information Sciences, vol. 604, pp. 115-141, 2022.

[16] J. Anju and R. Shreelekshmi, "A faster secure content-based image retrieval using clustering for cloud," Expert Systems with Applications, vol. 189, p. 116070, 2022.

[17] M. K. Kelishadrokhi, M. Ghattaei, and S. Fekri-Ershad, "Innovative local texture descriptor in joint of human-based color features for content-based image retrieval," Signal, Image and Video Processing, vol. 17, no. 8, pp. 4009-4017, 2023.

[18] M. J. Iqbal, M. Farhan, F. Ullah, G. Srivastava, and S. Jabbar, "Intelligent multimedia content delivery in 5G/6G networks: a reinforcement learning approach," Transactions on Emerging Telecommunications Technologies, vol. 35, no. 4, p. e4842, 2024.

A Secure and Efficient Framework for Multi-User Encrypted Cloud Databases Supporting Single and Multiple Keyword Searches

J V S Arundathi[1], Dr. K V V Satyanarayana[2]
Ph.D Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[1]
Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[2]

***Abstract*—Multi-user encrypted cloud databases have become essential for secure data storage and retrieval, especially when supporting both single and multiple keyword searches. Ensuring data confidentiality, integrity, and efficient access within such systems is paramount, particularly when dealing with multiple data owners and users. This paper presents a Secure Encrypted Trie-based Search (SETBS) method that significantly enhances multi-owner authentication, data secrecy, and data integrity in cloud environments. The SETBS framework leverages a sophisticated Merkle hash tree for dynamic maintenance and autonomous user verification, ensuring that the identity of users is reliable and that personal information remains protected across various ownership domains. By optimally utilizing resources, SETBS provides a robust and efficient solution for managing data in cloud environments. The framework addresses the bottleneck issue by distributing the workload among first-level owners, resulting in fair resource distribution and increased system efficiency. A key feature of the SETBS method is its ability to guarantee data integrity without compromising security. Users can be assured that their data remains unaltered and protected from unauthorized access, thanks to the integration of the Merkle hash tree. This mechanism enables clients to confirm the integrity of their data stored in the cloud, providing peace of mind regarding its security. Moreover, SETBS proves to be a flexible and scalable solution for large-scale cloud deployments, efficiently managing multiple data owners and parallelizing the processing load. The framework's focus on data privacy ensures that personal data remains secure during search operations. With lower encryption and decryption times compared to existing methods such as SPEKS, DSSE, and MKHE, SETBS demonstrates superior performance and is implemented in Python. This comprehensive approach offers an all-encompassing solution for businesses seeking to enhance their cloud security architecture while ensuring efficient data management, from processing to real-time or batch data analysis.**

Keywords—Secure keyword search; encrypted search; multi-user framework; encrypted cloud database; single and multiple key users

I. Introduction

The well-known advantages of outsourced data retrieval and sharing—convenient keeping and on-demand access—have made it popular in the big data era. The global web has developed so quickly that the huge data age has arrived. With the increasing data generation in daily life, cloud storage technology is developing. Cloud computing's exceptional benefits (including reduced expenses, improved work, effectiveness, and safety) have made it popular in the last few years [1]. Databases, which are storage, and servers are among its offerings. With a cloud storage system, individuals and file holders can examine the data remotely once it has been stored on cloud servers. This raises concerns about the confidentiality of data because the cloud server has access to the data and searches [2]. Employing safe cloud databases was the strategy of this paper in offering a hybrid model that tackles the issue of safe and effective information recovery in multi-user environments. The very purpose is the production of a structure that will maintain the confidentiality of information and will at the same time be able to be searched for by consumers using one or more keywords. Credible authentication techniques, which are very efficient with protected information, and work across many people, are only one of the main goals. Atlantis is a unique structure whose sole purpose is to increase the functionality and security of storing, retrieving data, and enabling research over the network, while maintaining safety, ensuring the prompt and precise recovery of data that is well-chosen, encrypted. Besides, it shields out hacker attempts by allowing several individuals to do searches at the same time. The recommended method is created to be the easiest and the one that is best for use in business settings where the efficiency and confidentiality are essential. The goal of this system is to provide a safe and reliable way for the recovered data from the cloud to combine the effective search algorithms as well as protection approaches. Along with these select search protections, developers must ensure that they choose the right encryption methods, have strict security measures in place, use fast, to create this innovative hybrid architecture, the organization would need to factor in the implementation and optimization of the algorithms and protecting the confidentiality of the information, as well as ensuring parallelism and adaptability, proper installation, and compliance with safety standards, amongst others.

Security and functionality are the goals of computer network models that let multiple users search encrypted data in cloud databases. Han et al. [3] developed a novel software-assuring technique which employs a combination of customizable encryption techniques attached to a public Blockchain to enhance strength against attacks on servers. Blockchain's unique approach to decentralized computing secures protection from system intrusions and only allows restricted access-awareness. Investigated on the Ethereum blockchain, the plan turned out to be much better in terms of effectiveness than traditional methods, the one big advantage was the cost-effective searchable encryption. However, issues such as heavy file processing and the demand for additional internet resources are the main imperfections. The issue of cumbersome computation needs and internet resources is a great problem in the cloud. Only the challenges are alluring Han et al.'s approach introduces a step closer to the full implementation of the said cloud technology. It already provides rapid and secure data transfer to online users who maybe just residents in New York or the Illinois area. Moreover, under security concerns cryptography protocols were applied. Cui et al. [4] gave the first account of the Multi-User Safe and Verifiable K Nearest Neighbor (MSVkNN) search idea that deals with the privacy and integrity of the multithreaded areas as included among the K Nearest Neighbor queries in the cloud-based location services. A concise and improved version is here the abstract of this framework, which makes use of, the Verifiable and Secure Index (VSI) structure, and its corresponding protocols have been designed to safeguard the issue of data, the queries, the results, and the access patterns. Thanks to VSI technology, it is also possible to delink the ownership of private information, which makes it a privacy-protecting method in both respects, that is, it allows anonymity of the query and the completeness verification of the result. Critically, the MSVKNN method supported the possibility of using multiple keys for different users, which led also to improving the security through rigorous security proofs and empirical testing. In the words of Cui et al., the solution provided by their approach is exhaustive and it not only protects the privacy and integrity of the searches of cloud systems which are based on the new trends in IAR but it can also be applicable in the more complex environments of multi-user.

Blockchain technology is applied in the experimenting of the use of database architectures together with indexed structures by the authors in the ongoing review articles. New approaches are proposed to be used in the data retrieval process as well as through the privacy issue. The integrity of data and confidentiality become stronger when blockchain becomes the security platform, therefore, important data such as those regarding shipments of essential supplies are protected. The fault is, thereby, shifted by the use of data-mining technologies to resolve usually requiring operations away. Traditional research problems such as allocating resources and processing overhead are overcome by the research groups through their continuous improvement of these strategies. The success of search processes involving encrypted data in cloud computing through the deployment of more effective and more secure encryption techniques is based on the success or failure of the initiatives. In this context, the enhancements offered by data encryption search systems also bring about cloud computation systems to the right procedure, which is more secure and more efficient, hence, promising additional security for increasingly expanding Data access/retention embedded in the changing research.

The main breakthrough of that study is the formulation of a newly established hybrid model meant for the secure and efficient management of cryptographic information in a cloud database that is accessible by multiple users search by single or multiple keywords. The design has been modified such that the latest security requirements are met, and the design that already exists will be coordinated with the current cloud service. This allows it to be the proof that not only do those regulations guarantee the protection of the data but also meet the industry standards and the operability with several cloud providers that in turn create the possibility for using it in different corporate environments without risk of both safety and productivity.

- Introduces an authenticated Search method for multi-owner cloud environments namely Secure Encrypted Trie-based Search that keeps the data secrecy and integrity up to the mark.
- Implements a Merkle hash tree of the latest generation for dynamic operation and self-organizing user verification, and guarantees the authenticity of users and the confidentiality of their data.
- Enhances the management of resources in SC environments, and solves bottleneck problems through an equitable share of the tasks with the first-level owners enhancing the system efficiency.
- Offers a strong and reliable avenue in which the users can verify data stored in the cloud and the security of such data.
- Shows that SETBS is easily scalable and implements well for large-scale applications in the cloud to handle numerous data owners and distribute loads.

The paper is organized as follows: Section II presents other works regarding multi-user encrypted cloud databases and the requirement of having a secure database and efficient storage of data. Problem statement of the paper is given in Section III. The proposed SETBS method is described in the subsequent Section IV. The findings of the paper have been presented and analyzed in the last section known as Section V. Section VI closes with an evaluation of the framework to resource allocation, followed by a consideration of the former for scalability of resource management for massive Cloud provisioning.

II. Related Work

Yoon et al., [5] articulated one of the most innovative advances in the state of the data leveraging Intel SGX as a trusted executor platform. This method decreases computation and communication costs by practically rendering cryptographic processes to its specific enclave in combination with a trusted execution environment (TEE) to prevent cloud-centric side-channel attacks. The architecture benefitting from SGX, SPEKS (Secure and Privacy-Enhancing Keyword Search), is a cryptosystem that allows safe keyword search over the cryptographic texts through the SGX enclave that stores the data. Encryption and geological data security as well as the cryptographical technique is said to be the study's solution which is the way of the magnification of secret key cryptography in computer-generated and general intellectual processes even though it has no reference to the database. The problem with the SPEKS is data security and its necessity for an SGX-capable device that may hamper its usage in devices that do not have these features. Accordingly, the SPEKS approach allowed the computing time for the PEKS, Trapdoor, and Search algorithms to be reduced dramatically. For example, the PEKS calculation durations have gone down from 8.123 ms in previous systems to just 0.0919 ms. Furthermore, it adjacently embeds franchised transmission and volatile space, therefore, posted offers secured keyword searches in encrypted settings. These developments clarify that SPEKS is a dependable choice that will be - ever since it is - safe for choice for user-friendly and malware-free keyword searches in the latest cloud system.

Liu et al., [6] The IKGAs were scrutinized closely using a lattice-like PEKS system chip to examine the advanced IKGAs. The researchers pay special attention to the PEKS method by Zhang et al. which is called FS-PEKS PEKS for the generation of random files of equal length code for the IKGAs. On the one hand, as said by Liu et al., the security layer, [fs-peks] based on the generation of several derived vulnerabilities to the design is why the intrinsic vulnerabilities of the secure layer of the FS-PEKS that can be derived are the main reasons for the success of the vulnerability research done by the two researchers. Also, the security layer that can be removed at several points of the design, has been referred to as various derived vulnerabilities by the calling. As for the testing of the protocol's consistency and the recognition of threats, the investigation applies such algorithms by changing variables and using some technical skills. A security threat concerning the IKGA inside the system architecture is being reported. The basic principle of dishonest employees' designs to detect the keywords in the system is explored in the paper. They pointed out that none of these technologies can be used in reality, and even if they were used, it would take nature longer time to recover than usual. Their findings imply that more robust PEKS that can deal with insider attacks are requisite for a secure keyword search on encrypted data in sensitive and risky areas like the IoT.

Bulbul et al., [7] suggested a study that tests the suggested plan using the Enron dataset, which consists of a substantial amount of email correspondence between Enron personnel. Dynamic Searchable Symmetric Encryption (DSSE) is the framework that is being employed; it is intended for usage in multiple-user scenarios. Through the use of symmetrical encryption, keyword-based combat, and random number generation for key updates following every query, it guarantees forward as well as reverse secrecy. One of the noted limitations is that creating a doorway requires a longer period than alternatives since reward positions for particular keywords must be determined before the analysis. Despite this, the DSSE method boasts from low connectivity overhead, database creation, and query effectiveness. This studies' findings show that the DSSE approach works better than other approaches in terms of index creation and query effectiveness, proving its usefulness and low-weight architecture in networked settings. Extensive empirical assessments confirmed the approach's exceptional effectiveness in real-life situations and its extraordinary effectiveness in index production.

Li et al., [8] explained a strategy for multi-key homomorphic encryption, which expanded key in the technique, which is based on the DGHV encoding framework, enables secure homomorphic calculations involving various consumers. The current MKHE systems and this improved DGHV program are contrasted. The computational efficiency measurements of many MKHE systems across different security parameter levels (Toy, Small, Medium, Large) make up the assessment dataset. The research's algorithm is an enhanced version of the DGHV technique that has been optimized to decrease the public key space and boost computational efficacy. According to the findings, the suggested MKHE scheme is more effective concerning of computational difficulty and capacity. Compared to current systems, the period needed for key generation and homomorphic operations (ciphertext extension, addition, multiplication, and decryption) is greatly decreased. For example, the suggested system takes 0.05 seconds to decode at the Moderate protection point, whereas ciphertext extension and multiplication take 0.98 and 3.34 seconds, correspondingly. The necessity to interface with CP-ABE for improved DU assign reliability and optimize the ACC's attribute verification algorithm to save petrol expenses are obstacles, nevertheless.

Liu et al., [9] suggested a plan using the Enron dataset, an extensive set of emails exchanged between Enron workers. For analysing retrieval of data algorithms, this dataset offers an accurate and varied measurement. The Key Generation Centre, Cloud Platform, Internal Servers, Data Providers, and Request Users are some of the stakeholders involved in the decentralized framework that is being deployed. The technique maintains privacy and confidentiality while supporting inquiries with multiple keywords. Potential access pattern leakage when users obtain records from search outcomes is one of the suggested scheme's limitations. Furthermore, relative to different systems, the method of searching could take longer, especially as the number of terms rises. The findings demonstrate that the recommended approach outperforms compared approaches in these areas and is efficient in index creation and gateway computing. However, owing to network delay, the search procedure can take longer. Because the technique protects data, search, and search pattern privacy, it may be used in actual-life situations where effectiveness and safety are crucial.

Cui et al. [4] proposed the Multi-User Safe and Verifiable k Nearest Neighbor search (MSVkNN) which is a k-nearest neighbor query architecture in a cloud database environment that is both secure and verified. At this time, large-scale data that might be brought into play for KNN searches. The proposed model takes advantage of multiple users joining the process of inquiring cloud storage data and integrates public-private cryptographic methods to secure data and verify query results. The writer, however, emphasizes that the design also carries some drawbacks, such as discontinuity problems that have an impact on real-time applications and the computational costs of encryption and verification procedures. The result shows that the architecture maintains strong security and verifiability while achieving high query precision. Nevertheless, the proposed model lacks to access the pattern of privacy protection.

Several methods have been developed recently for safe keyword searches on encrypted data. SPEKS architecture lowers computing expenses and improves secrecy by using hardware that supports SGX. Vulnerabilities in lattice-based PEKS systems are assessed, exposing shortcomings in existing methods. Despite requiring longer setup times, DSSE exhibits efficient index generation and query performance in multiple user scenarios. Enhanced MKHE technique improves effectiveness and safety in cloud computing. Issues about access frequency leaking are addressed by a decentralized design that prioritizes capability and privacy in keyword searches utilizing the Enron dataset.

III. Problem Statement

Data privacy and query confidentiality in cloud databases are guaranteed by an effective encrypted multi-user system. The current approaches frequently have issues with scalability, safe efficient keyword search through encrypted data and control of access. The other existing models have limitations such as only SGX-enabled hardware is applicability[5], susceptible to insider threats, jeopardizing the integrity of the network [6], longer index construction time as a result of determining keyword placements [7], Problems with combined integration and attribute confirmation optimization [8], Possible problems with computing costs and applicability in real time [4]. The proposed approach is perfect for commercial applications requiring high secrecy and efficiency since it guarantees authorized, secret data access for numerous users, securely enables concurrent queries, and handles cloud-based multi-user problems.

IV. Proposed Secure Encrypted Trie-Based Search (SETBS) Method

Secure Encrypted Trie-based Search (SETBS) improves multi-owner authentication, data secrecy, and data integrity in cloud contexts. It uses a sophisticated Merkle hash tree for dynamic maintenance and autonomous user verification. The framework presented is a way to make the identity of a person more reliable and to protect personal information from a variety of different owners that in the end make available fine and secure data search operations. SETBS is a very robust framework for data management in cloud environments that are both safe and efficient since it optimally uses the resources and guarantees the data processes. Through the distributed scheme of the work over the first-level owners, the contribution to the fair resource distribution becomes the highest, the bottleneck issue is resolved and the efficiency of the whole system increases. A Merkle hash tree is a vital element of this approach. This algorithm helps users to know their data integrity without endangering the security. This shows that clients might have a guarantee and confirmation of the data stored in the cloud still being the same without anybody unauthorized touching it and the data is fully protected in this way. Set-Aside Transaction Broker Services (SETBS) are a flexible and feasible choice for big cloud setups as they may manage a multitude of data owners efficiently and parallelize the load. What is more, the focal point of privacy of data presupposes that personal data will be secure in the exploration process. SETBS thoroughly complicates the safety and authenticity of cloud computing systems by harmonizing these traits together. When businesses want to enhance their cloud security architecture, one all-inclusive approach offered for data management in the cloud consisting of such issues as data integrity, confidentiality, and efficient resource use ranging is from data processing to analyzing real-time or batch data.

The procedure of the suggested SETBS model should be presented. The graphic of the data processing pipeline is shown in this Fig. 1, which comprises of data collection, AES encryption, trie-based keyword search, cloud storage, and performance analysis.

A. Data Collection

The KOSARAK dataset provided by an online news portal from Hungary was an unstructured dataset of 41,269 unique items and 990,000 entries. Each record corresponds to the interaction of user interaction. The dataset consists of smallest and largest sequences, which are 1 and 2,498 items long, respectively, along with an average sequence length of 8.10 items per sequence. At the news portal, this data set tell about the order in which pages or articles are viewed, thus, in brief, it gives details about navigation patterns by portraying user movements [1]. User engagement is recognized, and the effectiveness of content, in addition, website techniques and content recommendation software are evaluated using this data [10].

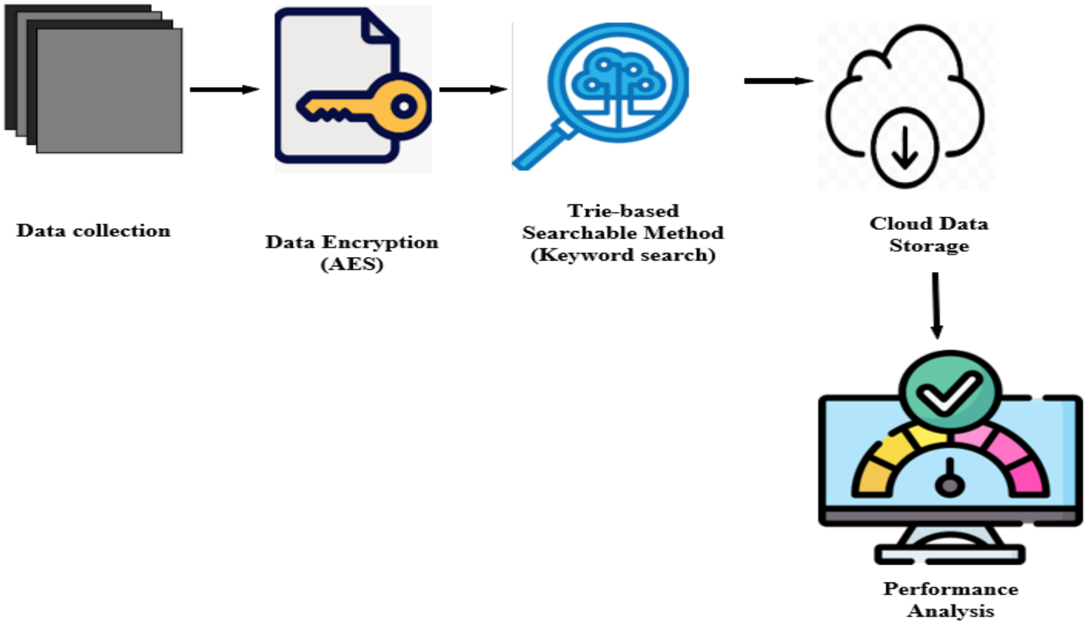


Fig. 1. Block diagram of proposed SETBS model.

The experimental setup included a thorough description of the methodologies employed, focusing on the KOSARAK dataset obtained from a Hungarian online news portal. This dataset, consisting of 41,269 unique items and 990,000 entries, captures user interactions, reflecting the order in which pages or articles are viewed. With sequence lengths ranging from 1 to 2,498 items and an average length of 8.10 items, the dataset effectively illustrates user navigation patterns. Additionally, the analysis emphasizes user engagement and evaluates the effectiveness of content and website strategies, providing a robust foundation for assessing the proposed model's performance and validity.

B. Data Encryption

Data encryption is the process of converting plaintext information into an unreadable format using an algorithm and an encryption key, which can only be decrypted by authorized parties with the help of the decryption key. The purpose of this is to protect your data from unauthorized access and cyber threats [11]. This survey introduces Advanced Encryption Standard (AES) for data encryption. Encryption algorithms are often based on mathematical calculations to convert plaintext to ciphertext. The following is an often-used symmetric encryption Eq. (1) that is inserted below:

$$Cliphertext = AES_{Encrypt}(Plaintext, Key) \qquad (1)$$

where, is a symmetric encryption algorithm called AES Advanced Encryption Standard.

1) Advanced Encryption Standard: AES - Advanced Encryption Standard is a symmetric encryption algorithm, and one of the most commonly used ones when it comes to the transfer of data of a sensitive nature. In the context of encrypting data in a dataset as large as KOSARAK, such an algorithm is operated by splitting the data into fixed-size blocks, for instance 128 bits each in the case of AES-128 [12]. The Eq. (2) for AES encryption is described in below:

$$E_k(P) = C \qquad (2)$$

where E_k represents encryption with key k, P is the plaintext, and C is the resulting ciphertext.

Here's a detailed explanation of how AES encrypts data:

a) Key expansion selection: AES requires a key for encryption and decryption, typically 128, 192, or 256 bits in length. Before encryption begins, the key undergoes an expansion process to generate a set of round keys, which are used in each round of the encryption process [13].

b) Initial round key addition: Each block of plaintext (or data) is divided into smaller blocks called state arrays. AES starts by performing an initial round key addition. Here, the state array is combined with the first-round key using a bitwise XOR operation.

c) Rounds of substitution and permutation: AES operates through multiple rounds (10 rounds for AES-128) of substitution and permutation.

d) Substitution: Bytes in the state array are substituted with corresponding bytes from a substitution box (S-box), which is pre-defined in the AES specification. This step adds confusion to the data.

e) Permutation: Rows in the state array are shifted cyclically, and columns are mixed using a matrix multiplication operation known as the Mix Columns step. These operations introduce diffusion in the data.

f) Final round: The final round of AES excludes the Mix Columns step to simplify the implementation. Instead, it consists of substitution, permutation, and a final round key addition.

g) Cipher text generation: After completing all rounds, the resultant state array, now transformed through substitution, permutation, and key addition, is the cipher text. This encrypted data is output as the final result of the AES encryption process [14].

C. Multi-Owner Authentication

This solution guarantees the cloud server database by applying a highly protected multi-owner authentication method. The data proprietor first uploads the data in an encrypted way employing an enhanced Merkle hash tree technique to the cloud server. The user is given a public key to see and download the data. The data owner uses the public key to confirm if the user is authorized or not. The data owner gives the user a decryption key if the user permits so they may decode the data. To process the work that the user wants done, a load-balancing notion is also implemented. The user suggestion is finally sent to the cloud server. The user research is answered by the cloud server if the user has effectively [15].

1) Data owner: The data owner must first register on the server of the cloud service provider. Following registration, the data owner receives both public and private keys generated by the cloud service provider. The relevant data are uploaded to the cloud server after being encrypted using the Enhanced Secure storing a lot of data, the cloud service provider also controls user and data owner identification [17]. The cloud server routes the user-requested work to any queue to process it. Virtual machines are used to handle user requests in the queue.

2) Third-party auditor: To verify the integrity of the data stored in the cloud, TPA is applied to encrypted cloud data. The data is submitted by the data owner, and the TPA audits the data upon request. To audit user-requested data, the TPA must register with the cloud server. After bilinearly mapping the user data, the cloud provides the evidence to the TPA. The data supplied by the user and the data from the cloud are compared by the TPA. The file's security is safeguarded by the encrypted data.

3) Bilinear mapping: The first stage of encryption to express the data in mapping form is called bilinear mapping. The input of the data is its cyclic group, and the bilinearly mapped output is denoted by the letter e. Think of group G as a prime order p gap Diffie-Hellman group. Assuming that GT is a prime order multiplicative cyclic group, a bilinear map is created [18]. The following characteristics of a useful $e: G \times G \rightarrow GT$, where GT is a prime order, multiplicative cyclic group. A useful e has the following (3), (4), (5) properties shown in below: To change the default, adjust the template as follows.

4) Encrypted Trie-based search (SETBS) technique[16]. A public key is used by the sender to encrypt data, and only the holder's private key may be used to decode the appropriate data.

5) User: To utilize the network, the user must first register for an account. The user creates an account, signs in, and asks the cloud service provider to review their account. The cloud service provider will process the work at the request made by the user. Languages for programming like Java and NET are used by the network to communicate with the cloud server. By contacting the cloud service provider, the user may obtain the desired data. The client retrieves cloud data using the provided private key for Secure Encrypted Trie-based search (SETBS).

6) Cloud service provider: The cloud service provider offers flexible online processing and storage of data by combining hardware and software resources. In addition to

$$Bilinearity - \forall m,n \in G \Rightarrow e(ma,nb) = e(m,n)^{ab} \quad (3)$$

$$Non - degeneracy — \forall m \in G, m = 0 \Rightarrow e(m,m) = \quad (4)$$

$$Computability — e \ should \ be \ efficiently \ computable \quad (5)$$

It is possible to express data using a two-dimensional vector is known as bilinearity; the ability to degenerate data back to its original form is known as non-degeneracy; and the capacity to solve a problem effectively when a and b are real random values is known as computational capability [19].

7) SETBS: Cloud databases with encryption provide a complete solution for secure and effective multi-user keyword searches with the Secure Encrypted Trie-based Search (SETBS) paradigm. The SETBS model offers different ways of protection and is, therefore, more functional and secure, since it is secured through keyword indexing, access control, and trie data structure [20]. Its ability to handle searches using one or more keywords and provide search results that others can check makes it a key addition to searchable encryption in cloud computing. The Secure Encrypted Trie-based Search (SETBS) system offers a reliable and scalable answer for encrypted cloud databases. This model uses a trie data structure because it works well and has potential. To protect the security and integrity of search queries and results, it also combines trie data structure with encryption methods. The SETBS model tries to solve the challenge of supporting both single and multiple keyword searches while meeting strict function and safety needs [21].

8) Trie-data structure: The SETBS model relies on the trie-data structure to organize and index keywords. In this structure, each node stands for a character in a term. Complete keywords are shown by paths that end at leaf nodes. Large volumes of encrypted data may be easily managed with the trie's hierarchical structure, which enables quick keyword lookup and retrieval. Verified Data Structure MTrie. Our proposal for MTrie, a new authenticated data structure based on MHT and Trie, is presented in this part of the paper. This study provides the relevant analysis of every node in the Trie, by the idea that MHT uses an organizational hash mechanism to offer query-based verification.

To obtain the root signature Sroot, as indicated in the equation below, DO first computes the hash value of each MTrie node. Then, using the secret key sk, it signs the hashed level of the MTrie root or Sroot in Eq. (6).

$$s_{root} = si_{sk}^{g}(h_{root}) \quad (6)$$

where s_{root} is produced by the use of a secret key signature technique si_{sk}^{g} the hash of the root value h_{root} The legitimacy and reliability of the root value are guaranteed by this procedure.

D. VO Construction

CSP employs Algorithm 1 to compute the query findings create the VO according to the query outcomes and execute the approximation string query q. In particular, VO contains the following four categories of data: utilizes Algorithm 1 to extract the last VO in the following manner, considering the query string q = "inf" and the distance to edit threshold d = 1.

$$VO = \left[I\left[*n[*n,*t]\right],t\left[(e,h6),(o,h9)\right]\right] \quad (7)$$

The query outcome set {"in","inn","int"} that satisfies the distance to edit threshold criterion has been incorporated in the VO, which is at last returned to the user via CSP together with the sign of the root node.

Initializing keys and an empty trie structure is the first step in the process. Before entering a keyword into the trie, it is hashed and encrypted. To find correspondence, search tokens are created and compared to the trie. Results include decrypting, compiling, and presenting matching data. Before revealing results or pointing up security flaws, tests make sure that reliability, honesty, and secrecy are met. This technique guarantees reliable data recovery while upholding strict security protocols during the process of searchable encryption.

Fig. 2 is a visual representation of the interaction between a client and an encryption module that encrypts data and places it in a trie data structure of the cloud server. The user's search queries are sent to the server, which processes these queries. Decrypted answers are sent back to the users by means of a decryption module.

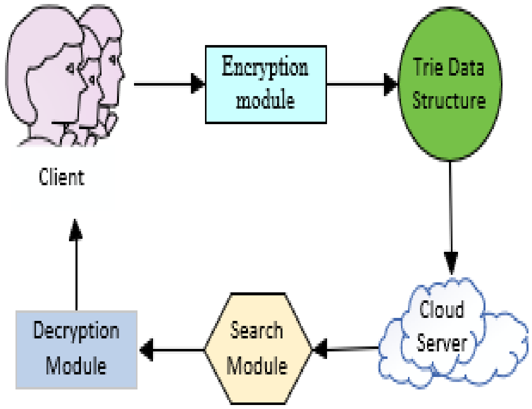


Fig. 2. System model.

The focal point of the proposed framework is bettering the safety through the gaining of a protected storage and data tool as well as the realization of the abilities of cloud databases to be a source of the information. The try-crypto-organized data structure, which does the encryption of the keywords hierarchically, that, allows for the function of search in fast and effective manner is one of the major building blocks of the system. Incoming up with an encrypted word search system where each keyword is encrypted using a safe cryptographic procedure before being put in the trie is an effective strategy. Users have to prove their authentic identity and then are assigned only the authority to execute the search a part of the procedure in order to make sure that only those with the right to a specific access can carry out searches and see data. This is accomplished by constructing the tree structure, which can go through the encrypted keywords and the matching procedure for both single and multiple keywords. The structure guarantees that the process of searching is not discerned and the operation goes on without the user decrypting the data, hence the data is safe from being read by unauthorized people. The protected query is then run through hashing to the encrypted trie. After that, the user receives the encrypted results which they may return to their plaintext using their private key [22]. One of the features of the infrastructure is a powerful access control system. It is the system which manages the different rights and access of the users, and it supports the presence of many users as well. It even ensures that the users' search queries and results are top secret and there are no unauthorized accesses. Additionally, this approach is useful in that it increases the safety of the cloud database applications but on the other hand, it also optimizes the search performance.

V. Results and Discussion

The result section provides a holistic overview of the study findings of Secure Encrypted Trie-based Search (SETBS) technology. It describes the method's advantages over SPEKS, DSSE, and MKHE. Since it is clear. Speaking about the result, which is the performance of SETBS in comparison to the other ones, SETBS being properly clear and fast on its own was more par than this one. Now, it confirms the right way for the technical development to ensure that SETBS excels in achieving fast and scalable cloud database processing. The set of techniques specifies SETBS's benefits when compared to the traditional techniques and presents the application ideas in all areas. For enterprises who want faster search and high scalability in a cloud environment, SETBS is a potential alternative since it enables faster search speeds and scalable without damaging privacy. These results demonstrate that this feature of SETBS is very useful and the injecting of the searchable security algorithms into data-intensive computing.

A. Performance Evaluation

The performance evaluation of the framework is achieved through these metrics in cloud systems operated by more than one person. The accuracy point guarantees correct results, the scalability point determines the system's possibilities of development, the security point stands for the data that is safe and the retrieval time fact checks the operational efficiency. They jointly offer a profound insight into the usability and robustness of the system in cloud settings.

TABLE I. Performance Metrics of Proposed Model

Metrics	SETBS (proposed)
Single Keyword Search Time (ms)	3.456
Multiple Keyword Search (ms)	5.678

Table I easily demonstrates the performance measures of the systematic mechanisms being the empirical results of SETBS. The search times are enclosed in the measurements both for single and multiple keywords. The search time of a single keyword is 3.456 milliseconds and this figures out the framework single-keyword queries quickly. The average response time for hunts with multiple keywords is longer at 5.678 milliseconds, once again showing its ability to handle not only simple queries but also more complex ones very fast. This shows the complete accuracy of SETBS which is used in cloud storage solutions and gives a snapshot of the search process to the user even in the shortest time.

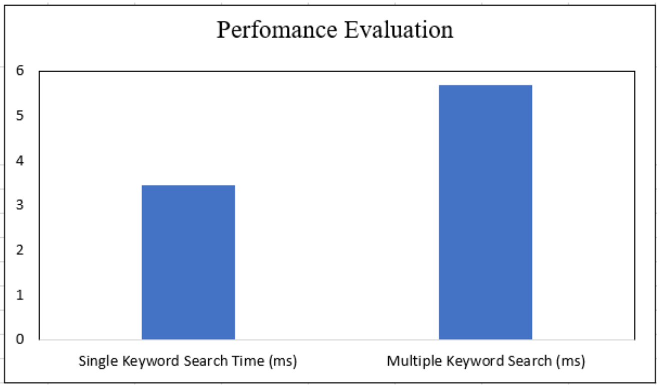


Fig. 3. Performance evaluation of proposed model.

In Fig. 3, the diagram symbolizes the search times of the search engine model SETBS, which was designed because of the one-day motivational improvement of, search times for both single and multiple keywords. The x-axis represents the type of search, while the y-axis denotes the search time in milliseconds. Two bars demonstrate the search times: 3.456 ms for a single keyword search and 5.678 ms for several keyword searches. The chart below indicates the performance of the SETBS model in terms of how quickly it handles searches with several keywords, which is explained above.

TABLE II. Proposed Model Comparison with Existing Models

Methods	Single Keyword Search Time (ms)	Multiple Keyword Search (ms)
SPEKS	8.123	10.234
DSSE	5.789	7.345
MKHE	6.912	8.567
SETBS (proposed)	3.456	5.678

The SETBS methodology greatly reduced the time required to get data for both single and multiple keyword searches when compared to other methods. The comparison of retrieval times for various approaches is displayed in Table II. This table includes SPEKS [23], DSSE [24], MKHE [25], and the suggested SETBS, compares the millisecond search times of many models for single and multiple keyword searches. The speed at which each of the models can execute a search query is used to gauge its performance. The suggested SETBS model outperforms the existing models with the fastest search time of 3.456 ms for single-term searches. After the DSSE model once again, the times of 7.345 ms, 8.567 ms for MKHE, and 10.234 ms for SPEKS are the fastest. These findings show that, for both single and multiple keyword searches, the SETBS framework outperforms other models in terms of efficiency. SETBS appears to be a more efficient method, as seen by the notable decrease in searching times. This might result in enhanced performance and faster retrieval of data for systems that use encrypted keyword searches. SETBS continues to do exceptionally well when it comes to multiple keyword searches, recording the fastest search time of 5.678 ms.

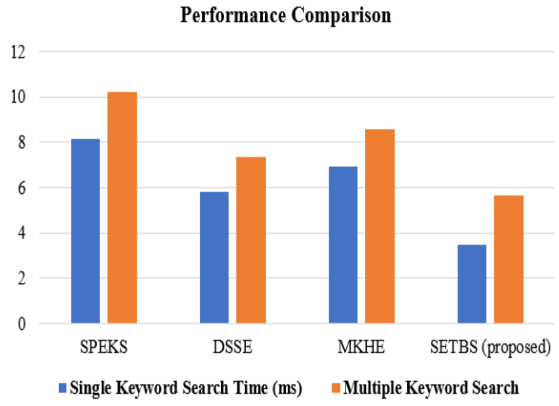


Fig. 4. Proposed model comparison with existing models.

The search durations for single and multiple keyword searches across four different frameworks —SPEKS, DSSE, MKHE, and the suggested SETBS—are graphically shown in Fig. 4. Two bars show the performance of each framework: an orange bar shows the multiple keyword search duration and a blue bar shows the single keyword search time (measured in ms). SPEKS shows comparatively long search times in this data, ranging from around 8 ms for single-word searches to about 10 ms for multiple-word searches. With a search duration of around 6 ms for single keyword searches and a little over 7 ms for multiple keyword searches, DSSE performs better than SPEKS. With a search duration of about 7 ms for individual keyword searches and 8.5 ms for multiple keyword searches, MKHE has search times that are lower than SPEKS but higher than DSSE. The proposed framework exhibits the fastest search times and greatest efficiency. For single keyword searches, it takes around 3.5 milliseconds, while for multiple keyword searches, it takes about 5.5 milliseconds. The SETBS model is a superior choice for quick and effective encrypted keyword searches than the other models, as this figure demonstrates. It performs better than the existing models in both single and multiple-keyword search cases.

B. *Running Time of Proposed Model*

The running time of the Proposed SETBS method depends on several factors including the size of the trie, complexity of the search queries, and efficiency of cryptographic operations. Typically, it involves constructing the encrypted trie, which may incur an initial setup cost, and then performing encrypted search operations, which are generally efficient due to trie's logarithmic search time relative to the size of the trie. Overall, SETBS aims for practical efficiency while ensuring secure keyword search capabilities in encrypted data environments.

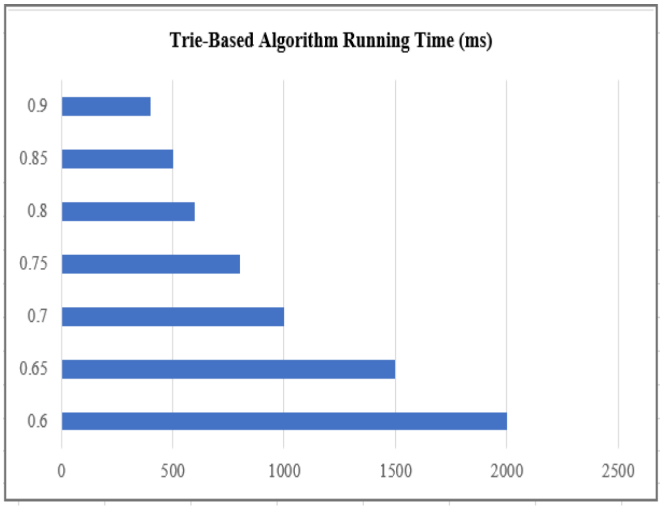


Fig. 5. Running time of the proposed model.

Fig. 5 demonstrates the given values denote the trie-based searchable encryption framework's running time for different values of (t). The duration of the run reduces with increasing (t), indicating that greater amounts of (t) optimize the search process and shorten the time needed to retrieve keywords. The trend of decreasing running time with increasing (t) presents the trie-based approach's efficiency in analysing encrypted search queries.

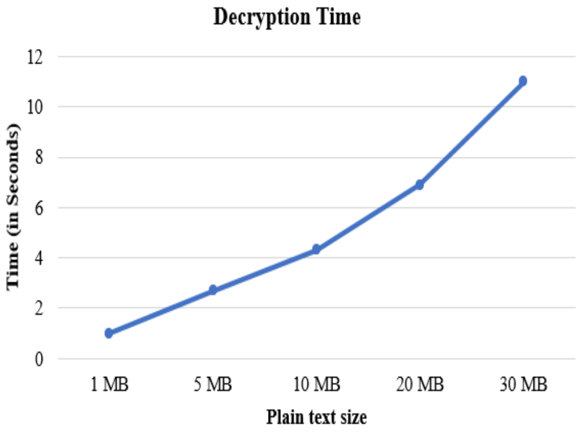


Fig. 6. Decryption time of AES.

The trie-based searchable decryption framework's processing time for varying data input sizes is depicted in Fig. 6. The framework takes one second for a one MB input, 2.7 seconds for a five MB input, and 4.3 seconds for a ten MB input. Processing 20 MB and 30 MB of bigger data inputs takes 6.9 and 11 seconds, respectively. This illustrates how the framework can be scaled and used effectively to handle different data quantities in encrypted cloud settings.

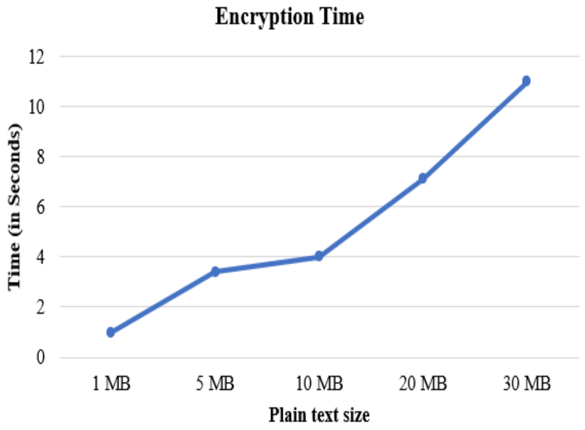


Fig. 7. Encryption time of AES.

Fig. 7 shows how long the trie-based searchable encryption system takes to handle different amounts of input data. The processing time is one second for an input of one MB. The timings grow to 3.4 seconds and 4 seconds, respectively, as the data input size increases to 5 MB and 10 MB. The processing speeds are 7.1 seconds and 11 seconds, respectively, for bigger inputs of 20 MB and 30 MB. These outcomes demonstrate how well the framework scales to accommodate varying data sizes.

C. SETBS Framework

Compared to conventional searchable encryption techniques, which frequently struggle to balance security and speed, the SETBS framework offers an important improvement in secure keyword searches for encrypted cloud databases. Its use of a trie data structure ensures efficient indexing and recovery of encrypted keywords; moreover, the hierarchical nature of the trie permits for rapid traversal, making both single and multiple-keyword searches practicable without sacrificing speed; additionally, the use of Advanced Encryption Standard (AES) ensures robust data protection against unauthorized access; finally, performance metrics reveal that SETBS achieves a single keyword search time of 3.456 milliseconds and a multiple keyword search time of 5.678 milliseconds, outperforming. For applications that need adequate safety and quick data retrieval, such as those in commercial cloud settings, this upgrade is essential. To further improve security, the framework's robust access control system makes sure that only authorized users may conduct searches. Overall, SETBS tackles the main issues with searchable encryption in multi-user cloud environments, providing a scalable and effective approach that satisfies strict security standards.

In evaluation to current searchable encryption techniques, SETBS extensively outperforms its friends in phrases of search performance and protection. For instance, whilst the SPEKS approach suggests seek instances of eight.123 milliseconds for unmarried key-word searches and 10.234 milliseconds for multiple key phrases, SETBS achieves a first rate three.456 milliseconds and 5.678 milliseconds, respectively. Similarly, other fashions like DSSE and MKHE display longer retrieval instances, with DSSE recording 5.789 milliseconds for unmarried keywords and seven.345 milliseconds for more than one key phrases, and MKHE at 6.912 milliseconds and 8.567 milliseconds. These metrics genuinely suggest that SETBS not best hurries up the search manner however additionally keeps robust security through its AES integration, making it a greater powerful alternative for applications in encrypted cloud environments. The better overall performance of SETBS positions it as a superior desire for customers desiring fast get right of entry to touchy data even as ensuring strict privacy and safety features.

To offer a greater complete and illustrative assessment, it's miles important to give no longer only the uncooked overall performance metrics but additionally contextualize them within actual international scenarios wherein every technique may be carried out. For instance, at the same time as SETBS boasts astonishing search times of three.456 milliseconds for single key phrases and five.678 milliseconds for multiple keywords, it's far important to spotlight that these rapid retrieval abilities enable users to speedy get admission to important statistics in time-touchy environments, which includes healthcare or economic offerings. In evaluation, SPEKS, with seek times exceeding 8 milliseconds, may additionally avert performance in programs requiring instant statistics retrieval, doubtlessly leading to delays in decision-making methods. Similarly, even as DSSE and MKHE display advanced overall performance over SPEKS, their retrieval times of 5.789 milliseconds and 6.912 milliseconds, respectively, nevertheless lag behind SETBS, which could be unfavourable in situations in which massive volumes of queries need to be processed quickly, consisting of in big statistics analytics. Moreover, incorporating qualitative components, together with personal experience and ease of integration into existing structures, similarly underscores SETBS's advantages, making it now not simplest a quicker choice but additionally a more sensible and consumer-pleasant solution in the realm of steady key-word searches in encrypted cloud databases.

D. Discussion

The SETBS framework represents a tremendous advancement in steady keyword searches for encrypted cloud databases, efficiently addressing the common demanding situations of balancing safety and pace that conventional searchable encryption strategies frequently face. By utilizing a trie records structure, SETBS enables efficient indexing and retrieval of encrypted keywords, facilitating short single and a couple of keyword searches without compromising overall performance. The hierarchical nature of the trie allows for speedy traversal, while the combination of Advanced Encryption Standard (AES) ensures strong records safety in opposition to unauthorized access. Performance metrics display that SETBS achieves marvelous seek times of three.456 milliseconds for single key phrases and 5.678 milliseconds for more than one key phrases, surpassing current strategies. This enhancement is mainly important for packages requiring both excessive protection and rapid facts retrieval, together with those in commercial cloud environments. Additionally, the framework's stringent get entry to manipulate system ensures that handiest authorized users can carry out searches, in addition strengthening its security posture. Overall, SETBS successfully resolves key issues related to searchable encryption in multi-person cloud settings, offering a scalable and efficient answer that clings to rigorous protection requirements [3].

VI. Conclusion and Future Work

In this paper, we present the SETBS method, which extends the security, as well as improves the efficiency of the multi-user encrypted cloud databases. In SETBS, the Merkle hash tree is well employed to enable dynamic addition/subtraction of owners and self-verification of the owners together with secure and efficient multi-owner authentication, non-disclosure of data, and data tamper proofing. The above framework shows significant enhancements compared to the previous techniques as follows: the encryption and decryption time is less therefore the framework proves to be more effective and viable solution on the cloud for managing the encrypted data. Some important issues like resource constraint, task performance consequences of bottlenecks, and inadequate data management are well solved by SETBS by properly controlling the load and work distribution to the first level owners. But it does so in a way that optimizes the general operational efficiency of the system in question; it also addresses the system bottleneck problem. The fact that one can verify the data consistency with reasonably high security gives additional confidence in data protection. In addition, the scalability of SETBS also allows for large scale and multiple data owners in a cloud and parallel processing in case of large volume. There are several directions for the further development of the SETBS framework which need to be investigated in future research: Firstly, it is possible to enhance the security aspects and performance of the SETBS with the help of the further incorporation of innovative cryptographic approaches. Moreover, applying the framework to analytically more demanding forms of queries and data representations might expand the area of its utilization and relevance. It was seen that extending SETBS to various cloud environment as well as putting it through different loads could give valuable indication of its stability and versatility. However, when linked to other cloud-based services and platforms, SETBS could present a more global solution of storing and protecting data. Finally, user studies and real-world implementations may offer insights into further improvements of the proposed framework and deal with certain issues that may arise with other applications.

The future scope of this studies encompasses several avenues for enhancement and application. Firstly, the framework can be improved to aid more complicated question kinds beyond unmarried and multiple key-word searches, doubtlessly integrating herbal language processing techniques to enhance user interaction. Additionally, the incorporation of device mastering algorithms ought to decorate the framework’s capacity to adaptively optimize seek efficiency based totally on user conduct and alternatives. Exploring the integration of blockchain technology could similarly bolster safety and transparency in multi-user environments. Furthermore, applying the framework to diverse domains, which include healthcare, finance, and e-commerce, can provide precious insights into its versatility and robustness in managing sensitive facts. Finally, carrying out good sized actual-international user studies will help validate the framework’s overall performance and user satisfaction in diverse operational contexts.

References

[1] Xu and Shiyuan, “Lattice-based Public Key Encryption with Authorized Keyword Search: Construction, Implementation, and Applications,” 2023.

[2] Y. Wang and D. Papadopoulos, “Multi-user Collusion-Resistant Searchable Encryption for Cloud Storage Yun Wang, and Dimitrios Papadopoulos,” 2023.

[3] J. Han, “Attribute-Based Access Control Meets Blockchain-Enabled Searchable Encryption: A Flexible and Privacy-Preserving Framework for Multi-User Search Jiujiang,” 2022.

[4] N. Cui, “Towards Multi-User, Secure, and Verifiable kNN Query in Cloud Database,” 2023.

[5] H. Yoon, “SPEKS: Forward Private SGX-Based Public Key Encryption with Keyword Search,” Nov. 2020.

[6] Z.-Y. Liu, “Cryptanalysis of ‘FS-PEKS: Lattice-based Forward Secure Public-key Encryption with Keyword Search for Cloud-assisted Industrial Internet of Things,’” Jun. 2021.

[7] S. S. Bulbul, “Fast Multi-User Searchable Encryption with Forward and Backward Private Access Control Fast Multi-User Searchable Encryption with Forward and Backward Private Access Control,” 2024.

[8] X. Li, “Privacy preserving via multi-key homomorphicencryptionincloud computing,” 2023.

[9] X. Liu and Y. Guomin, “Privacy-Preserving Multi-Keyword Searchable Encryption for Distributed Systems,” 2020.

[10] L. Jia, “A Trie Based Set Similarity Query Algorithm.” 2023.

[11] M. N. Ramachandra, “An Efficient and Secure Big Data Storage in Cloud Environment by Using Triple Data Encryption Standard,” 2022.

[12] M. Azhari, “Implementasi Pengamanan Data pada Dokumen Menggunakan Algoritma Kriptografi Advanced Encryption Standard (AES),” 2022.

[13] A. Kumar and C. Shantala, “An extensive research survey on data integrity and deduplication towards privacy in cloud storage,” Int. J. Electr. Comput. Eng., vol. 10, no. 2, p. 2011, 2020.

[14] F. S. Abas and R. Arulmurugan, “Radix Trie improved Nahrain chaotic map-based image encryption model for effective image encryption process,” Int. J. Intell. Netw., vol. 3, pp. 102–108, 2022.

[15] Y. WANG, “A Trie-Based Authentication Scheme for Approximate String Queries,” 2024.

[16] J. S. Jayaprakash, “Cloud Data Encryption and Authentication Based on Enhanced Merkle Hash Tree Method,” 2022.

[17] V. Melnyk, “Data Structures and Lookup Algorithms Investigation for the IEEE 802.15. 4 Security Procedures Implementation.,” in IntelITSIS, 2021, pp. 494–513.

[18] R. Subrahmanyam, N. R. Rekha, and Y. S. Rao, “Authenticated distributed group key agreement protocol using elliptic curve secret sharing scheme,” IEEE Access, vol. 11, pp. 45243–45254, 2023.

[19] X. Yang, T. Li, X. Pei, L. Wen, and C. Wang, “Medical Data Sharing Scheme Based on Attribute Cryptosystem and Blockchain Technology,” IEEE Access, vol. 8, pp. 45468–45476, 2020, doi: 10.1109/ACCESS.2020.2976894.

[20] L. Jia, “ATrie Based Set Similarity Query Algorithm,” 2023.

[21] C. Ghasemi, “Content Distribution over Named-Data Networks,” 2020.

[22] H. Zhong, Z. Li, J. Cui, Y. Sun, and L. Liu, “Efficient dynamic multi-keyword fuzzy search over encrypted cloud data,” J. Netw. Comput. Appl., vol. 149, p. 102469, 2020.

[23] S. Syväniemi, “Evaluating the fabrication and performance of sulfonated biochar composite membranes for copper redox flow battery,” 2024.

[24] M. R. Ahmed, J. M. Cano, P. Arboleya, L. S. Ramón, and A. Y. Abdelaziz, “DSSE in European-type networks using PLC-based advanced metering infrastructure,” IEEE Trans. Power Syst., vol. 37, no. 5, pp. 3875–3888, 2022.

[25] J. Park, “Homomorphic encryption for multiple users with less communications,” Ieee Access, vol. 9, pp. 135915–135926, 2021.

Exploring the Application of Neural Networks in the Learning and Optimization of Sports Skills Training

Dazheng Liu*
Yangzhou Polytechnic Institute, Jiangsu 225127, China

***Abstract*—Sports skills training is a crucial component of sports education, significantly contributing to the development of athletic abilities and overall physical literacy. It is essential to utilized neural networks to optimize traditional training methods that are inefficient and rely on subjective assessments. This paper develops methods for sports action recognition and athlete pose estimation and prediction based on deep neural networks. Given the complexity and rapid changes in sports skills, we propose a multi-task framework-based HICNN-PSTA model for jointly recognizing sports actions and estimating human poses. This method leverages the advantages of Convolution and Involution operators in computing channel and spatial information to extract sports skill features and uses a decoupled multi-head attention mechanism to fully capture spatio-temporal information. Furthermore, to accurately predict human poses to avoid potential sports injuries, this paper introduces an MS-GCN prediction model based on the multi-scale graph. This method utilizes the constraints between human body key points and parts, dividing the 2D human pose into different levels, significantly enhancing the modeling capability of human pose sequences. The proposed algorithms have been thoroughly validated on a basketball skills dataset and compared with various advanced algorithms. Experimental results sufficiently demonstrate the effectiveness of the proposed methods in sports action recognition and human pose estimation and prediction. This research advances the application of deep neural networks in the field of sports training, providing significant reference value for related studies.**

***Keywords*—Deep neural network; action recognition; 2D pose prediction; pose estimation; sports skill training; attention mechanism**

I. Introduction

Sports play an indispensable role in the cultural development of nations, serving not only as a key factor of citizen welfare but also as an important vessel for cultural identity. Recently, numerous policies have been published to encourage public participation in sports activities, with an increasing number of individuals seeking to alleviate stress and release emotions through sports [1]. As societal enthusiasm for sports activities grows, the learning and optimization methods of sports skills have gained more attention. Traditionally, this process has been predominantly governed by the professional capabilities and personal experiences of coaches, considering as an individual-dependent method that lacks objectivity and is resource-intensive. Therefore, it is essential to explore how advanced artificial intelligence algorithms can be utilized to enhance the efficiency of the sports skill learning and optimization process. In recent years, deep neural networks have been widely applied in various fields, such as speech recognition, fault monitoring, and text analysis. Notably, in the field of image recognition, deep convolutional neural networks (DCNNs) [2] have demonstrated the ability to effectively process unstructured image inputs and uncover latent features within massive datasets, providing a novel approach for sports training.

Employing neural network algorithms to identify sports skills presents an intriguing research problem. Such methods leverage the powerful image recognition capabilities inherent in deep learning algorithms to analyze the types of movements performed by athletes, detect key points in the human body and postural information, and thereby aid athletes in enhancing their motor skills and improving the quality of their movements. Additionally, by extracting temporal information from continuous inputs, deep neural networks can effectively predict future movements, thereby preventing potential risks and avoiding injuries resulting from improper actions. Therefore, the accurate recognition of sports actions and estimation of human poses can not only enhance the efficiency and quality of motor skill learning but also provide sports enthusiasts with more effective training methods.

In sports training, accurately identifying and predicting sequences of athletic movements poses a significant challenge. This challenge arises from the inherent complexity of human posture, the diversity of athletic skills, and the uncertainty in the execution of movements. To address the aforementioned challenges and enhance the feature extraction capability possessed by neural networks, it is imperative to comprehensively capture the temporal-spatial relationships inherent in sports movements. To this end, this paper proposes a novel multitask framework to jointly recognize sports actions and estimate human poses based on the hybrid deep neural network that integrates the Involution operator and Convolution operator. This approach significantly enhances the model's ability to capture spatial information, surpassing the performance of traditional convolutional neural networks. A parallel spatial-temporal attention mechanism is further designed to operate in a decoupled manner and focused separately on temporal and spatial dimensions. It facilitates the neural network's ability to identify crucial movements and detect subtle variations across different frames. Finally, a sports pose prediction method is proposed based on the multiscale graph convolutional networks, thereby optimizing the effectiveness and practicability when applied to sports skill training.

*Corresponding Author

This work is supported by University-level scientific research project (Social science) in 2023: Research on the construction of National fitness public service system embedded in digital technology in Yangzhou (Project Number: 2023xjsk003)

II. Literature Review

This section will present existing works that are the most relevant methods related to our work, including human action recognition, human pose estimation, and human motion prediction.

A. Human Action Recognition Methods

Human action recognition is a vision task centered on humans, aiming to identify the classification results corresponding to the input action sequences. It has extensive applications in fields such as human-computer interaction, sports training, and smart security [3]. Human action recognition has evolved from the use of hand-crafted features to features automatically obtained via deep neural networks. Early research primarily focused on extracting shallow information from input images, such as angles, edges, or contours [4]. To effectively recognize the motion information contained in human actions, optical flow and Histograms of Oriented Gradients (HOG) are often adopted as part of the feature set. For instance, FarajiDavar et al. [5] utilized HOG3D features to describe tennis actions and explored feature re-weighting and feature translation methods based on these features. Calandre et al. [6] employed optical flow information to detect table tennis stroke actions, thereby identifying the most relevant frames in the input videos. However, the use of manual features and machine learning methods suffers from poor generalization, complex feature extraction processes, and reliance on shallow features that inadequately describe the action information reflected in the original inputs, especially in the domain of sports action recognition.

Currently, deep learning techniques, represented by deep neural networks, have become the predominant method for human action recognition. By establishing methods for human action recognition based on deep learning, it is possible to construct more efficient and comprehensive sports training systems, pushing the development of sports skills learning in a more intelligent direction. Simonyan et al. [7] proposed a two-stream convolutional network for action recognition, which utilizes both input RGB images and optical flow images to extract spatial and temporal features respectively, with the recognition results obtained through the fusion at the decision layer. Moreover, utilizing human posture information to enhance action recognition results is also considered an effective means. Nie et al. [8] proposed a hierarchical structure to capture the geometric and appearance variations in posture. Notably, the lateral connections between adjacent frames were considered to describe the action-specific information. Furthermore, Lin et al. [9] developed a Temporal Shift Network (TSN), which can switch feature channels along the temporal dimension to exchange temporal information between adjacent frames. This module can also be embedded as an independent structure into any deep convolutional network model and can significantly improve recognition performance while maintaining lower FLOPs.

On the other hand, as the primary data for action recognition often comprises video data, deep models based on 3D convolution have also received considerable attention. Cao et al. [10] proposed a dual-stream bilinear 3D-CNN model that utilizes selective convolutional layer activations to form discriminative descriptors for videos, ultimately achieving a recognition accuracy of 95.3% on the PENN Dataset. Additionally, Baradel et al. [11] introduced a novel attention mechanism known as Glimpse Clouds, which learns to focus on specific image patches in space and time, aggregating the patterns and softly assigning each feature. Overall, action recognition methods based on deep learning, allowing for fully automated feature extraction and action classification, avoiding the influence of subjective factors, and having high accuracy and generalizability, becoming the mainstream approach in action recognition tasks.

B. Human Pose Estimation Methods

Human pose estimation involves determining the location of body joints and the connections between various body parts in 2D/3D spaces. This field has been actively researched over the past few years, evolving from conceptual frameworks such as Pictorial Structures [12] to recent deep neural network-based approaches. An effective approach considers human pose estimation as a detection task, specifically by obtaining heatmaps at body joints based on detection scores. Newell et al. [13] proposed a novel CNN architecture that employs repeated bottom-up and top-down processing. The proposed Stacked Hourglass networks were evaluated on the FLIC and MPII benchmarks, demonstrating improvements in 2D pose estimation. Pishchulin et al. [14] introduced the DeepCut method, which initially detects regions potentially containing human key joints, followed by creating a connection graph encompassing all regions. However, detection-based methods do not directly provide coordinates of human keypoints and instead infer them indirectly by maximizing the posterior probability.

Regression-based approaches involve projecting input actions onto desired keypoint coordinates through nonlinear functions. Toshev et al. [12] were pioneers in proposing a method for human pose estimation based on DNNs and cascade regression, which avoids the need for explicitly designing feature representations or detectors for body parts. Cheng et al. [15] addressed the issue of scale variation in multi-person pose estimation by proposing a method that utilizes a high-resolution feature pyramid to learn scale-aware representations, thereby achieving more precise keypoint localization in multi-person pose estimation. The proposed method achieved an Average Precision (AP) of 67.6% in the CrowdPose test.

In recent years, a key research focus has been on calculating the spatial information of each keypoint based on their 2D coordinates to obtain the 3D position of human posture. With the release of more high-accuracy 3D data, it has become feasible to train 3D human pose estimation models using deep neural network algorithms. Chen et al. [16] innovatively decomposed the 3D pose estimation problem into a 2D estimation based on camera coordinates and a 2D-to-3D matching using a non-parametric shape model. Pavllo et al. [17] proposed a multi-view fusion 3D pose estimation algorithm based on 2D keypoint trajectories, utilizing 2D keypoints to estimate 3D poses and back-projecting to 2D space to enable semi-supervised training. The proposed method reduced the error by 11% compared to the previous state-of-the-art on the Human3.6M dataset.

C. Research Gaps

Although deep learning has achieved significant success in human-centered fields such as action recognition, its application in sports skill training still presents numerous challenges, particularly when dealing with rapid pose movements and the diversity of actions under the same sports skill. Based on these considerations, this paper aims to address the following key issues:

1) Limitations of traditional CNN models in sports skills training: The utilization of traditional Convolutional Neural Networks (CNNs) in the domain of sports skill training, particularly for sports action recognition and human pose estimation, has exposed specific deficiencies. Standard CNN architectures are characterized by fixed, limited receptive fields and inherent spatial invariance, which compromise their capacity to effectively model the contextual nuances of complex athletic movements and limit their sensitivity to variations in spatial configurations. Furthermore, the generalization of conventional CNNs is predominantly contingent upon the original training dataset, typically necessitating substantial retraining or fine-tuning to adapt these models for diverse sports training applications.

2) Lack of Multitask Framework for Sports Action Recognition and Pose Estimation: Contemporary studies in implementing action recognition and human pose estimation for sports skill training typically utilize independent operational frameworks. While this method permits the tailored algorithmic development specific to each task, it frequently neglects the potential synergistic interactions between these intimately connected tasks. Furthermore, both pose estimation and action recognition generally share analogous feature extraction phases and operating these models independently leads to repetitive processing steps, thereby diminishing computational efficiency and increasing the complexity of real-time applications. Employing a multitask framework to concurrently learn shared features from pose estimation and action recognition can facilitate the acquisition of more robust and extensible features, offering a more holistic comprehension of sports actions and markedly enhancing the support for the learning and optimization of sports skills.

3) Limitations of CNN models in sports pose prediction: Although CNN backbone networks are widely used in action recognition and pose estimation, the inherent non-Euclidean nature of human keypoints makes it challenging for CNN models to achieve satisfactory results in human pose prediction. Particularly when dealing with the relationships or constraints between human keypoints and body parts, models based on CNN backbones struggle to incorporate such information in a priori manner, which is crucial for accurate human pose prediction. Representing the human pose in the form of a graph allows for a more precise reflection of the structural and functional relationships between different body parts. Thus, constructing a backbone model based on GCNs (Graph Convolutional Networks) to model the spatiotemporal relationships of human poses better meets the requirements of motion pose prediction.

In summary, future studies should concentrate on creating innovative models and approaches that tackle the existing challenges in sports skill training, aiming to not only boost the practicability but also enhance the efficacy and accuracy of sports training.

III. Research on Recognition Methods of Sports Action and Human Pose based on Deep Neural Networks and Attention Mechanism

In this section, a novel multitask framework based on deep neural networks and attention mechanisms is proposed for sports action recognition and human pose estimation. Furthermore, a novel multiscale model is proposed for human pose prediction based on the estimated body keypoints. The preliminary knowledge of Hybrid Involution and Convolution Neural Networks (HICNN) is first introduced, followed by the proposed multitask framework and parallel spatial-temporal attention (PSTA). Finally, the multiscale Graph Convolutional Network (MS-GCN) is designed to predict human poses for sports skill training.

A. Hybrid Involution and Convolution Neural Network

As a primary component of deep neural networks, Convolutional Neural Networks (CNNs) utilize the spatial invariance and channel specificity of convolution kernels to enhance computational efficiency and the ability to interpret translation equivalency. However, these characteristics hinder the adaptability of convolution kernels to different spatial positions, and their limited receptive fields pose challenges in modeling long-distance relationships. To address these issues, the involution operator, which possesses symmetrically inverse inherent characteristics, has been proposed. Specifically, the involution operator shares weights across different channels while varying spatially, thereby compensating for the deficiencies of convolution kernels in capturing long-distance relationships [18].

Given the input feature map as $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, C represent its height, width, and channels. By applying multiply-add operations in a sliding-window manner, the output feature map can be expressed as

$$\boldsymbol{Y}_{i,j,k} = \sum_{c=1}^{C} \sum_{(u,v)\in\Delta_K} \Theta_{k,c,u+\lfloor K/2 \rfloor, v+\lfloor K/2 \rfloor} \boldsymbol{X}_{i+u,j+v,c} \tag{1}$$

where $\Theta \in \mathbb{R}^{C_0 \times C_i \times K \times K}$ represents the convolution filters with the fixed kernel size of $K \times K$, and $\Delta_K \in \mathbb{R}^2$ refers to the set of offsets in the neighborhood considering convolution conducted on the center pixel, written as

$$\Delta_K = [-\lfloor K/2 \rfloor, \cdots, \lfloor K/2 \rfloor] \times [-\lfloor K \times 2 \rfloor, \cdots \lfloor K/2 \rfloor] \tag{2}$$

Compared to the standard convolution kernel, involution kernels are devised with the inverse characteristics in the spatial and channel dimension, expressed as $\Psi \in \mathbb{R}^{H \times W \times K \times K \times G}$. Specifically, an involution kernel is tailored for the pixel $\boldsymbol{X}_{i,j} \in \mathbb{R}^{C}$ but shared over different channels. The output feature map of involution can be obtained by applying multiply-add operations with involution kernels, that is

$$\boldsymbol{Y}_{i,j,k} = \sum_{(u,v)\in\Delta_K} \Psi_{i,j,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor,\lceil KG/C\rceil} \boldsymbol{X}_{i+u,j+v,k} \tag{3}$$

To fully leverage the capabilities of convolution and involution, this section explores the alternating stacking of these two types of operators, constructing the Hybrid Incolution and Convolution Neural Networks (HICNN). This hybrid architec- -ture serves as the backbone extractor aiming to enhance the model's ability to discern complex spatial relationships while maintaining computational efficiency.

B. HICNN-PSTA for Sports Action Recognition and Pose Estimation

In this section, we introduced the HICNN-PSTA (Hybrid Involution and Convolution Neural Network with Parallel Spatial-Temporal Attention (PSTA), which is specially designed for joint sports action recognition and human pose estimation, as shown in Fig. 1.

Different from the previous work, this section attempts to establish a multitasking framework by predicting human poses and recognizing sports actions in parallel. The input RGB frames are first fed into the HICNN model to extract low-level visual features. Besides, a novel two-pathway attention mechanism, namely PSTA, is proposed to model spatial and temporal information in parallel. The PSTA mechanism significantly enhances the processing capabilities for both single-frame image and image sequences, which is crucial for multitask frameworks. Specifically, spatial attention aids in focusing on critical human-related information within individual frames, thereby increasing the accuracy of human pose estimation. Temporal attention, on the other hand, concentrates on the continuity of actions, which is essential for understanding action sequences and patterns.

The proposed PSTA is illuminated in Fig. 2 that originates from the vanilla multi-head self attention module, defined as

$$MSA(Q,K,V) = \mathrm{softmax}(\frac{Q \cdot K^T}{\sqrt{C}}) \cdot V \tag{4}$$

where $Q,K,V \in \mathbb{R}^{N\times C}$ denote the queies, keys and values. In PSTA module, the input embedding $E \in \mathbb{R}^{T\times N\times C}$ are firstly mapped into queries, keys and values with the same dimensions. Then, the mapped tensors are evenly divided into two groups along the channel dimension, results in time group $\{Q_T, K_T, V_T\}$ and space group $\{Q_S, K_S, V_S\}$. To model the spatial-temporal dependencies between joints avoiding the quadratic computation, the temporal and spatial correlations are calculated in two separate self-attention modules, which can be expressed as,

$$\begin{aligned} H_T &= MSA(Q_T, K_T, V_T) \\ H_S &= MSA(Q_S, K_S, V_S) \\ \boldsymbol{H} &= cat(H_T, H_S) \end{aligned} \tag{5}$$

Based on the latent representation, the multi-task prediction block produces single frame features, multi-task features, and image sequence features, which is defined as $\chi_t \in \mathbb{R}^{H_f\times W_f\times N_f}, \delta_t \in \mathbb{R}^{H_f\times W_f\times N_f}, v_t \in \mathbb{R}^{T\times N_j\times N_v}$. For pose estimation, prediction blocks take as input the multi-task features to predict body joint probability maps, expressed as

$$h_t = \Phi(W_h * \delta_t), h_t \in \mathbb{R}^{H_f\times W_f\times N_j} \tag{6}$$

The elastic net loss between the predicted human poses and ground-truth values is adopted for model training, which is defined as

$$L_p = \frac{1}{N_j}\sum_{j=1}^{N_j}(\|\hat{p}^j - p^j\|_1 + \|\hat{p}^j - p^j\|_2^2) \tag{7}$$

Furthermore, the multi-task features are multiplied by probability maps h_t at channel dimension to obtain the appearance features $V \in \mathbb{R}^{T\times N_j\times N_f}$ that describe the entire image sequences, thus recognizing sports actions by categorical cross-entropy loss on predicted actions.

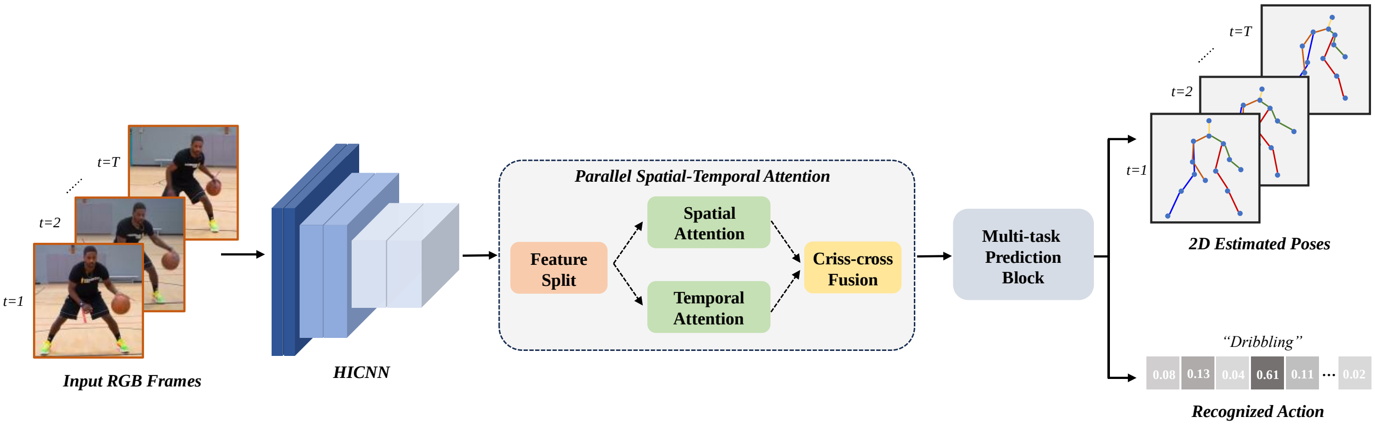


Fig. 1. Network structure diagram of sports action recognition and pose estimation based on HICNN-PSTA.

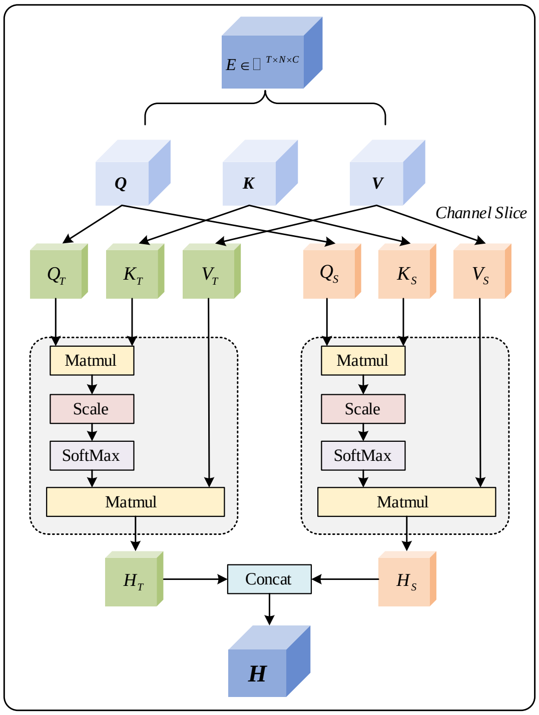


Fig. 2. An overview of the proposed PSTA.

C. MS-GCN for Sports Pose Prediction

Human pose prediction is of paramount importance in the field of sports skills training, as it assists participants in optimizing their skills and avoiding potential hazards. The key to effective prediction of human poses lies in a comprehensive understanding of the intrinsic correlations among sequences of human keypoints. To address these issues, this section introduces a Multi-Scale Graph Convolution Network (MS-GCN) that predicts future human poses based on the estimated sequences of human keypoints. MS-GCN extends conventional keypoints analysis by integrating single-scale graph and multi-scale graphs at various levels to connect body components. The single-scale graph provides a multi-granularity representation of the body skeleton, while the multi-scale graph, initialized by predefined physical connections, reflects the interconnections between different single-scale graphs and adjusts to poses sensitivity during training.

Suppose the estimated 2D skeleton-based poses are $\boldsymbol{P}_{-T_h:0} = [P_{-T_h}, \ldots, P_0] \in \mathbb{R}^{M \times (T_h+1) \times 2}$ and the future poses are $\boldsymbol{P}_{1:T_f} = [P_1, \ldots, P_{T_f}] \in \mathbb{R}^{M \times T_f \times 2}$. The goal of pose prediction is to generate future poses by the past observed ones, which can be expressed as $\hat{\boldsymbol{P}}_{1:T_f} = M_{pred}(\boldsymbol{P}_{-T_H:0})$. To construct the MS-GCN, two body scales are first initialized, expressed as a trainable adjacency matris $A_s \in \mathbb{R}^{M_s \times M_s}$ at scale *s*. Based on the single-scale graph, the GCN block extract spatial features of body components as well as temporal features from poses sequences, defined as,

$$\boldsymbol{P}_{s,sp} = \mathrm{ReLU}(A_s P_s W_s + P_s U_s) \tag{8}$$

where, W_s, U_s are trainable parameters. To enable information exchange across scales, a cross-scake fusion block is adopted to convert features from one scale to another. The cross-scale graph is a bipartite graph that corresponds the nodes in one single-scale graph to the nodes in another graph. Assuming the cross-scale graph with adjacent matrix as $A_{s_1 s_2}$, and the vectorized features of human joint and part are defined as $v_{s_1,i} = vec(conv_{s_1,\tau}((P_{s_1})_{:,i,:}; \mu)), v_{s_2,k} = vec(conv_{s_2,\tau}((P_{s_2})_{:,k,:}; \mu))$ to leverage temporal information, where τ, μ represent the temporal convolution kernel size and stride. Then, the edge weight between the joint and part can be inferred as

$$\begin{aligned} r_{s_1,i} &= \sum_{j=1}^{M_{s_1}} f_{s_1}(v_{s_1,i}, v_{s_1,j} - v_{s_1,i}) \\ h_{s_1,i} &= g_{s_1}([v_{s_1,i}, r_{s_1,i}]) \\ r_{s_2,k} &= \sum_{j=1}^{M_{s_2}} f_{s_2}(v_{s_2,k}, v_{s_2,j} - v_{s_2,k}) \\ h_{s_2,k} &= g_{s_2}([v_{s_2,k}, r_{s_2,k}]) \\ (A_{s_1 s_2})_{k,i} &= \mathrm{softmax}(h_{s_2,k}^T h_{s_1,i}) \in [0,1] \end{aligned} \tag{9}$$

where, the $f(\cdot), g(\cdot)$ denotes multi-layer perceptrons. Given the joint features at a certain time stamp, the part-scale feature can be updated by the edge weight.

To accurately predict the future human poses, the MS-GCN adopted encoder-decoder architecture, where a graph-based GRU is utilized to learn and update hidden states with the guide of a graph. Let $A_H \in \mathbb{R}^{M \times M}$ be the adjacent matrix of the inbuilt graph, which is initialized with the skeleton-graph, and $H^0 \in \mathbb{R}^{M \times D_h}$ be the initial state of GRU. The processing procedures of GRU are defined as

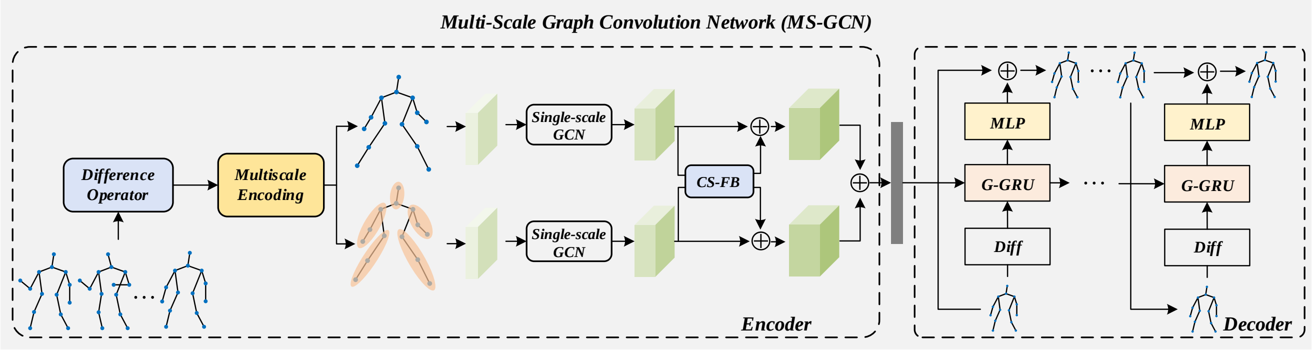


Fig. 3. Network structure diagram of pose prediction based on MS-GCN.

$$\begin{aligned} r^{(t)} &= \sigma(r_{in}(I^{(t)}) + r_{hid}(A_H H^{(t)} W_H)) \\ u^{(t)} &= \sigma(u_{in}(I^{(t)}) + u_{hid}(A_H H^{(t)} W_H)) \\ c^{(t)} &= \tanh(c_{in}(I^{(t)}) + r^{(t)} \odot c_{hid}(A_H H^{(t)} W_H)) \\ H^{(t+1)} &= u^{(t)} \odot H^{(t)} + (1-u^{(t)}) \odot c^{(t)} \end{aligned} \quad (10)$$

At the end, the future pose is predicted by the decoder as

$$\hat{P}^{(t+1)} = \hat{P}^{(t)} + f_{pred}(\mathrm{GRU}(diff(\hat{P}^{(t)}), H^{(t)})) \quad (11)$$

where *diff* represents difference operator that calculate velocity and acceleration of human body keypoints. The entire structure diagram is shown in Fig. 3, which is trained end-to-end by the loss function as

$$L_{pred} = \frac{1}{N}\sum_{n=1}^{N} \left\| (\boldsymbol{P}_{1:T_f})_n - (\hat{\boldsymbol{P}}_{1:T_f})_n \right\|_1 \quad (12)$$

IV. Case Verification

This section will validate the effectiveness of the proposed method based on a self-made experimental dataset.

A. Dataset Preparation and Experimental Environment

To advance the application of action recognition and human pose understanding in sports skills training, this paper has developed a basketball motion dataset by collecting internet data and filming original content. This dataset comprises 400 RGB videos of various basketball actions such as dribbling, passing, shooting, and dunking, performed by different individuals in diverse settings. From this collection, approximately 6,000 images were meticulously selected and manually annotated with human keypoints for 2D pose estimation and prediction. The dataset is divided into training, validation, and testing sets in a ratio of $[6:2:2]$. Effective data augmentation techniques, including random flipping and the addition of random noise, have been applied. This dataset served as the basis for validating the proposed HICNN-PSTA and MS-GCN models. Detailed information on the hardware and software used in the experiments is provided in Table I.

TABLE I. Experimental Software and Hardware Environment Table

CPU	*Intel(R) Core(TM) i5-13400F*
GPU	*NVIDIA GeForce RTX 4070*
Operating System	*Ubuntu 18.04*
CUDA	*11.1*
Programming	*Pytorch1.10.0, Python 3.8*

To fully demonstrate the effectiveness of the proposed methods, this study conducted a series of comparative experiments to comprehensively evaluate the performance of the proposed algorithms in sports action recognition and human pose estimation and prediction. Given that the initial part of this research utilized a multi-task framework, we selected appropriate models for comparison based on the specific problems addressed. For sports action recognition, the AGC-LSTM [19] model was chosen as the benchmark, whereas the HPRNet [20] model was used as the comparative standard in the domain of 2D human pose estimation. The specific content of the experiments is as follows: the action recognition accuracy for different basketball skills, the results of 2D human keypoint estimation and pose prediction for various basketball skills, and ablation studies conducted on the proposed algorithm. During the training process, an Adam optimizer with an initial learning rate of 0.001 was employed, accompanied by a linear learning rate decay coefficient set at 0.95. The training batch size was configured to 64, and the number of epochs for iteration was set to 50.

B. Experimental Results

To fully demonstrate the effectiveness of the proposed algorithm, this section first explores the performance of the proposed HICNN-PSTA in basketball skill action recognition, building on the experimental setup described above. The model was trained using a supervised learning approach on the dataset constructed for this study, and the classification cross-entropy error variation curve during the training process is shown in Fig. 4. It is evident that the error rapidly decreased and stabilized shortly after training commenced, reflecting the model's capability to effectively adjust model weights using error gradients and ultimately achieve convergence. Additionally, the proposed HICNN-PSTA was also subjected to quantitative experiments, as shown in Table II, which includes the recognition accuracies of various models for different basketball skill actions. The proposed algorithm achieved the highest recognition accuracy across all skills, likely benefiting from the PSTA module's superior ability to capture temporal-spatial information, particularly crucial for the rapid and complex movements characteristic of basketball and other sports skills.

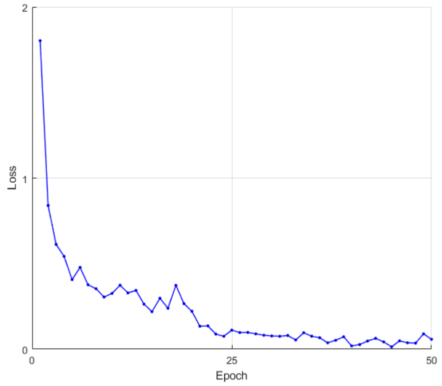


Fig. 4. Training loss rate curve.

TABLE II. Quantitative Comparison of Sports Action Recognition

Method	**Accuracy**			
	Dribbling	*Shooting*	*Passing*	*Dunk*
AGC-LSTM	94.53	90.24	92.50	85.02
Proposed	**96.11**	**91.15**	**93.87**	**91.52**

Furthermore, to delve deeper into the performance of the proposed method in sports action recognition, we conducted several ablation experiments, the results of which are presented in Table III. This experiment compared the proposed model with two variants: one substituting the backbone network with ResNet, referred to as ResNet-PSTA, and another omitting the PSTA module, referred to simply as HICNN. The assessment criterion was the F1-score on the test set, and performance across four types of sports skills was evaluated. The results indicate that the proposed HICNN-PSTA model achieved the best performance in recognizing different sports skills. This demonstrates the superior capability of the proposed modules in capturing latent movement information and modeling spatio-temporal relationships, which are crucial for sports action recognition. The effectiveness of the HICNN-PSTA model underscores its significant role in accurately recognizing complex sports actions.

TABLE III. ABLATION STUDY OF SPORTS ACTION RECOGNITION

Method	F1-score			
	Dribbling	*Shooting*	*Passing*	*Dunk*
ResNet-PSTA	0.81	0.79	0.70	0.68
HICNN	0.76	0.73	0.61	0.71
Proposed	**0.88**	**0.81**	**0.76**	**0.80**

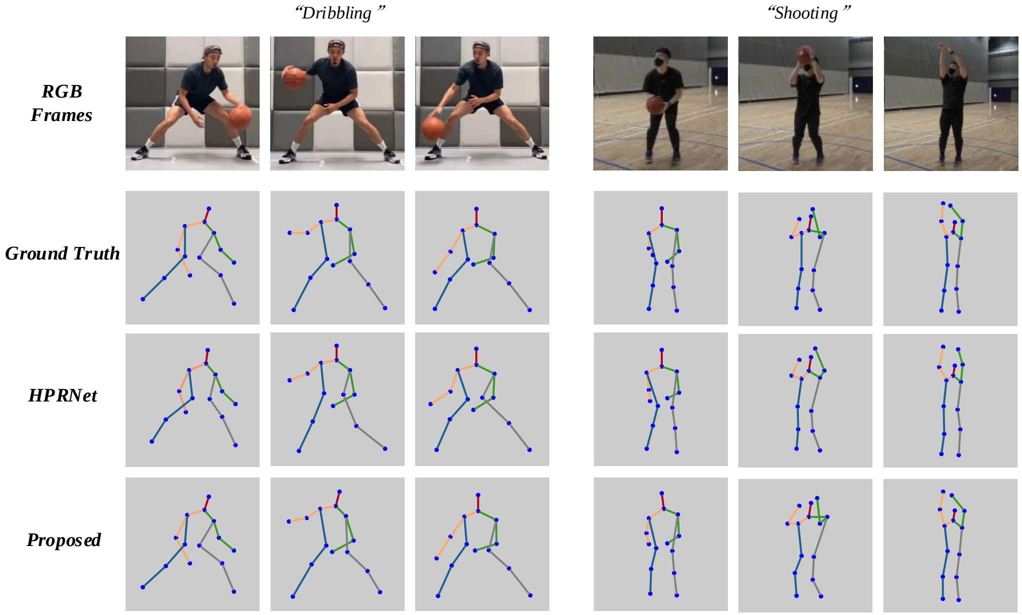


Fig. 5. Qualitative comparison of sports pose estimation based on the HICNN-PSTA.

Additionally, we further explored the performance of HICNN-PSTA in sports pose estimation, as shown in Table IV and Fig. 5. Table IV employs the Percentage of Correct Keypoint Percentage (PCK) as the evaluation metric under thresholds of [0.2, 0.1, 0.05], demonstrating the model's performance in dribbling pose estimation. It is observable that HICNN-PSTA achieved the best recognition results across all thresholds, reflecting the model's capability to utilize effective information from action recognition to enhance the accuracy of human pose estimation in a multi-task framework. The results of human pose estimation for dribbling and shooting are illustrated in Fig. 5.

In addition, we conducted multiple experiments to evaluate the performance of the proposed MS-GCN in human pose prediction and compared it with two widely-used models, TP-RNN [21] and Traj-GCN [22], as shown in Table V. This table employs the Mean Angle Error (MAE) as the evaluation metric, detailing the prediction results for different sports skills over various time intervals. It is evident that, compared to the other two algorithms, MS-GCN achieved the best prediction results in the pose prediction for shooting and dunking across different time intervals. Although Traj-GCN outperforms MS-GCN in predicting these two skills, overall, MS-GCN still demonstrates substantial potential and practicality in predicting sports skills.

TABLE IV. QUANTITATIVE COMPARISON OF SPORTS POSE ESTIMATION

Method	PCK@0.2	PCK@0.1	PCK@0.05
HPRNet	82.11	75.8	70.01
Proposed	**90.08**	**81.47**	**79.34**

TABLE V. QUANTITATIVE COMPARISON OF SPORTS POSE PREDICTION

Sports Skills	ms	TP-RNN	Traj-GCN	Proposed
Dribbling	80	0.34	**0.32**	0.33
	160	0.61	0.50	**0.42**
	320	1.25	1.19	**0.88**
Shooting	80	0.56	0.45	**0.41**
	160	1.48	0.86	**0.78**
	320	1.97	1.28	**1.01**
Passing	80	0.66	0.59	**0.47**
	160	1.01	1.13	**0.92**
	320	1.68	**1.45**	1.47
Dunk	80	0.30	0.38	**0.28**
	160	0.75	0.49	**0.50**
	320	1.32	1.06	**0.92**

V. CONCLUSION

This study aims to explore how to better utilize neural network models to optimize sports skill training, with a focus on achieving sports action recognition and the estimation and prediction of athletes' poses, thereby advancing the application of neural networks and other artificial intelligence algorithms in the field of sports training. To address these challenges, we first propose a multi-task framework-based HICNN-PSTA model. This model enhances the feature extraction capabilities of the conventional CNN by integrating the Involution operator into the backbone network. Additionally, this study constructs a PSPA module based on the attention mechanism to fully capture the latent spatio-temporal information of sports actions, thereby improving the efficiency of the algorithm with the help of the multi-task framework. Furthermore, to accurately predict future poses of athletes and provide training recommendations, this paper introduces an MS-GCN model based on a multi-scale graph. This algorithm considers the constraints between human body keypoints and segments, significantly enhancing the capability to model the complex sports skills. Detailed experiments validate that the proposed algorithms can effectively recognize sports actions and also demonstrate excellent performance in human pose estimation and prediction. In the future, we plan to integrate more advanced neural network algorithms to address the generalization deficiencies across different sports, thereby further optimizing sports skill training.

REFERENCES

[1] P. Wang, “Research on Sports Training Action Recognition Based on Deep Learning,” Scientific Programming, vol. 2021, pp. 1–8, Jun. 2021, doi: 10.1155/2021/3396878.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[3] K. Host and M. Ivašić-Kos, “An overview of Human Action Recognition in sports based on Computer Vision,” Heliyon, vol. 8, no. 6, p. e09633, Jun. 2022, doi: 10.1016/j.heliyon.2022.e09633.

[4] K. Soomro and A. R. Zamir, “Action Recognition in Realistic Sports Videos,” in Computer Vision in Sports, T. B. Moeslund, G. Thomas, and A. Hilton, Eds., in Advances in Computer Vision and Pattern Recognition. , Cham: Springer International Publishing, 2014, pp. 181–208. doi: 10.1007/978-3-319-09396-3_9.

[5] N. FarajiDavar, T. De Campos, J. Kittler, and F. Yan, “Transductive transfer learning for action recognition in tennis games,” in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain: IEEE, Nov. 2011, pp. 1548–1553. doi: 10.1109/ICCVW.2011.6130434.

[6] J. Calandre, R. Péteri, and L. Mascarilla, “Optical Flow Singularities forSports Video Annotation: Detection of Strokes in Table Tennis”.

[7] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos”.

[8] B. X. Nie, C. Xiong, and S.-C. Zhu, “Joint action recognition and pose estimation from video,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 1293–1301. doi: 10.1109/CVPR.2015.7298734.

[9] J. Lin, C. Gan, and S. Han, “TSM: Temporal Shift Module for Efficient Video Understanding,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 7082–7092. doi: 10.1109/ICCV.2019.00718.

[10] C. Cao, Y. Zhang, C. Zhang, and H. Lu, “Body Joint Guided 3-D Deep Convolutional Descriptors for Action Recognition,” IEEE Trans. Cybern., vol. 48, no. 3, pp. 1095–1108, Mar. 2018, doi: 10.1109/TCYB.2017.2756840.

[11] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, “Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 469–478. doi: 10.1109/CVPR.2018.00056.

[12] A. Toshev and C. Szegedy, “DeepPose: Human Pose Estimation via Deep Neural Networks,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE, Jun. 2014, pp. 1653–1660. doi: 10.1109/CVPR.2014.214.

[13] A. Newell, K. Yang, and J. Deng, “Stacked Hourglass Networks for Human Pose Estimation,” in Computer Vision – ECCV 2016, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9912. , Cham: Springer International Publishing, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.

[14] L. Pishchulin et al., “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 4929–4937. doi: 10.1109/CVPR.2016.533.

[15] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, “HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 5385–5394. doi: 10.1109/CVPR42600.2020.00543.

[16] C.-H. Chen and D. Ramanan, “3D Human Pose Estimation = 2D Pose Estimation + Matching,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 5759–5767. doi: 10.1109/CVPR.2017.610.

[17] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7745–7754. doi: 10.1109/CVPR.2019.00794.

[18] D. Li et al., “Involution: Inverting the Inherence of Convolution for Visual Recognition,” Apr. 11, 2021, arXiv: arXiv:2103.06255. Accessed: Aug. 21, 2024. [Online]. Available: http://arxiv.org/abs/2103.06255

[19] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1227–1236. doi: 10.1109/CVPR.2019.00132.

[20] N. Samet and E. Akbas, “HPRNet: Hierarchical point regression for whole-body human pose estimation,” Image and Vision Computing, vol. 115, p. 104285, Nov. 2021, doi: 10.1016/j.imavis.2021.104285.

[21] H.-K. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, “Action-Agnostic Human Pose Forecasting,” in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA: IEEE, Jan. 2019, pp. 1423–1432. doi: 10.1109/WACV.2019.00156.

[22] W. Mao, M. Liu, M. Salzmann, and H. Li, “Learning Trajectory Dependencies for Human Motion Prediction,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 9488–9496. doi: 10.1109/ICCV.2019.00958.

Birth Certificates Delivery, Traceability and Authentication Using Blockchain Technology

Tankou Tsomo Maurice Eddy*, Bell Bitjoka Georges, Ngohe Ekam Paul Salomon, Ekani Mebenga Vianney Boniface
Laboratory of Electrical Engineering Mechatronics and Signal Processing ENSPY, University of Yaoundé, Cameroon

***Abstract*—Now-a-days, the vast majority of birth certificate registration systems are paper-based and managed independently by administrative communities. This means that birth information only exists at the place where the birth is registered, which facilitates the counterfeiting or falsification of such identity documents. Therefore, the implementation of a system for the issuance, traceability, and authentication of birth certificates is imperative. Blockchain, characterized by transparency, immutability, protection, privacy, and autonomy, makes this technology the ideal solution for implementing a birth certificate registration, traceability, and authentication system. This article presents a decentralized system for the registration, traceability, and authentication of birth certificates based on Hyperledger Fabric private blockchain deployed in a Virtual Private Network - Multi-Protocol Label Switching (VPN-MPLS) network. This birth certificate is characterized on one hand by the attributes of its owner and on the other hand by a Quick Response (QR) code containing the digital signature of its signer and the unique identifier of the birth certificate. Within the network, the unique identifier of the generated document is hashed and stored using the Secure Hash Algorithm-256 (SHA-256) hash function to optimize storage space and enhance security. Furthermore, the proposed platform includes an application designed using Docker Compose, Apache CouchDB, NodeJS, Go, and Hyperledger Explorer. The designed model is a birth certificate registration platform that ensures enhanced security and transparency.**

Keywords—Birth certificates; blockchain; security; traceability; authentication; counterfeiting; falsification; hyperledger fabric

I. Introduction

For many decades, centralized architectures have been used in the deployment of network infrastructures. These architectures are characterized by a main server responsible for establishing a direct trust relationship among all network participants [1]. However, this type of architecture has limitations in that the failure of the central server leads to the paralysis of the entire network, and the malfunction of a node causes a break in the communication chain without prior notification to the system's participants [2, 3].

Distributed networks, on the other hand, do not require a single trusted authority. Each node in the network acts as both a client and a server, dynamically discovering and connecting with each other while relaying requests from terminal to terminal. This architecture is robust but requires significant traffic, as searching for a file takes more time. Each request is sent to all connected users, who do the same, which can lead to a long wait for a response to a request if thousands of users are connected [4]. This horizontal chain of trust model is the foundation of blockchain technology. Blockchain is a storage and transmission technology that operates transparently without a central control body [5, 6].

Document storage techniques are generally carried out within centralized systems. When a modification, deletion, or update is made in this centralized database, it affects the entire system. However, when it comes to a fraudulent act, the entire network is compromised [7, 8]. Considering the advantages of the decentralized model, where any modification, deletion, or update of information requires the approval of at least 51% of the network nodes.

In this article, a decentralized platform based on the Hyperledger Fabric private blockchain for the production, traceability, and authentication of birth certificates is proposed. The smart contract generates a Unique Identification Number (UIN) that identifies the birth certificate. This UIN, along with the identifiers and the digital signature of the civil registrar, is contained in a QR code that is affixed to the birth certificate. To optimize storage space and ensure security, this UIN is hashed using the SHA-256 hashing function. To verify the authenticity of a birth certificate, the hash of the UIN scanned via the QR code is compared with the one stored in the blockchain database; if they match, authenticity is guaranteed; otherwise, it is compromised.

The remainder of the document is organized as follows: Section II focuses on general knowledge. In Section III, the related work is presented. Section IV deals with the research method. Section V presents the results and discussions. Finally, conclusions and future work are addressed.

II. Background

This section presents general knowledge on the main concepts addressed in this article. The primary objective of this section is to facilitate the understanding of the key concepts used. The following outlines represent the subsections:

A. Blockchain

1) Definition of blockchain: A blockchain is a secure and decentralized distributed ledger technology. It is best known for its role in cryptocurrency systems to maintain a secure and decentralized record of transactions, but it also has many other potential applications. It is a chain of blocks, where each block contains a timestamp, transaction data, and a cryptographic hash of the previous block. This hash is a unique digital fingerprint that links the blocks together and makes it very difficult to tamper with the data. When a new transaction is made, it is broadcast to the network of computers that manage the blockchain. These computers verify the transaction and add it to a new block. Once the new block is created, it is added to the end of the chain. This process is continuously repeated, creating an ever-growing chain of blocks that is constantly verified and updated. Thus, the blockchain is a very secure and tamper-proof record of transactions [9-13].

2) Blockchain characteristics: Blockchain technology has several key features that make it unique and valuable for various applications. [14]:

a) Immutability: Once data is recorded on a blockchain, it cannot be altered or deleted. This ensures a permanent and tamper-proof record of transactions.

b) Decentralization: A blockchain operates on a decentralized network of nodes, eliminating the need for a central authority. This enhances security and reduces the risk of single points of failure.

c) Transparency: All participants in the network have access to the same data, promoting transparency and trust. Every transaction is visible to all nodes, ensuring accountability.

d) Security: Blockchain employs cryptographic techniques to secure data. Each block is linked to the previous one by a cryptographic hash, making it extremely difficult to modify data without detection.

e) Consensus mechanisms: Blockchains rely on consensus algorithms, such as Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), and Proof of Authority (PoA), to validate transactions and maintain the integrity of the ledger.

f) Distributed ledger: The ledger is distributed across all nodes in the network, ensuring that all participants have an up-to-date copy of the data. This distribution enhances reliability and reduces the risk of data loss.

g) Smart contracts: These are self-executing contracts with the terms of the agreement directly written into lines of code. They automatically execute the agreement when predetermined conditions are met, reducing the need for intermediaries.

3) Blockchain Applications: Blockchain technology has a wide range of applications across various industries. Here are a few examples:

a) Finance: Blockchain enhances the security, transparency, and efficiency of financial transactions. It is used for cross-border payments, smart contracts, and decentralized finance (DeFi) applications [15].

b) Healthcare: Blockchain can securely store patient records, ensuring data integrity and confidentiality. It also facilitates the sharing of medical data among various healthcare providers [16, 17].

c) Supply chain management: Blockchain ensures the transparency and traceability of supply chains, helping to track the origin and journey of products, which is crucial for quality control and fraud prevention [18, 19].

d) Real estate: Blockchain can streamline real estate transactions by reducing paperwork and enabling secure, transparent, and tamper-proof records of property ownership and history [20].

e) Voting systems: Blockchain can be used to create secure and transparent voting systems, reducing the risk of fraud and ensuring the integrity of election results [21].

f) Digital identity: Blockchain can provide secure and verifiable digital identities that can be used for various purposes, including online authentication and Know Your Customer (KYC) processes [22].

g) Intellectual property: Blockchain can protect intellectual property rights by providing a secure and immutable record of ownership and creation [23].

h) Internet of Things (IoT): Blockchain can improve the security and efficiency of IoT [24].

i) Blockchain technology offers numerous advantages across various sectors. Some of the key benefits include:

j) Enhanced security: Blockchain uses advanced cryptographic techniques to secure data. Each transaction is encrypted and linked to the previous one, making it extremely difficult for unauthorized parties to modify data [25, 26].

k) Greater transparency: As blockchain operates on a decentralized network, all participants have access to the same data. This transparency ensures that all transactions are visible and verifiable by all network members.

l) Improved traceability: Blockchain creates an immutable record of transactions, which is particularly useful in supply chains. It allows for tracking products from their origin to their final destination, reducing the risk of fraud and ensuring authenticity.

m)Increased efficiency and speed: By eliminating intermediaries and automating processes through smart contracts, blockchain can significantly speed up transactions and reduce the time required for various operations.

n) Reduced costs: Blockchain can reduce costs by eliminating the need for third-party intermediaries and reducing the amount of paperwork and administrative tasks required.

o) Decentralization: The decentralized nature of blockchain means there is no single point of failure. This enhances the system's resilience and reduces the risk of data loss or corruption.

p) Improved privacy: Blockchain can enhance privacy by allowing users to control their account data.

4) Blockchain typology: There are three main types of blockchain:

a) Public blockchain: A public blockchain allows transactions to be recorded and validated by the entire network. This type of blockchain can be compared to a tamper-proof ledger maintained by all its participants. A public blockchain is a decentralized network that operates on a peer-to-peer basis. Peer-to-peer means exchanging between two actors without an intermediary through a relationship of trust.

b) Private blockchain: In contrast to a public blockchain, a blockchain is considered private if the consensus principle is verified by a limited and predefined number of participants. The ability to participate in transactions is defined by an organization, as is the verification work [27].

c) Consortium blockchain: A consortium blockchain brings together several private actors who have an interest in working together. Decisions (block validations) are made by the majority of the most important members and not by the entire network as in a public blockchain. Only the decision-makers can verify the validity of the blocks [28].

5) Hyperledger fabric: Hyperledger Fabric is an open-source private blockchain platform developed by the Linux Foundation and IBM. Unlike Bitcoin and Ethereum, it does not require a virtual currency, the consensus is open, and smart contracts can be written in multiple programming languages including Go, Java, Python, etc., with tokenization through smart contracts [29]. The particularity of this technology lies mainly in the fact that, depending on the information, transactions can be viewed by everyone (public) or restricted to a group of organizations (confidential). It allows for the deployment of blockchain applications.

A Hyperledger Fabric blockchain network is composed of several nodes or peers, which host the blockchain and execute smart contracts, or chaincode. These smart contracts are executed by the peers of the network specific to a group of organizations and are inaccessible to members who are not part of that organization. Thanks to channels in Fabric, all smart contracts and data are only accessible to members who are part of the channel.

In addition to the network peers, a Fabric network can include an ordering service/ordered that performs the total ordering of transactions accepted by the Fabric network.

As for smart contracts, they are executed in a Docker container, in order to isolate them from the Fabric code and other smart contracts running on the same machine. Each smart contract has a persistent state called a key-value store [30].

Smart contracts manipulate key-value pairs using the put and get methods, which allow reading from the key-value store. These key-values are stored internally in a Level DB database on the same node, and CouchDB is used for the implementation of the database. This database allows storing key-value pairs [31-34].

In Hyperledger Fabric, a transaction goes through the following steps:

- Client proposes the transaction: This transaction is a request to invoke a smart contract. The request is signed by the client and sent to the channel where the chaincode is deployed. The number of endorsements it expects to receive is in accordance with the chaincodes endorsement policy.
- Endorsing peers verify the signature and execute the transaction: Peers verify the authenticity, form, replay protection, and client authorization.
- The client collects endorsements and sends them to the ordering service: The client examines and compares all endorsements and verifies that they meet the criteria defined in the smart contract. If the request is a read-only type, no request is sent to the ordering service. If the request is a smart contract invocation or write request, the endorsements are gathered into a transaction and sent to the ordering service which, in turn, includes it in the blockchain. The transaction is validated and recorded: The transactions ordered in the blocks are transmitted to all peers on the organization's channel via the ordering service. Peers perform a verification of the transaction and apply the endorsement policy by the endorsing peers. If all verifications are successful, the endorsing peers add the block to the distributed ledger.

6) Birth certificate: A birth certificate is a legal document that proves a person's civil status. Indeed, a birth certificate contains the name, first name(s), date, time, and place of birth, etc. A birth certificate is used to prove one's identity and family status.

However, the procedure for obtaining one is governed by a law that varies from country to country. For many years, birth certificates were produced manually by the responsible parties. The complexity, the burden, and the falsification or fraud in the establishment process have motivated the migration towards the digitalization of birth certificates. The completed birth certificate form is then recorded in a central or non-central database. One of the limitations of the vertical trust chain is that the central node can be compromised. Moreover, the administrations in charge of certification do not have a means of authenticating certificates produced at the time of their certification.

III. Literature Review

The secure production of birth certificates is paramount for both individuals and the organizations that interact with them. There is a risk of fraudulent activities within these databases, and the authorities responsible for authenticating birth certificates may not be able to verify in real-time the identity of the person who produced a given document. Falsification can involve altering the date of birth to appear younger or older, changing the names of the parents, the place of birth, etc. Table I refers to the most recent literature review.

In [35], the authors study the cumbersome procedures caused by the manual registration of births and deaths, and they propose a decentralized, simplified, and transparent application based on the Ethereum blockchain, guaranteeing immutability and providing the true and irrefutable origin of records. However, a problem arises with the consensus algorithm due to the slowness of the transaction verification system, compromising availability, and high energy consumption, leading to economic and environmental issues. Furthermore, less developed chains are highly susceptible to 51% attacks.

In [36], the authors propose a traceable online will system based on the Ethereum blockchain and smart contract technology to address the issues of complexity, falsification, slowness, and high cost in the establishment of wills. To this end, they propose a traceable online will system integrating an arbitration strategy in case of disputes based on blockchain technology. The main problems lie in the flaw in the consensus algorithm used, which is proof of work, and the solution is energy-intensive.

TABLE I. SYSTEMATIC LITERATURE REVIEW

Authors (Year)	Model+Applications	Advantages	Disadvantages
Shah et al., (2020) [35]	Ethereum-based, simplified and transparent application for birth and death registration	Accelerated registration processes	Slow system, susceptible to 51% attacks, energy-intensive
Chen et al., (2021)[36]	Ethereum-based online will creation and storage system	Traceability of wills, accelerated processes, simplified procedures, reduced costs	System complexity, slow system, susceptible to 51% attacks, energy-intensive
Okoth et al.,(2023) [37]	Centralized architecture-based security policy for civil registration systems	Data confidentiality and access control	Failure of the central node paralyzes the entire system
Danquah et al., (2022) [38]	Ethereum-based platform for birth registration	Accelerated birth registration process	Lack of experimental data
Sharma et al., (2020) [39]	Ethereum and blockcerts-based application for issuing birth certificates	Simplified birth certificate acquisition process	Monetization of generated transactions, system complexity, slow system, susceptible to 51% attacks, energy-intensive
Bennett et al., (2022) [40]	Ethereum and Corda-based application for birth registration and certificate issuance	Simplified birth registration and certificate issuance process	Incompatibility with existing systems

The author in [37] presents the security vulnerabilities faced by civil registration systems, compromising the integrity and confidentiality of information. Moreover, the digitalization of civil records is accompanied by multiple threats related to cyber threats, and malicious actors can compromise the security of the database to manipulate or steal civil records. Cybersecurity measures such as encryption, firewalls, intrusion detection systems, and regular security audits can protect data and ensure the continuity of civil registration operations. The limitation of this approach is the use of a centralized architecture.

The author in [38] presents a statistic that approximately one-third of births are not registered. The proposal of a model aims to improve the birth registration process. The platform is based on Ethereum technology, which uses an energy-intensive consensus. Additionally, an experimental study could be conducted to confirm the validity of the processes described in the study.

Another author in [39] mentions that half of the world's population faces a problem with the complexity of the procedure for obtaining a birth certificate, and the authentication of these certificates is tedious. The Ethereum blockchain coupled with blockcerts is used to issue and verify a transaction. A major difficulty in using blockcert technology lies in the monetization of generated transactions. Other authors mention that civil registry data mostly exists in paper form, which can differ from one place to another without any form of interoperability within the same country, consequently increasing the falsification of birth certificates since information about births only exists where the birth is registered. A system for registering births and issuing certificates, storing a digital copy of the certificate, and verifying the validity of the certificate is proposed using the private blockchain Corda. The major difficulty of this blockchain lies in its difficulty of integration into existing information systems, and the lack of documentation on the technology is an obstacle to the development of the Corda community [40].

IV. METHODOLOGY

The primary objective of the literature review is to identify, evaluate, and analyze existing research related to the production of birth certificates. In this section, the research question, the research methods aimed at reducing fraud in the birth certificate issuance process, and the gap observed in the authentication process of these certificates by the responsible organizations are presented.

Initially, an algorithmic method is used to write smart contracts that are subject to prior verification by certain peer approvers within the network. The cryptographic method, which relies on asymmetric cryptography, such as RSA encryption and electronic signatures, is also employed. Finally, a virtual simulation method based on containerization (Docker) is used to create network nodes, define the protocol, and the consensus algorithm.

A. Valuable Research Questions

This article aims to develop a blockchain-based system for issuing, tracking, and authenticating birth certificates, ensuring traceability between products from the operational birth certificate from the production facility and the actors involved in their creation without compromising confidentiality due to transparency.

B. Research Methods

1) Network architecture: The agglomerations are interconnected via VPN/MPLS links over an operator network whose backbone is based on MPLS technology. The specificity of MPLS lies in label switching along the path that packets must take to reach the destination network. Since the path is predetermined, routers only need to read the label and do not need to check the packet's IP address. This allows for faster and more efficient routing. Additionally, to enhance VPN security, the IPSec protocol is a complement, as it is widely deployed to restrict access or selectively apply security operations in VPN implementations.

As illustrated in Fig. 1, the edge routers of each agglomeration have interfaces addressed on the operator side with public addresses and on the side of internal local networks with private addresses. In the firewalls, access control lists are configured to filter incoming and outgoing packets to the different networks. The interfaces of the switch connected to the firewall are in different VLANs from the local network VLANs for added security. In each internal network, workstations of decentralized territorial authorities (town halls, civil registry offices, prefectures, sub-prefectures) are connected. The operator network is represented by a cloud due to the security policy.

Fig. 1 presents the network architecture interconnecting 10 agglomerations.

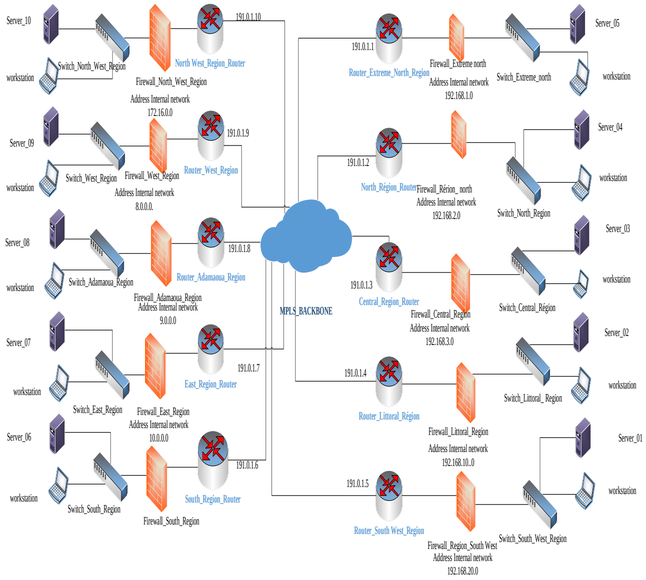


Fig. 1. Physical network architecture of the proposed solution.

2) Smart contract specifications: In this section, the technical specifications of the smart contract for document registration and verification, ensuring immutability and transparency are presented. Developed in Golang due to its simplicity, performance, compatibility with the Hyperledger Fabric blockchain platform, and extensibility, the smart contract offers high performance and efficient concurrent management, making birth certificate management reliable and secure. These specifications cover registration and verification, node transaction approval, transparency and traceability, security, compatibility and extensibility, architecture, and data structure.

C. Registration and Verification

The smart contract incorporates several technical features to ensure optimal and secure management of birth certificate documents:

1) JSON serialization: Documents are serialized into JSON format for easy storage and manipulation on the blockchain.

2) Authenticity: When a birth certificate is submitted to the network, the smart contract verifies its authenticity. A digital fingerprint of the document is then generated and recorded on the blockchain, ensuring data integrity and immutability.

D. Transaction Approval

The smart contract is deployed across multiple nodes in the blockchain network. This decentralized approach ensures an equitable distribution of responsibilities and enhances system security. Additionally, transaction validation is contingent upon approval from all validating nodes in the network.

E. Transparency and Traceability

Every interaction with the smart contract is time-stamped and recorded on the blockchain, creating an immutable audit trail. This enables full traceability of document lifecycle events, from creation to deletion, and facilitates compliance with regulatory requirements.

1) Action logging: Each interaction with the smart contract, including the creation, update, and deletion of documents, is recorded on the blockchain. This comprehensive audit trail ensures full transparency and accountability.

F. Security

The smart contract is deployed across multiple nodes of the blockchain network. This decentralized approach ensures an equitable distribution of responsibilities and enhances system security. Additionally, transaction validation requires the approval of all validating nodes within the network.

G. Cryptographic Method

RSA asymmetric cryptography is used to generate public and private key pairs, as well as digital signatures to ensure the authenticity and integrity of transactions. The public key must be shared with others to receive birth certificate data. The private key is used to sign the transactions within the platform. The key generation process is as follows: [41]

- Choose two distinct, large prime numbers p and q;
- Calculate their product $n = p \times q$ which is called the encryption modulus.;
- Calculate $\varphi(n) = (p-1)(q-1)$ which is Euler's totient function;
- Choose an encryption exponent e such that the greatest common divisor (GCD) of $(e, \varphi(\mathrm{n})) = 1$ i.e., they are coprime;
- Calculate the multiplicative inverse d of e modulo φ(n) using the extended Euclidean algorithm
    - $d \times e \equiv 1 modulo\ \varphi(\mathrm{n})$;
- The public key is the pair (n, e), and the private key is d, which must be kept confidential.

H. Data Security

Blockchain data is immutable and resistant to unauthorized modification or disclosure.

I. The System Architecture

The smart contract architecture is designed to maximize efficiency, security, and transparency in managing birth certificate documents. The key components of this architecture are:

1) Imported packages:

- *Encoding/json:* Used for JSON data manipulation, allowing for easy serialization and deserialization of documents;
- *Fmt:* Used for input and output formatting, facilitating debugging and maintenance.
- *Time:* Used for date and time management, crucial for time-stamped records;
- *github.com/hyperledger/fabric/core/chaincode/shim:* Used for interactions with the Hyperledger Fabric framework, allowing the smart contract to communicate with the blockchain;
- *github.com/hyperledger/fabric/protos/peer:* Used for protocol definitions, ensuring compatibility with Hyperledger Fabric blockchain standards.

2) Data Structures: The smart contract's data structures are designed to provide a granular representation of birth certificates. The following outlines the data structure for birth registration and the resulting birth certificate.

- *BirthRegistration:* a data structure defining the attributes of a birth registration, such as child's information, parental details, and registration metadata.

```
BirthRegistration struct type {
Number       string `json:"N°"`
Type         string `json:" Document type"`
Registration string `json:" Registration designation"`
The          string `json:"Registration date"`
At           string `json:"Registration place"`
Hour         string `json:"Registration hour"`
Surname:  string `json:"Child's surname (s)"`
Name         string `json:" Child's name (s)"`
Sex        string `json:"Child's sex"`
Of0       string `json:"Father's name"`
Sur0      string `json:"Father's surname"`
BornOn0      string `json:"Father's date of birth"`
BornAt       string `json:"Father's place of birth"`
Job string `json:"Father's job"`
IdCartP      string `json:"Father's Id cart"`
Of1        string `json:"Mother's name"`
Pre1       string `json:"Mother's surname"`
BornOn1      string `json:"Mother's date of birth"`
BornAt               string `json:"Mother's place of birth"`
Job2 string `json:" Mother's job "`
Dom        string `json:"Parents' residences"`
SM         string `json:"Marital status of the parents"`
IdCartP        string `json:"Mother's ID card"`
City              string    `json:"City where the declaration was
made"`
Tem              string    `json:"Witnesses"`
DECDate           string    `json:"Date of the declaration"`
Signature of the declarant string    `json:"Signature"`
Signatory        string    `json:"Signatory of the declaration"`
NumCNIDEC        string    `json:"Declarant's national identity
card number"`
ProfDEC        string    `json:"Declarant’s job"`
DateENREG               time.Time  `json:"Date of definitive
registration"`
}
```

- *Birth certificate:* A legal document detailing the birth of a child, including the child's name, date and place of birth, gender, as well as the parents' names, dates and places of birth, nationalities, addresses, and occupations, and any associated declarations.

```
Type of birth certificate struct {
Number        string `json:"Nº"`
Type          string `json:"Type of document"`
PROVINCE      string `json:"PROVINCE"`
DEPARTMENT     string `json:"DÉPARTMENT"`
BOROUGH string `json:" BOROUGH "`
At            string `json:"De"`
Child's surname (s)"`  string `json:"Child's surname (s)"`
Child's name (s)"` string `json:"Child's name (s)"`
Was born      string `json:"Birthplace"`
The          string `json:"Date of birth"`
Sex        string `json:" Child sex "`
At0       string `json:"Nom du père"`
NeLe0     string `json:"Father's date of birth"`
NeA       string `json:"Father's place of birth"`
Natio0    string `json:"Father's Nationality"`
Residences A string `json:"Parents' residences"`
Job string `json:"Father's job"`
At1        string `json:"Mother's name"`
The1       string `json:"Mother's date of birth"`
Born in       string `json:"Mother's place of birth"`
Natio1     string `json:"Mother's nationality"`
Residences string `json:"Mother’s residences"`
Job2 string `json:"Profession de la mère"`
Declaration of    string    `json:"On the declaration of"`
By us string    `json:"By us"`
Center          string    `json:"Registering officer of "`
Assisted by string    `json:"Assisted by"`
Civil registrar's signaturestring       `json:" Civil registrar's
signature"`
DEC Number          string     `json:"Original birth certificate
number"`
Dated          time.Time `json:"Dated"`
```

3) Main Functions:

- *Initialization Function (Init):* Initializes the smart contract. It requires no input parameters and returns a success response if the initialization is successful.
- *Invocation Function (Invoke):* Routes the request to the appropriate function based on the invoked method's name. It requires the parameters stub, function, args as input and returns the response of the invoked method as output Fig. 1.

4) Functions of a Birth Certificate:

- Create Birth Certificate: Creates a new birth certificate. As input, it requires an array of 29 strings representing the details of the birth certificate and returns a success response if the birth certificate is successfully created.
- Retrieve Birth Certificate: Retrieves a birth certificate by its unique number. As input, it takes an array with one string (the unique number) and returns the details of the birth certificate if found.
- *Consult All Birth Certificates:* Retrieves all birth certificates. As input,

5) Functions of a birth certificate

- CreateBirthCertificate: Creates a new birth certificate. As input, it takes an array of strings representing the birth details (name, date of birth, place of birth, parents' names, etc.) and returns the unique identifier of the newly created certificate.
- RetrieveBirthCertificate: Retrieves a birth certificate based on its unique identifier. As input, it takes a string representing the unique identifier and returns an array of strings containing the birth certificate details.
- ListAllBirthCertificates: Lists all birth certificates. It doesn't require any input parameters and returns a list of all birth certificates in JSON format.

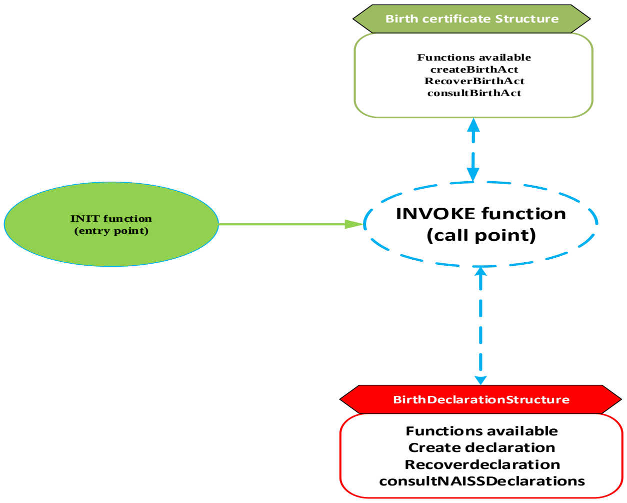
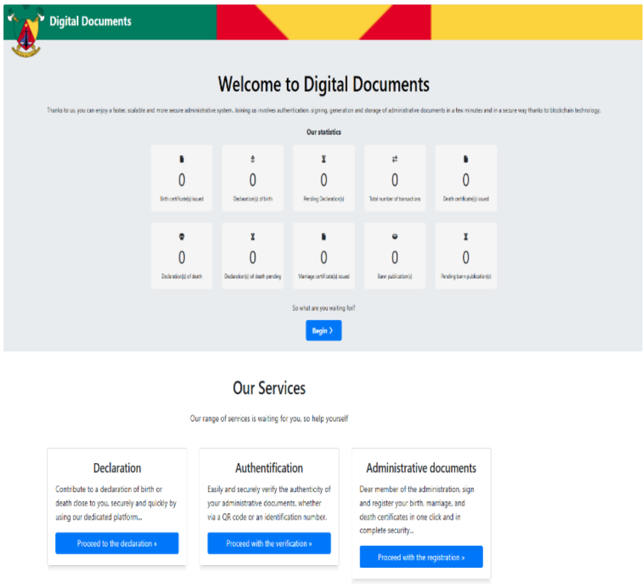


Fig. 2. Components of a smart contract.

V. Results and Discussion

In this part, the effectiveness of the methodology outlined in the preceding section is demonstrated. Through a series of tests, the performance of birth certificate registration, authentication processes, and the generation of birth declaration statistics are evaluated. To accomplish this, the capabilities of the Hyperledger Fabric private blockchain platform are leverage. Fig. 3 showcases the user interface of the government portal specifically designed for civil registry and authentication. This portal facilitates the creation, verification, and statistical analysis of civil status records, providing granular data at weekly, monthly, and annual levels, segmented by region.



Welcome to Digital Documents

Thanks to us, you can enjoy a faster, scalable and more secure administrative system. Joining us involves authentication, signing, generation and storage of administrative documents in a few minutes and in a secure way thanks to blockchain technology.

Our statistics

0 Birth certificate(s) issued	0 Declaration(s) of birth	0 Pending Declaration(s)	0 Total number of transactions	0 Death certificate(s) issued
0 Declaration(s) of death	0 Declaration(s) of death pending	0 Marriage certificate(s) issued	0 Bann publication(s)	0 Pending bann publication(s)

So what are you waiting for?

Our Services

Our range of services is waiting for you, so help yourself

Declaration

Contribute to a declaration of birth or death close to you, securely and quickly by using our dedicated platform...

Authentification

Easily and securely verify the authenticity of your administrative documents, whether via a QR code or an identification number.

Administrative documents

Dear member of the administration, sign and register your birth, marriage, and death certificates in one click and in complete security...

Fig. 3. Government portal homepage.

To enhance traceability and accountability, we recommend the following roles:

Healthcare professional: Verifies births and submits declarations to approving nodes.

Civil registrar: Creates and digitally signs birth certificates.

After a healthcare professional's profile is created, they submit a birth declaration for approval (Fig. 4). The declaration includes a QR code with a unique identifier, the signatory's name, and ID number. It contains details such as...

Establishment name: This is the name of the hospital that issued the birth certificate;

Child's name and surname: This is the first name and last name of the infant;

Date of birth: This refers to the date the baby was born;

Place of birth: This is the location where the baby was born;

Gender: This refers to the sex, in this case, male.

Subsequent to the infant's details, we have the following parental information:

Name and surname of father: Father's full name;

Father's occupation: Father's profession;

Father's date of birth: Father's date of birth;

Father's place of birth: Father's place of birth;

Father's Id card number: Father's national ID number;

Mother's full name: Mother's full name;

Mother's occupation: Mother's profession;

Mother's date of birth: Mother's date of birth;

Mother's place of birth: Mother's place of birth;

Mother's ID card number: Mother's national ID number;

Parent’s address: Parents' address;

Marital status: Marital status. In this case, the parents are married.

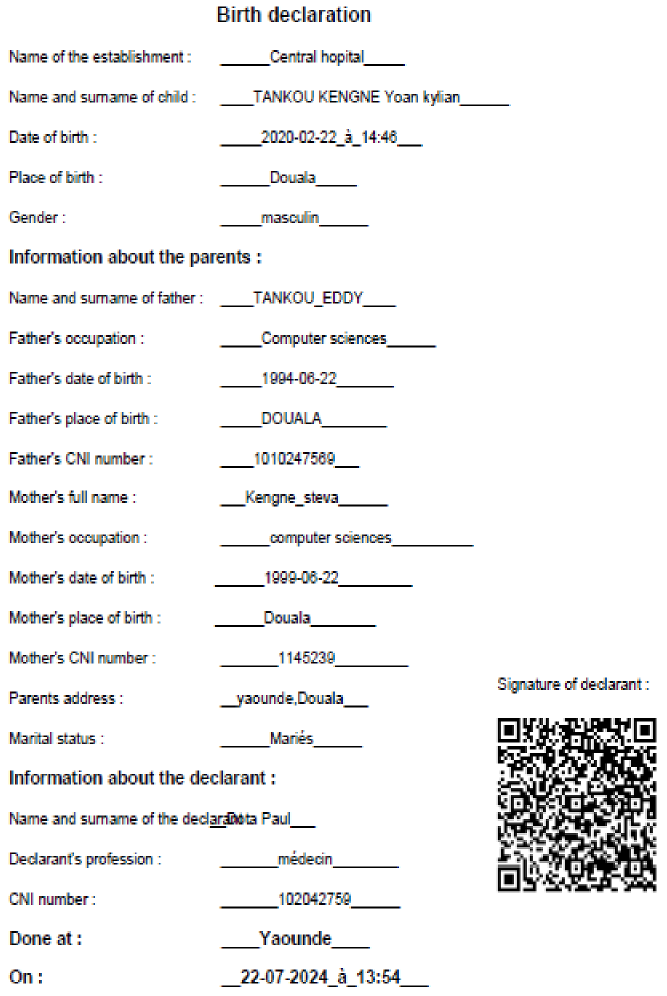
Subsequent to the parental data, the registrar's information is provided:

Name and surname of the declarant: Dota Paul;

Profession: Healthcare professionals;

ID card number;

Place and date of the declaration: [Location], on [Date] at [Time];



Birth declaration

Name of the establishment :	_____Central hopital_____
Name and surname of child :	____TANKOU KENGNE Yoan kylian______
Date of birth :	_____2020-02-22_à_14:46___
Place of birth :	______Douala_____
Gender :	_____masculin______
Information about the parents :	
Name and surname of father :	____TANKOU_EDDY____
Father's occupation :	_____Computer sciences_____
Father's date of birth :	_____1994-06-22______
Father's place of birth :	_____DOUALA_______
Father's CNI number :	____1010247569___
Mother's full name :	___Kengne_steva_____
Mother's occupation :	______computer sciences_________
Mother's date of birth :	______1999-06-22_________
Mother's place of birth :	______Douala________
Mother's CNI number :	_______1145239_________
Parents address :	__yaounde,Douala___
Marital status :	______Mariés______
Information about the declarant :	
Name and surname of the declar	__Dota Paul___
Declarant's profession :	_______médecin________
CNI number :	_______102042759______
Done at :	____**Yaounde**____
On :	__**22-07-2024_à_13:54**__

Signature of declarant :

Fig. 4. Birth certificate.

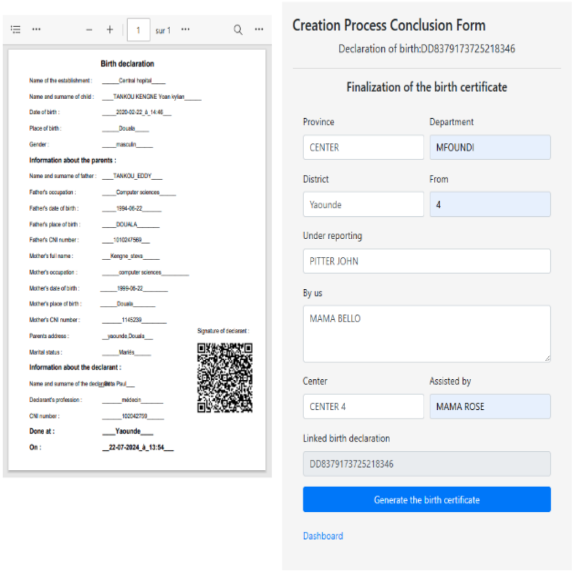
Finally, there is the QR code, which is a type of two-dimensional barcode typically composed of square modules arranged in a square on a white background. It contains the unique identification number (UIN), the name of the document signer, and their national identity card number. Additionally, this QR code generated on the form bears the signature of the civil registrar, allowing for verification of the signer and the document's author. This signature incorporates the SHA-256 cryptographic hash function, which generates a unique 64-character hexadecimal hash for each block of transactional data.

The security of the forms primarily relies on a UIN (unique identification number) designed using a pseudorandom number generator. At the end of the declaration form completion process, a UIN issued by the aforementioned function serves as a unique identifier for each document.

This cryptographic hash function dates back to 2001 and originated from the NSA (National Security Agency). It is increasingly used in blockchain technology today due to its high level of security. Furthermore, its algorithm is undeniably complex, as the generation of a SHA-256 hash requires a highly sophisticated binary calculation, utilizing several compression functions (addition, substitution, and rotations). Ultimately, it relies on the Merkle tree to provide a reliable guarantee of the integrity of the manipulated data.

Once the birth declaration is established, a user account is created for the civil registrar to finalize the birth certificate issuance process. The civil registrar profile is authorized to issue civil status acts. Additionally, the national identity card number entered during account creation uniquely identifies the civil registrar within the platform. Once the user account is created, it is hashed using the SHA-256-bit hash function and stored in the distributed node database for strong authentication. Thus, during account authentication, the account hash is compared to the hash stored in the database; if there is a match, the authenticity of the account owner is guaranteed.

The civil registrar logs into the platform using these created user account parameters and finalizes the birth certificate issuance process. Fig. 5 illustrates this process.



Creation Process Conclusion Form

Declaration of birth:DD8379173725218346

Finalization of the birth certificate

Province: CENTER

Department: MFOUNDI

District: Yaounde

From: 4

Under reporting: PITTER JOHN

By us: MAMA BELLO

Center: CENTER 4

Assisted by: MAMA ROSE

Linked birth declaration: DD8379173725218346

Fig. 5. Birth certificate finalization form.

When the finalization form is opened, the civil registrar completes the missing fields and generates the final act. They perform the final checks on the document and submit the transaction via the API (Application Programming Interface) to the network for consensus. All nodes with a copy of the smart contract execute the transaction, approve it, and return the result to the API. In turn, the API sends it to the scheduling service, which hashes the transaction and generates the block. Having only one transaction per block in the Hyperledger Fabric platform makes it faster compared to the Ethereum blockchain. Fig. 6 shows the generated birth certificate.

The left figure represents the previously established birth declaration, and the right figure represents the birth certificate finalization form corresponding to the previously established birth declaration. On this right-hand form, there are the following information: Province: the province where the birth certificate is issued; Department: the department where the certificate is issued; District: the district where the certificate is issued, as well as the district number, which is 4; Sub-declaration: the name of the person who establishes the act; By us: it represents the name of the civil registrar who signs the birth certificate, in the locality of the center and the district of Yaoundé 4; Assisted by: represents the civil registrar's secretary who assisted the civil registrar in this process; Linked birth declaration: is the UIN generated on the previously established birth declaration, this number links the birth declaration to a birth certificate and is not modifiable.

The civil registrar finally generates the birth certificate transaction and redirects us to the interface allowing us to view the transaction identifiers of the birth declaration and the corresponding birth certificate issuance in the blockchain. A SHA-256-bit hash function is finally generated and stored in the distributed node database for strong authentication. Thus, during document authentication, the document hash is compared to the hash stored in the database; if there is a match, the authenticity is guaranteed.

In a blockchain context, each actor who wants to interact with the network needs an identity. In this context, one or more certification authorities can be used to define the members of an organization from a digital perspective. It is the certification authority that provides the basis for actors in an organization to have a verifiable identity.

Hyperledger Fabric provides a built-in certification authority component to enable the creation of certification authorities in formed blockchain networks. This component, called Fabric CA, is a private root certification authority provider capable of managing the digital identities of Fabric participants in the form of X.509 certificates.

In the operational platform for producing birth certificate forms, an X node acts as the root certification authority (Fabric_CA), and the intermediate certification authorities are the local servers with respective users (Fabric client) being hospitals and community users.

Certificates are generated offline by the root certification authority (X). Once these certificates are created, they are securely distributed to the intermediate certification authorities, which in turn issue certificates to the platform users, namely hospitals, health centers, and communities.

To evaluate the performance of the root certification authority, the SCP protocol via SSH (Secure Copy Protocol) allows us to develop a bash script that runs intermittently to ensure high data availability in case of failure of the main root_CA for each organization. This SCP protocol allows the copying of information from a root_CA1 to a root_CA2 every three seconds (3s) through a bash script [42].

The civil registrar profile is authorized to establish birth certificates. Furthermore, the national identity card number provided during account creation uniquely identifies the civil registrar within the platform.

Once the user account is created, it is hashed using the SHA-256-bit hashing function and stored in the distributed database of nodes for strong authentication purposes. Thus, during account authentication, the hash of the account is compared to the hash stored in the database; if there is a match, the authenticity of the account holder is guaranteed.

On this birth certificate (Fig. 7), there are the following information:

Birth certificate number: A unique identifier characterized by the initials of the locality that issued the certificate.

Province: The province where the birth certificate was issued.

Department: The department where the certificate was issued.

District: The district where the certificate was issued, specifically District number 4.

Name and surname of child: The given name(s) and surname of the newborn.

Place of birth: The location where the newborn was born.

Date of birth: The date and time of the newborn's birth.

Gender: The sex of the child, in this case, male.

Of: The given name and surname of the newborn's father.

Born on: The date of birth of the newborn's father.

At: The place of birth of the newborn's father.

Nationality: The nationality of the newborn's father.

Resident at: The residence of the newborn's father.

Occupation: The profession of the newborn's father.

By us: The name of the civil registrar who issued the certificate.

Under declaration of: The person who assisted the civil registrar in this process.

Office: The district of the locality and the number associated with that district in that locality.

Assisted by: The name of the civil registrar's secretary who assisted in the issuance of the birth certificate.

The QR code contains a unique identification number (NIU), the name of the document signer, and their national identity card number. This QR code is a type of two-dimensional barcode typically composed of black square modules arranged in a square on a white background, encoding the unique identification number (NIU), the name of the document signer, and their national identity card number.

Additionally, the QR code generated on the form bears the signature of the civil registrar, allowing for verification of the signer and the document's author. This signature incorporates the SHA-256 cryptographic hash function, which generates a unique 64-character hexadecimal hash for each block of transactional data.

The security of the forms initially relies on a NIU (unique identification number generated using a pseudo-random number generator). At the end of the declaration form completion process, a NIU issued by the aforementioned function serves as a unique identifier for each document. This cryptographic hash function dates back to 2001 and originated from the National Security Agency (NSA). It is now increasingly used in blockchain technology due to its high level of security. Furthermore, its algorithm is undeniably complex as the generation of a SHA-256 hash requires a sophisticated binary calculation, employing various compression functions (addition, substitution, and rotations). Ultimately, it relies on the Merkle tree to provide a reliable guarantee of the integrity of the manipulated data.

All data contained in the act is hashed using the SHA-256 hash function. The resulting hash is signed by the civil registrar using their private key. The obtained signature is finally recorded in the QR code.

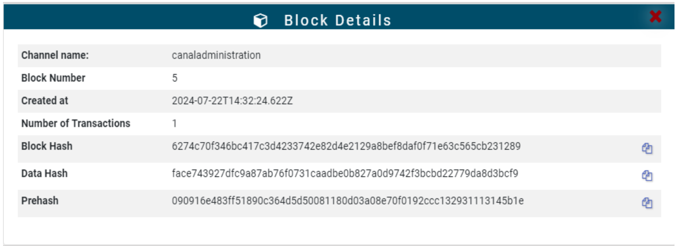
On the generated birth certificate, all contained data is hashed using the SHA-256 hash function. The resulting hash is signed by the civil registrar using their private key. The obtained signature is finally recorded in the QR code. Fig. 8 illustrates this solution.

The details of the block containing the transaction are shown in Fig. 6.

Fig. 8 provides an experimental history of the number of blocks generated per hour. The figure visualizes the hourly and minutely block count over time. In terms of scalability, measured by the number of transactions processed per unit of time, this work demonstrates approximately 3500 transactions per second on the Hyperledger Fabric blockchain, compared to 38 transactions per second on the Ethereum platform referenced in the literature. This exponential increase in transaction processing speed makes the proposed platform a suitable solution for the digitalization of birth certificates in localities with multiple civil registry centers, including secondary centers and affiliated civil registry offices.

A. Comparison of Works with those of Other Authors

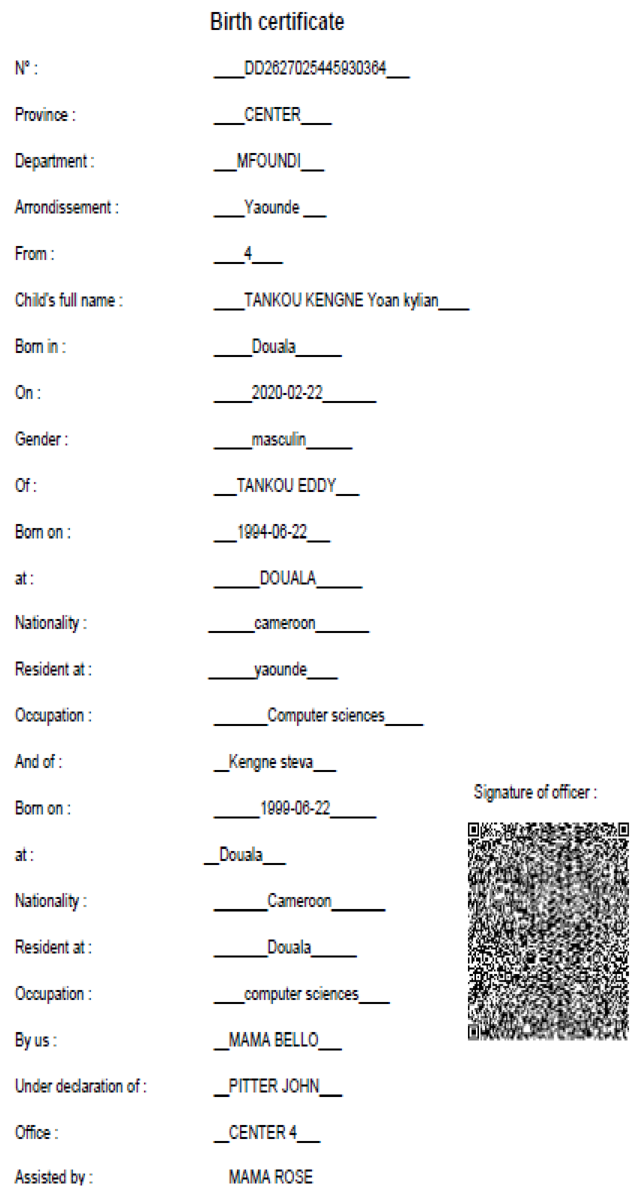
Table II compares the proposed scheme with those of previous studies to identify both areas of convergence and the unique contributions of this research.



Block Details

Channel name:	canaladministration
Block Number	5
Created at	2024-07-22T14:32:24.622Z
Number of Transactions	1
Block Hash	6274c70f346bc417c3d4233742e82d4e2129a8bef8daf0f71e63c565cb231289
Data Hash	face743927dfc9a87ab76f0731caadbe0b827a0d9742f3bcbd22779da8d3bcf9
Prehash	090916e483ff51890c364d5d50081180d03a08e70f0192ccc132931113145b1e

Fig. 6. Details of the transaction block.



Birth certificate

N° : ___DD2627025445930364___

Province : ___CENTER___

Department : __MFOUNDI__

Arrondissement : ___Yaounde ___

From : ___4___

Child's full name : ___TANKOU KENGNE Yoan kylian___

Born in : ____Douala_____

On : ____2020-02-22_____

Gender : ____masculin_____

Of : __TANKOU EDDY__

Born on : __1994-06-22__

at : _____DOUALA_____

Nationality : _____cameroon_____

Resident at : _____yaounde___

Occupation : _____Computer sciences____

And of : __Kengne steva__

Born on : _____1999-06-22_____

at : __Douala__

Nationality : _____Cameroon_____

Resident at : _____Douala_____

Occupation : ___computer sciences___

By us : __MAMA BELLO__

Under declaration of : __PITTER JOHN__

Office : __CENTER 4__

Assisted by : MAMA ROSE

Signature of officer :

Fig. 7. Birth certificate.

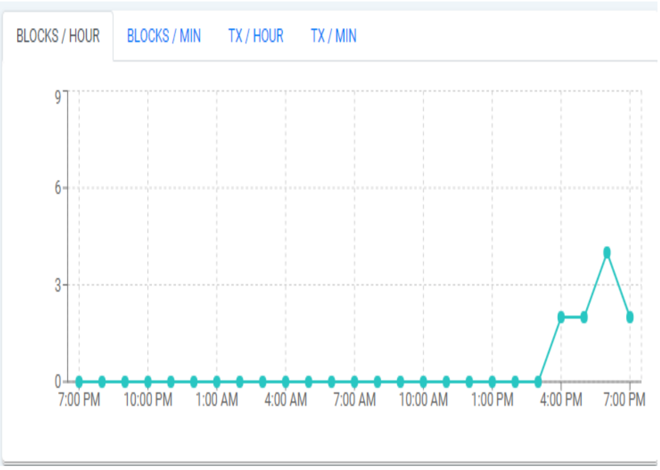


Fig. 8. Hourly blockchain statistics.

TABLE II. COMPARISON OF WORKS WITH RECENT LITERATURE

Authors **Criteria**		**Proposals**	**Shah, et al. [14]**	**Mthethwa, et al. [43]**	**Thamrin, et al. [44]**	**Shi, et al. [38]**
Energy consumption		▬	+++	▬	+++	+++
Blockchain	**Blockchain type**	Private	Public	Public	Public	Public
	Technology used	Hyperledger fabric	Ethereum	Ethereum	Ethereum	Ethereum
Security features	Unique ID for civil status documents	☑	☒	☒	☒	☒
	QR code on a civil registry form	☑	☒	☑	☒	☒
	Transaction ID	☑	☑	☒	☑	☑
	Digital signature on civil status documents	☑	☑	☒	☑	☑
	SHA-256	☑	☒	☒	☒	☒
	Consensus	Open	Closed	☒	Closed	Closed
	Distributed data storage	☑	☑	☒	☑	☑
	Data privacy in channels	☑	☒	☒	☒	☒
Functionality	Birth certificate	☑	☑	☒	☒	☑
	Death certificate	☑	☑	☒	☒	☑
	Marriage banns publication	☑	☒	☒	☒	☒
	Birth record establishment	☑	☒	☒	☒	☒
	Death record establishment	☑	☒	☒	☒	☒
	Marriage record establishment	☑	☒	☒	☒	☒
Scalability (Transactions per second)		**3500**	**38** [45, 46]	☑	**38** Berné [45] **[46]**	**38** Berné [45] **[46]**
Mining		☒	☑	☒	☑	☑
Evolvability		☑	☒	☑	☒	☒
Distributed architecture		☑	☑	☒	☑	☑
License (open source)		☑	☒	☒	☒	☒
From a demographic standpoint		☑	☒	☒	☒	☒

This section benchmarks the results obtained against existing literature to elucidate both shared insights and the original contributions to the field.

Legend: The symbol "☑" indicates the presence of criterion "☒", while its absence signifies the lack of this criterion. The sign "▬" means "less than" and the sign "+++" means "more than" compared to the negation.

Until 2021, the Ethereum platform used the Proof of Work (PoW) consensus mechanism. This mechanism had the major drawback of being very energy-intensive and susceptible to 51% attacks [123]. In contrast, the consensus used in this work, which is based on the Hyperledger platform, is open and consumes less energy, making it an ideal solution for countries with significant energy deficits. In other words, developers can define their own consensus rules or choose between PoW, PoS, PoA, etc.

From a security standpoint, the results obtained in this work are varied. Forms are equipped with a unique identification number obtained after the generation of the act form, thus characterizing the uniqueness of a document in the distributed database. The QR code printed on this form allows verification of the signatory author of a document as well as the associated key pair (public key and private key). The transaction identifier recorded in the blockchain database guarantees the tamper-proof nature of documents, not to mention the SHA-256 cryptographic hash for document signing. Hyperledger Fabric has a particular characteristic in that the created channels can be accessible or visible to a category of actors in the private blockchain, thus guaranteeing confidentiality in transactions; unlike the public blockchain used by Ethereum where channels are visible to all network members, and anyone can create and read blockchain transactions [47].

Based on the offered functionalities, this work proposes a complete system for managing civil status acts, including declarations (birth declaration, death declaration, marriage ban publication) and the establishment of civil status acts themselves (birth certificate, death certificate, marriage certificate), unlike the works of [14, 38, 44] which focused solely on the security of the form by QR code on the one hand, and on birth and death certificates on the other.

In terms of scalability, which represents the number of transactions processed per unit of time, this paper presents statistics of the order of 3500 transactions per second for the Hyperledger Fabric blockchain compared to 38 transactions per second for the Ethereum platform used in the works [14, 38, 44]. This exponential speed in transaction processing makes the platform a suitable solution for the digitalization of civil status in regions of the territory that have several civil status centers, secondary centers, and affiliated centers.

Mining, on the other hand, is a process of creating a block of transactions and adding it to the Ethereum blockchain. However, since it is computers that execute instructions, they are rewarded at the end of the transaction. For a government platform for the digitization of civil status acts, this solution is inefficient in that it will be necessary to reward the authors of transactions each time, unlike the Hyperledger platform which does not require mining for transaction validation. Finally, the evolutionary factor allows a government authority to consider a possible extension without making major changes. To add an organization, simply create the organization's channel, and its members have access to this channel according to their level of accreditation. The open-source nature of the proposed solution allows developers to make modifications to the source code as they see fit, which is not the case with the Ethereum technology used by [35, 48].

VI. Future Work

In remote regions lacking administrative infrastructure and electrical grids, establishing birth certificates is often challenging or even impossible. The use of embedded devices like the Raspberry Pi, which are inexpensive and energy-efficient nano computers capable of operating on alternative power sources such as batteries and solar panels, enables the creation of autonomous systems. Equipped with blockchain software, these systems can record birth data locally without requiring a permanent internet connection. When an internet connection becomes available, this data can be synchronized with a secure blockchain, ensuring data integrity [49, 50]. This innovative solution offers numerous advantages, including reliability (data is secure and immutable), accessibility (the solution is suitable for rural areas), equity (all individuals, even in the most remote areas, have access to an official birth certificate), and sovereignty (local communities have greater control over their data). As a result, it contributes to digital inclusion and improves the living conditions of the most vulnerable populations.

Based on the relevance of the research findings, this section outlines future directions for the design of other identity documents and official certificates to reduce document fraud. These include: Marriage certificates, which are issued by a civil registrar and serve as legal proof of marital status; death certificates, which are official administrative documents attesting to a person's death and issued by the civil registrar of the municipality where the death occurred; land titles, which are official certifications of real estate ownership and are considered irrevocable and inalienable; and diplomas, which are official documents conferring a degree or qualification.

VII. Conclusion

The birth certificate is a legal and vital document that proves an individual’s civil status. The fact that it certifies an individual’s identity and family situation makes it a major element generally required in administrative procedures such as national identity cards, passports, diplomas, bank transactions, etc. While blockchain technology is inherently secure, it remains vulnerable to side-channel attacks that exploit indirect information to compromise security[51, 52]. These attacks can leverage variations in execution time to extract cryptographic keys, potentially leading to the alteration of transactions. To mitigate these threats, techniques such as obfuscation, which aims to make code difficult to understand, analyze, or modify, are crucial. As a result, continuous vigilance is necessary to safeguard blockchain systems against such vulnerabilities. This article aims to contribute to the development of a system for issuing, tracking, and authenticating birth certificates based on Hyperledger Fabric blockchain technology. To achieve this, the network architecture model is proposed, it based on VPN/MPLS on one hand, and on the other hand, the design of the smart contract characterized by the creation, registration, and verification of birth certificates, the approval of transactions by approving nodes, the transparency and traceability of the various processes, the security throughout the network, and the architecture composed of imported packages, Data Structures, main Functions, Birth Declaration Functions, and Birth Certificate Functions. The proposed platform is an application for creating, registering, and verifying the authenticity of birth certificates, ensuring robust and efficient security-and transparency. This platform is particularly important for organizations responsible for producing birth certificates, diplomatic representations, judicial and administrative authorities, etc.

References

[1] J. Arkko, "The influence of internet architecture on centralised versus distributed internet services," Journal of Cyber Policy, vol. 5, no. 1, pp. 30-45, 2020.

[2] R. Heeks, "Centralised vs. decentralised management of public information systems: a core-periphery solution," Information Systems for Public Sector Management Working Paper, no. 7, 1999.

[3] K. K. Vaigandla, R. Karne, M. Siluveru, and M. Kesoju, "Review on blockchain technology: architecture, characteristics, benefits, algorithms, challenges and applications," Mesopotamian Journal of CyberSecurity, vol. 2023, pp. 73-84, 2023.

[4] J. Sacha et al., "Decentralising a service-oriented architecture," Peer-to-Peer Networking and Applications, vol. 3, pp. 323-350, 2010.

[5] L. Ghiro et al., "What is a Blockchain? A Definition to Clarify the Role of the Blockchain in the Internet of Things," arXiv preprint arXiv:2102.03750, 2021.

[6] C. Zadra-Veil et al., "Blockchain et Immobilier: Le Smart Bail," ed, 2021.

[7] A. S. Ghanghoria, A. S. A. Raja, V. J. Bachche, and M. N. Rathi, "Secure E-documents storage using blockchain," Int. Res. J. Eng. Technol.(IRJET), vol. 7, pp. 1972-1974, 2020.

[8] A. Zaky and I. G. B. B. Nugraha, "Increase activity time efficiency in official documents management using blockchain-based distributed data storage," in 2019 International Conference on Electrical Engineering and Informatics (ICEEI), 2019: IEEE, pp. 81-86.

[9] A. S. Rajasekaran, M. Azees, and F. Al-Turjman, "A comprehensive survey on blockchain technology," Sustainable Energy Technologies and Assessments, vol. 52, p. 102039, 2022.

[10] M. Belotti, N. Božić, G. Pujolle, and S. Secci, "A vademecum on blockchain technologies: When, which, and how," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3796-3838, 2019.

[11] D. Yaga, P. Mell, N. Roby, and K. Scarfone, "Blockchain technology overview," arXiv preprint arXiv:1906.11078, 2019.

[12] N. Mishra, S. Mistry, S. Choudhary, S. Kudu, and R. Mishra, "Food traceability system using blockchain and QR code," in IC-BCT 2019: Proceedings of the International Conference on Blockchain Technology, 2020: Springer, pp. 33-43.

[13] M. Darisi, O. Modi, V. Mistry, and D. Patel, "MapReduce-based framework for blockchain scalability," in IC-BCT 2019: Proceedings of the International Conference on Blockchain Technology, 2020: Springer, pp. 119-132.

[14] V. Shah, K. Padia, and V. B. Lobo, "Application of Blockchain Technology in Civil Registration Systems," in IC-BCT 2019: Springer, 2020, pp. 191-204.

[15] M. Hashemi Joo, Y. Nishikawa, and K. Dandapani, "Cryptocurrency, a successful application of blockchain technology," Managerial Finance, vol. 46, no. 6, pp. 715-733, 2020.

[16] T. McGhin, K.-K. R. Choo, C. Z. Liu, and D. He, "Blockchain in healthcare applications: Research challenges and opportunities," Journal of network and computer applications, vol. 135, pp. 62-75, 2019.

[17] P. K. Ghosh, A. Chakraborty, M. Hasan, K. Rashid, and A. H. Siddique, "Blockchain application in healthcare systems: a review," Systems, vol. 11, no. 1, p. 38, 2023.

[18] D. Dujak and D. Sajter, "Blockchain applications in supply chain," SMART supply network, pp. 21-46, 2019.

[19] N. Kawaguchi, "Application of blockchain to supply chain: Flexible blockchain technology," Procedia Computer Science, vol. 164, pp. 143-148, 2019.

[20] A. Saari, J. Vimpari, and S. Junnila, "Blockchain in real estate: Recent developments and empirical applications," Land Use Policy, vol. 121, p. 106334, 2022.

[21] J. Huang, D. He, M. S. Obaidat, P. Vijayakumar, M. Luo, and K.-K. R. Choo, "The application of the blockchain technology in voting systems: A review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-28, 2021.

[22] R. Rivera, J. G. Robledo, V. M. Larios, and J. M. Avalos, "How digital identity on blockchain can contribute in a smart city environment," in 2017 International smart cities conference (ISC2), 2017: IEEE, pp. 1-4.

[23] H. Song, N. Zhu, R. Xue, J. He, K. Zhang, and J. Wang, "Proof-of-Contribution consensus mechanism for blockchain and its application in intellectual property protection," Information processing & management, vol. 58, no. 3, p. 102507, 2021.

[24] X. Wang et al., "Survey on blockchain for Internet of Things," Computer Communications, vol. 136, pp. 10-29, 2019.

[25] M. A. Khan, S. M. Khan, and S. K. Subramaniam, "SECURITY ISSUES IN CLOUD COMPUTING USING EDGE COMPUTING AND BLOCKCHAIN: THREAT, MITIGATION, AND FUTURE TRENDS-A SYSTEMATIC LITERATURE REVIEW," Malaysian Journal of Computer Science, vol. 36, no. 4, 2023.

[26] M. Swan, "Anticipating the economic benefits of blockchain," Technology innovation management review, vol. 7, no. 10, pp. 6-13, 2017.

[27] D. Guegan, "Blockchain publique versus blockchain privee: Enjeux et limites," 2017.

[28] P. Zheng, Q. Xu, Z. Zheng, Z. Zhou, Y. Yan, and H. Zhang, "Meepo: Sharded consortium blockchain," in 2021 IEEE 37th International Conference on Data Engineering (ICDE), 2021: IEEE, pp. 1847-1852.

[29] M. Valenta and P. Sandner, "Comparison of ethereum, hyperledger fabric and corda," Frankfurt School Blockchain Center, vol. 8, pp. 1-8, 2017.

[30] A. Baliga, N. Solanki, S. Verekar, A. Pednekar, P. Kamat, and S. Chatterjee, "Performance characterization of hyperledger fabric," in 2018 Crypto Valley conference on blockchain technology (CVCBT), 2018: IEEE, pp. 65-74.

[31] G. Greenspan, "Multichain private blockchain-white paper," URl: http://www. multichain. com/download/MultiChain-White-Paper. pdf, vol. 85, 2015.

[32] M. Hearn and R. G. Brown, "Corda: A distributed ledger," Corda Technical White Paper, vol. 2016, p. 6, 2016.

[33] D. Level, "Level DB database," Level DB Database, 2018.

[34] D. Couch, "Couch DB database," Couch DB Database, 2018.

[35] V. Shah, K. Padia, and V. B. Lobo, "Application of Blockchain Technology in Civil Registration Systems," in IC-BCT 2019: Proceedings of the International Conference on Blockchain Technology, 2020: Springer, pp. 191-204.

[36] C. Chen, C. Lin, M. Chiang, Y. Deng, P. Chen, and Y. Chiu, "A traceable online will system based on blockchain and smart contract technology. Symmetry. 13 (3), 466 (2021)," ed.

[37] P. K. Okoth, "Security challenges in civil registration: safeguarding vital information in an evolving landscape," World Journal of Advanced Research and Reviews, vol. 19, no. 1, pp. 1051-1071, 2023.

[38] J. Shi, S. K. N. Danquah, W. J. I. J. o. E. R. Dong, and P. Health, "A Novel Block Chain Method for Urban Digitization Governance in Birth Registration Field: A Case Study," vol. 19, no. 15, p. 9309, 2022.

[39] N. Sharma, M. Afzal, and A. Dixit, "Blockchain-blockcerts based birth/death certificate registration and validation," International Journal of Information Technology (IJIT), vol. 6, no. 2, 2020.

[40] C. U. Bennett, O. A. Ojerinde, H. O. Aliyu, and A. S. Adepoju, "Registration and Verification of Birth Certificate using Blockchain Technology," 2022.

[41] M. M. Ebenezer, P. Félix, M. Yannick, S. N. P. Junior, N. N. J. I. J. o. A. C. S. Léandre, and Applications, "Contribution to the improvement of cryptographic protection methods for medical images in DICOM format through a combination of encryption method," vol. 12, no. 4, 2021.

[42] T. T. M. Eddy, B. B. Georges, and N. E. P. Salomon, "Towards a New Model for the Production of Civil Status Records Using Blockchain," Journal of Information Security, vol. 14, no. 1, pp. 52-75, 2022.

[43] S. Mthethwa, N. Dlamini, and G. Barbour, "Proposing a blockchain-based solution to verify the integrity of hardcopy documents," in 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), 2018: IEEE, pp. 1-5.

[44] R. M. Thamrin, E. P. Harahap, A. Khoirunisa, A. Faturahman, and K. J. A. j. o. r. i. Zelina, "Blockchain-based land certificate management in indonesia," vol. 2, no. 2, pp. 232-252, 2021.

[45] R. Berné. "Qu'est-ce que la scalabilité d'une blockchain ?" https://cryptoast.fr/scalabilite-definition-explication/ (accessed 11/11, 2023).

[46] J.-P. Delahaye, "Chapitre 7. L'univers des cryptomonnaies," in Au-delà du Bitcoin. Paris: Dunod, 2022, pp. 149-173.

[47] M. Castro and B. J. A. T. o. C. S. Liskov, "Practical byzantine fault tolerance and proactive recovery," vol. 20, no. 4, pp. 398-461, 2002.

[48] T. J. W. V. L. R. Cutts, "Smart contracts and consumers," vol. 122, p. 389, 2019.

[49] K.-K. R. Choo, M. M. Kermani, R. Azarderakhsh, and M. Govindarasu, "Emerging embedded and cyber physical system security challenges and innovations," IEEE Transactions on Dependable and Secure Computing, vol. 14, no. 3, pp. 235-236, 2017.

[50] A. Jalali, R. Azarderakhsh, M. M. Kermani, and D. Jao, "Towards optimized and constant-time CSIDH on embedded devices," in Constructive Side-Channel Analysis and Secure Design: 10th International Workshop, COSADE 2019, Darmstadt, Germany, April 3–5, 2019, Proceedings 10, 2019: Springer, pp. 215-231.

[51] D. Jauvart, J. J. Fournier, N. El-Mrabet, and L. Goubin, "Improving side-channel attacks against pairing-based cryptography," in Risks and Security of Internet and Systems: 11th International Conference, CRiSIS 2016, Roscoff, France, September 5-7, 2016, Revised Selected Papers 11, 2017: Springer, pp. 199-213.

[52] X. Lou, T. Zhang, J. Jiang, and Y. Zhang, "A survey of microarchitectural side-channel vulnerabilities, attacks, and defenses in cryptography," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-37, 2021.

Optimizing Customer Interactions: A BERT and Reinforcement Learning Hybrid Approach to Chatbot Development

Dr. K.R.Praneeth[1], Dr. Taranpreet Singh Ruprah[2], Dr. J Naga Madhuri[3], Dr. A L Sreenivasulu[4],
Syed Shareefunnisa[5], Dr. Vuda Sreenivasa Rao[6]

Assistant Professor, School of Management, The Apollo University, The Apollo Knowledge City Campus,
Chittoor, Andhra Pradesh, India[1]
Associate Professor, Amity University, Jaipur, Rajasthan, India[2]
Assistant Professor, Department of English, Velagapudi Rama Krishna Siddhartha Engineering College
(Deemed to be University), Kanuru, Vijayawada, India[3]
Professor of CSE, Vignana Bharathi Institute of Technology, Telangana, Aushapur(V), Ghatkesar(M) Medchal (D),
Hyderabad, Telangana, India[4]
Assistant Professor, VFSTR Deemed to be University, Guntur, Andhra Pradesh, India[5]
Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, AP, India[6]

Abstract—**In the case of chatbots massive progress has been made, but problems remain in handling the complexity of the sentence and in context relevance. Traditional models can be rather insufficient when it comes to providing various levels of detail in the responses to the end-users' questions, particularly when referring to customer support scenarios. To overcome these limitations, this research comes up with a new model which combines the BERT model with DRL. Through DRL, BERT pre-training is adding flexibility and correspondence to correctly perceive contextual delicate matters in the response. The proposed method includes the following pipeline where in; data tokenization, conversion to lowercase characters, lemmatization and then passes through the BERT fine-tuned model. DRL is utilized to optimize the chatbot's response in the light of long term rewards and the conversational history, the interactions are formulated as a Markov Decision Process with the reward functions based on cosine similarity of the consecutive responses. This makes it feasible for the chatbot to provide context based replies in addition to the option of constant learning for enhanced performance. It also proved that the accuracy and relevance of the BERT-DRL hybrid system were higher than traditional models according to the BLEU and ROUGE scores. The performance of the chatbot also increases with the length of the conversation and the transitions from one response to the other are coherent. This research contributes to the field through the integration of BERT in understanding language and DRL in the iterative learning process in the innovation within the flaws of chatbot technologies and establishing a new benchmark for conversational AI in customer service settings.**

Keywords—Chatbots; BERT (Bidirectional Encoder Representations from Transformers); RL (Reinforcement Learning); customer service; responsiveness

I. Introduction

A chatbot is a computer program that talks to people like a human. It can use text or speech to have conversations with users. In customer service, chatbots are really important in making customer support better and more efficient. These smart systems are used on websites, mobile apps, and messaging apps to help customers quickly with their questions. Chatbots use special tools to understand what users are saying, figure out what they want, and give them the right answers. They can do many different jobs, like answering questions, giving product information, helping with problems, and doing transactions [1]. Chatbots help make customers happy by answering quickly and being available all the time. This also helps businesses work better and faster. But, chatbots need to understand what users are saying, have good conversations, and adjust to different ways people talk and different situations in order to work well. Ongoing research and improvements in NLP, machine learning, and dialogue management are very important. They help make customer service chatbots better and make sure they can talk to users without any problems [2].

Chatbots are helpful in customer service because they can quickly and accurately respond to many different questions. By using smart computer programs, chatbots can understand what users want, get important information, and quickly give the right answers [3]. This helps businesses grow their customer support without sacrificing quality or speed, so they can use their resources better and work more efficiently. Additionally, chatbots help make customers feel more connected and loyal by giving them personalized experiences based on what they like and how they act. By using data analysis and machine learning, chatbots can use past conversations and customer information to predict what customers might need, suggest helpful products or services, and solve problems before they get worse [4]. Chatbots are important for businesses because they help to keep customers happy and coming back [5]. They do this by building good relationships and having positive conversations with people. Yet, to make chatbots work well in customer service, it's important to plan carefully, keep improving them, and follow the best ways of doing things. Businesses need to spend money on the right technology, talented people, and good ways of working to make their chatbots work as well as possible. This includes making them easy for people to use, creating good conversations for them, teaching them how to learn, and keeping an eye on how well they are doing. Moreover, as customers' needs change, chatbots need to change too. They need to keep up with the latest trends, preferences, and technologies to stay useful in the changing world of customer service [6].

Being quick and accurate when talking to a chatbot is really important. It helps make sure users have a good experience and that businesses and customers get what they want. Responsiveness means the chatbot can quickly answer the questions and requests from users. In the fast-paced digital world of today, people anticipate prompt assistance and solutions for their problems. A chatbot that responds quickly not only meets these expectations but also helps to make users less frustrated, keeps them interested, and makes them feel like they can trust it [7]. In the same way, it's really important for the chatbot to give accurate and helpful responses that match what the user needs. A good chatbot understands what users mean, understands their questions in the right way, and gives the right information or takes the right actions. Giving wrong answers can make users feel confused, misinformed, and unhappy. This can make the chatbot seem less trustworthy and helpful. Additionally, giving wrong answers can make users feel frustrated and uninterested, which can make them think negatively about the brand or service [8].

Research wants to use a combination of BERT and reinforcement learning to make chatbots better. Traditional chatbots have limitations, and research want to improve how quick and accurate they are. BERT is really good at understanding language and making relevant responses. By using BERT's smart understanding of words and meanings, chatbots can be better at understanding what users want and giving them the right answers. Alternatively, reinforcement learning is a useful way to improve how chatbots manage conversations and make decisions [9]. By teaching the chatbot by trying different things and learning from user feedback, reinforcement learning helps the system improve its responses over time. This smart way of learning helps the chatbot get better and better at talking to people over time. It gets faster and more effective at having conversations in real life. By using a mix of BERT and reinforcement learning, chatbots can better understand and respond using words and also improve how they manage conversations. This mixed method makes the chatbot better at understanding and answering questions from users. It also helps the chatbot to learn and change according to what users want and how they talk. Overall, the hybrid system offers a positive solution for developing smarter and more effective chatbots that can improve user experiences in various contexts, such as customer service.

The primary contribution of the work is as follows:

- Making of a new framework of the chatbot that integrates the ability of BERT to consider a context along with the reinforcement learning to improve the quality of the chatbot’s answers with time based on the feedback from the user.
- Substantial increase in the quality and pertinence of chatbot answers, as it was proven by higher BLEU and ROUGE scores in comparison to a basic chatbot model.
- Use of the progressive learning model in which the chatbot adjusts its response to the users depending on the results obtained in the process.
- Enhanced customer satisfaction and interactivity as concomitant with presenting actual and highly and closely related responses to specific queries, thereby making customer services more fit into the natural context.
- Establishment of a general framework of a chatbot that can be easily fitted into other domains, thus, can be applicable to a wide range of client services.

The order of the remaining sections is as follows. An introduction is given in Section I. The literary sections are shown in Section II. This is the issue within the conventional methods provided in Section III. The suggested methodology for the study's research design, data gathering strategies, and analytic approaches are covered in Section IV. The efficacy measures are presented, and the results are compiled in Section V. Section VI delivers more studies and a conclusion.

II. Related Work

Dhyani and Kumar [9] explains how to use deep learning to make a better Chatbot. Implementing the Neural Machine Translation (NMT) model using the TensorFlow software library. Learning and gathering information for creating a model is a very important but challenging task. Bidirectional Recurrent Neural networks with attention layers are employed to improve responses for lengthy or wordy sentences. The information used to teach the computer model in the paper comes from Reddit. The model was made to translate English to English. This work aims to make the model more confused and learn more quickly and to measure the translation quality using Bleu Score for the same language. Research did experiments with TensorFlow using python 3. 6 The confusion, learning speed, language evaluation score and average time for every 1000 steps are 56. 10, 00001, 3016 and 45 One period is finished after taking 23,000 steps. The paper also looks at how MacBook Air can be used for neural network and deep learning. In the future, research will also make a healthcare Chatbot to help patients with diseases like COVID-19, diabetes, high blood pressure, and heart problems. by giving details about the disease, suggesting what foods to eat, and explaining how to handle emergencies.

A chatbot is a computer program that talks like a person and can have a conversation with a human. One important job in artificial intelligence and understanding human language is to study how people talk to each other. Since artificial intelligence started, it has been really difficult to make a good chatbot. Even though chatbots can do many different things, their main job is to understand what people say and reply in a helpful way. In the past, chatbot designs were made using basic statistics or written rules and templates. Around 2015, end-to-end neural networks replaced other models because they can learn better. Right now, the encoder-decoder recurrent model is very important for modeling conversations. This design comes from the field of machine translation using neural networks and it worked very effectively in that area. So far, many new features and changes have made chatbots better at talking to people. In this paper, research studied a lot of new books and articles. Research looked at a lot of articles from the past five years about chatbots. Next, research talked about other research on the topic and the AI ideas needed to make a smart chatbot using deep learning. Then, research showed a plan for creating a helpful chatbot for healthcare. In the future, research will be better at spotting and diagnosing bots, as research will have more advanced tools to help us understand their symptoms, like how intense they are, how long they last, where they are happening, and a more detailed description of what is going on [10].

AI, ML, and NLP are changing the way organizations do things. As more data comes in and AI systems get better at using it to help businesses, people are getting more excited about AI. Large amounts of data, computer power, better ways to solve problems, easy-to-use tools, and systems have made companies use AI to make their business better and make more money. These technologies help all kinds of businesses, from farming to finance. AI, ML, and NLP are helping companies with things like customer service, predicting what might happen, making things more personal for customers, recognizing pictures, understanding emotions, and working with documents online and offline. This study had two main goals. First, research look at how AI is used in business. Then research check if these uses make customers more loyal by looking at data from 910 different companies. The data includes scores for four different AI features: AI customer service, predictive modeling, personalized machine learning, and natural language processing integration. The goal is to measure how loyal customers are using a simple yes or no answer. All the qualities are rated on a scale of 1 to 5. Research used six different types of computer programs to help us learn from and make predictions about data. These were Logistic regression, KNN, SVM, Decision Tree, Random Forest, and Ada boost Classifiers. The abilities of the different algorithms were tested using confusion matrices and ROC curves. The decision tree had an accuracy of 0. 532 and KNN had an accuracy of 0. 570This study shows that businesses can use AI, ML, and NLP to look at data and find important information. This can help them automate tasks and plan business strategies. In order to stay competitive and retain customer loyalty, companies should begin incorporating them into their business strategies [11].

It is important to understand how people feel when they interact with robots as customers. This can be seen from the reviews they share online. Knowing this is important for predicting whether people will want to use service robots in the future. Qualitative analysis gives us lots of helpful information from data, but it takes lots of time and effort to do. Experts have talked about how helpful it is to use algorithms to tell the difference between different emotions. This study looks at the good and bad sides of using qualitative analysis and machine learning methods together, using both human and machine intelligence. Research took 9,707 customer reviews from two big social media sites (Ctrip and TripAdvisor). The reviews were about 412 hotels in 8 different countries. The study found that customers really like service robots, and they feel happy, amazed, and excited when they interact with them. Customers are not happy when service robots don't work and they can't use them. Robots that help people can make them feel more emotions when they move. The results also show how different cultures can affect how customers feel about service robots. The research shows that using a mix of different methods can make machine learning faster and more efficient. This aids in addressing certain issues with machine learning, such as the difficulty in comprehending concepts and the limited range of emotions available [12].

Analyzing customer feelings is very important for understanding how customers feel about products, especially in online stores where there are a lot of customer reviews. The changes in e-commerce reviews, like adding pictures, videos, and emojis, makes it more complex to analyze how people feel about products. Old-fashioned text models might have a hard time understanding feelings shown in things that are not written. This paper suggests a better way to understand how people feel about products on online stores, to make it better for shoppers. The plan involves using Fejer Kernel filtering to estimate data points in the E-commerce dataset. Research use a fuzzy dictionary to find important words in the E-Commerce dataset. The information research used for the study was collected using a method called Optimized Stimulated Annealing to pick out the most important features. Customer opinions are sorted using the BERT deep learning model. The model gives us the opinions of consumers in the E-Commerce dataset. The way customers feel about products in the E-commerce data decides how they are grouped. The test showed that the new model is better at correctly categorizing things on the online shopping site. This study helps to improve how research understand and use sentiment analysis for online shopping. It is a milestone in the development of more perceptive systems that can understand and respond to customer input [13].

III. Problem Statement

Existing customer service chatbots regularly face several limitations that the proposed development of a Responsive Customer Service Chatbot using a BERT and RL hybrid system objectives to address. Traditional chatbots commonly depend on rule-based totally or statistical fashions that may conflict with information context, dealing with complicated queries, and generating coherent responses [12]. They can also fail to evolve to evolving consumer options and remarks, ensuing in a static person enjoy [11]. Also, existing models often lack customized interplay abilities and can exhibit limited mastering from user interactions because of insufficient training data and simplistic algorithms. These chatbots normally do now not leverage superior language expertise techniques, main to suboptimal handling of nuanced language and conversational context. The proposed hybrid system, combining BERT's deep contextual embeddings with RL's adaptive learning mechanisms, goals to overcome these obstacles through improving the chatbot's capability to apprehend and generate contextually relevant responses.

IV. Proposed Responsive Customer Service Chatbot Using A BERT and Reinforcement Learning Hybrid System

The proposed method for enhancing the performance of chatbots and virtual assistants in comprehending and responding to human language involves a systematic pipeline. Firstly, data is collected, followed by pre-processing steps like tokenization, lowercasing, removing stop words, and lemmatization to clean and prepare the data. Subsequently, the processed data is fed into a BERT model, renowned for its ability to understand contextual nuances in conversations, thus providing dialogue state representation. Finally, Deep Reinforcement Learning (DRL) is employed to optimize the relevance of responses generated by the BERT model, ensuring that the chatbot or virtual assistant delivers responses that are contextually appropriate and accurate. This comprehensive approach not only improves the understanding of conversation context but also ensures the generation of responses that are highly relevant to user queries or statements, thereby enhancing the overall performance of the conversational system. This is visually presented in Fig. 1.

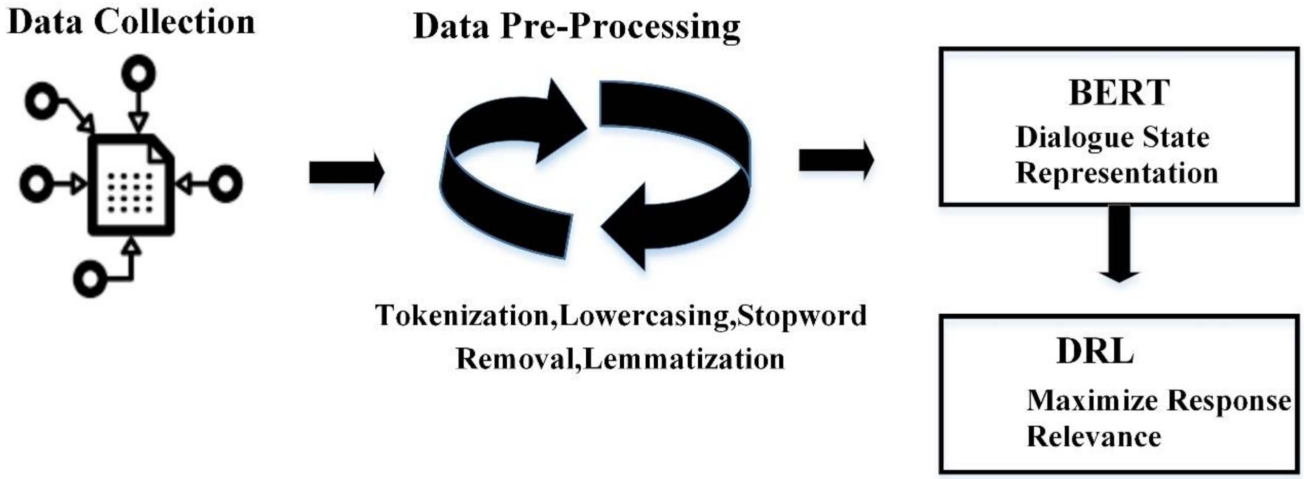

Data Collection
Data Pre-Processing
Tokenization,Lowercasing,Stopword Removal,Lemmatization
BERT
Dialogue State Representation
DRL
Maximize Response Relevance

Fig. 1. Proposed method.

A. Data Collection

Customer service chatbot data collected from Kaggle refers to a dataset sourced from the Kaggle platform, which contains conversational data typically used to train and evaluate customer service chatbots. This dataset usually includes text exchanges between customers and customer service representatives across various domains, such as retail, telecommunications, or technology, and serves as valuable input for developing and fine-tuning natural language processing models aimed at automating customer support interactions[10]. The dataset incorporates a cause tag for "good-bye," which encompasses styles like "Bye," "See you later," and "Goodbye," and is associated with responses which includes "See you later, thanks for travelling!" and "Have a nice day!" This established layout facilitates in training the chatbot to recognize and reply appropriately to one of a kind person interaction, ensuring that it could deal with common conversational exchanges effectively.

B. Pre-processing

1) Tokenization: Breaking down sentences into individual words or tokens allows for better analysis of the text. It helps in understanding the structure of the sentences and facilitates further processing steps.

2) Lowercasing: Converting all text to lowercase ensures consistency in the dataset, preventing the model from treating words with different cases as different entities. This step also reduces the vocabulary size and improves generalization.

3) Stopword Removal: Eliminating common stopwords such as "and", "the", "is", etc., helps in reducing noise in the dataset. Since these words occur frequently but often carry little semantic meaning, removing them can improve the quality of the text data.

4) Lemmatization: Reducing words to their base or dictionary form through lemmatization helps in normalizing the text and reducing dimensionality. It ensures that different forms of the same word are treated as one, thus improving the efficiency of downstream tasks like sentiment analysis or topic modelling [11].

C. BERT

In customer support chatbots, BERT's role is essential in understanding user queries and generating relevant responses. Through means of users interact with the chatbot, their queries undergo tokenization and are fed into the BERT model for natural language understanding. Bidirectional nature captures complicated word relationships, decoding the nuanced context within the user's message. This is essential user’s need, whether it involves in search of information, reporting an issue, or providing feedback. Once the user's intent is classified, the chatbot employs task-specific output layers to customize its responses. For instance, if a user seeks assistance with a technical issue, the chatbot may direct the query to a troubleshooting module trained specifically for such concerns. Similarly, if the user expresses dissatisfaction or provides feedback, the chatbot may route the query to a sentiment analysis module to assess customer sentiment and respond accordingly. Fig. 2 illustrates the framework of a pre-trained BERT model, highlighting its architecture and key components for natural language processing tasks.

Prior to implementation, the BERT model is a fine tune to domain-specific data on customer support services. This design fine-tuning ensures that the model adapts to complex customer enquiries and support interactions, improves the ability to provide accurate and helpful information and fine-tunes the model templates for tasks such as issue resolution, queries and sentiment analysis, thereby helping users maximize its effectiveness. The cross-entropy loss used during fine-tuning is represented in Eq. (1):

$$Loss\colon \sum Loss = \sum_i Y_{i\,\log(\widehat{Y_I})} \tag{1}$$

Using thorough express ratings or implicit indicators that indicate user involvement, the chatbot uses feedback to alter parameters and increase its ability to serve customers effectively. In simple terms, BERT serves as the foundation for customer service chatbots, allowing them to recognize user requests, classify intentions, generate contextually appropriate answers, and continuously enhance via comments. Utilizing BERT's simultaneous understanding and first-rate tuning capabilities, interaction with chatbots may provide personalized, effective support, enhancing the user experiences and increasing client pride is given in Eq. (2)

Feedback Loop:

$$Update\colon \theta_{BERT=}\theta_{BERT-\alpha\,\nabla\theta_{BERT}Loss} \tag{2}$$

The following formula depicts each parameter updating step in the continuous learning process, where α is the development rate and ∇θ_BERT signifies the change in gradient in relation to the BERT model parameters. Words in sentences are masked in BERT during pre-training; by predicting them, BERT learns bidirectional representations rather than reading text in a one-pass manner as most models do. Through having an interchange between the two directions, BERT is able to seize the contextual meaning of words making it very efficient in the interpretation of queries made in customer service chat bots. In the second step known as the 'fine-tuning' the BERT is trained for certain tasks such as for customer support, by feeding it with relevant data, which helps in providing appropriate responses. This, coupled with the task-specific fine-tuning, makes BERT ideal for this scenario, especially given its bidirectional context, out-competing models that do not have that or are trained in an entirely different model for an entirely different task.

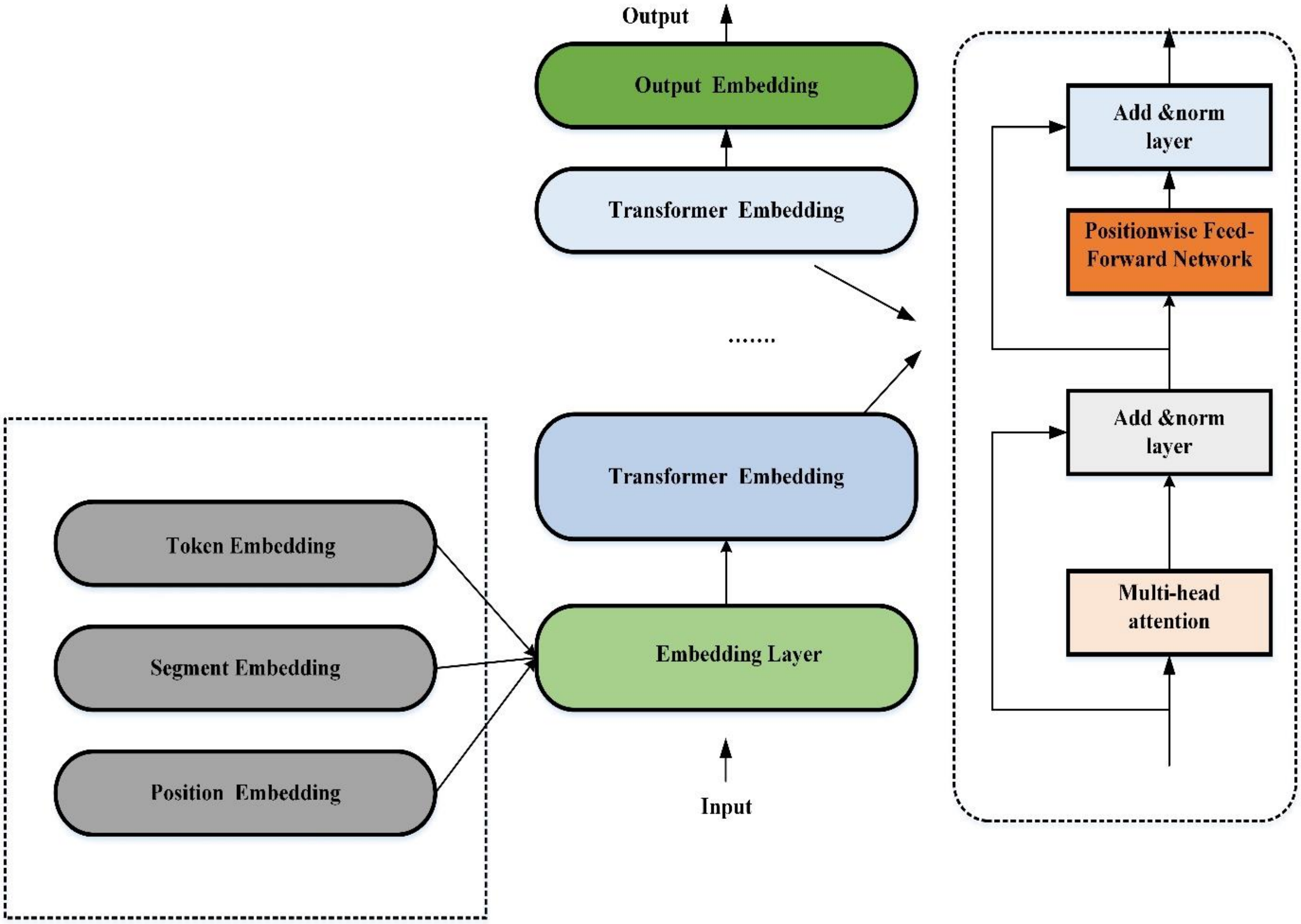


Fig. 2. Framework of a Pre-trained BERT model.

D. Deep Reinforcement Learning to Maximize Response Relevance

The development of a bidirectional contextual chatbot involves utilizing deep reinforcement learning (RL) to maximize response relevance and effectiveness. The chatbot operates within a Markov Decision Process (MDP), modeling the conversation as a reinforcement learning problem. It consists of states, actions, transition functions, and reward functions. The goal is to learn a policy that maximizes long-term rewards, ensuring responses are not only relevant to the current turn but also consider future discourse implications.

1) Markov Decision Process (MDP): A Markov Decision Process is used for modeling the discussion. MDP consists of states S, actions A, a transition function P, and a reward function R. Provided an MDP (S, A, P, and R), the algorithm is taught to identify an approach that resolves this issue. From an algorithm perspective, policy is a conditional likelihood distributed on the set of activities A. During the encounter, the agent takes action in accordance with the policy. The surrounding area changes state in response to the agent's behavior. Furthermore, the agent communicates with the surroundings at every discontinuous time step (t = 0, 1, 2, ...) is represented in Eq. (3).

$$p(s_{t+1,}r_{t+1,}|s_t,a_t) \qquad (3)$$

2) Reward Definition: The study suggests two rewards for eliciting targeted replies and achieving specified goals in conversations. It solves the problem of encoder-decoder-based models producing incoherent or insignificant outcomes. Prospective function are defined by sequential rotations and interactions among acts and prior declarations. The reward for every action is represented by r1, and the cosine similarity between actions determines the initial reward at the present state. Let h_t and h_{t+1} the graphical representations obtained from the agents for successive rounds. The cosine similarity among h_t and h_{t+1} yields a preliminary reward at present stage s_t is represented in Eq. (4).

$$r_1 = \cos(h_t, h_{t+1}) = \cos(\frac{h_t,h_{t+1}}{||h_t\, h_{t+1}||} \qquad (4)$$

3) Conversation Simulation: Although the pre-trained encoder-decoder enables the algorithm to produce coherent responses based on the discussion the past, applying RL improves the model's capacity to produce responses that are optimized for long-term objectives. To replicate the chat as follows. In the initial stage, researchers extract an input sentence via training dataset that includes conversational history as background data and give it to the initial agent. The initial agent encrypts the inputs as a vectorc_L and analyzes it to provide an outcome for the following round. The second agent modifies the simulation's condition by integrating the dialogue record and output. It instantly encodes the altered state into its visual form and interprets it to generate a new response, which is then passed again to the original agent and repeating. At the completion of the simulation, the right-context cR is a sequence of k successive utterances { $s_{t+1,}s_{t+2,} \dots\dots.. s_{t+k,}$ } to the correct portion of the produced response s_t. The change in distribution of probabilities, which represents the policy, has been initialized with a pre-trained BERT-based models. Candidate replies have been produced using the above distribution. This is represented in Eq. (5).

$$\pi = \rho_{bert2bert(a'_t}|[s'_{t,}c'_l]) \qquad (5)$$

The agent's goal is to maximize the expected accumulated reward through a series of actions (responses) during the entire discussion. To accomplish this, the parameters of the model are modified utilizing the policy gradients method. An approach for learning is employed, modeling debate for k turns and using policy gradient approaches to generate parameters that maximize the expected future reward. The purpose is to maximize the predicted cumulative value from the scheduled sequence of actions. The loss function is computed based on an anticipation for tasks and rewards in history. The upward slope of the loss function is generated using the chain rule, and this allows for adjustments to parameters to improve the model's effectiveness [12]. The MDP in the context of a Customer Service chatbot, one can refer to the following example of an interaction. Let a user ask about availability of a certain product or for some recommendations. Based on the state (the user query), the chatbot (agent) has the choice of the next action (response). For example, instead of ‘Yes, that is correct,’ it might say ‘Actually, let me check that for you.’ This response leads the conversation flow from one state to another – the user can then ask particular questions and elicit further from the chatbot. Their function is to achieve the maximum total reward that is to offer most helpful and pertinent answers. Cosine similarity is particularly important here: reward functions lie in this context. With reference to the ideal/expected response, cosine similarity makes ensures that the flow of the conversation is based on the similarity index between the chatbot and the user’s input. High cosine similarity reveals that the chatbot’s response is on the right path, the low one means that the chatbot should improve its response. While the reinforcement learning continues, the conversation patterns of the chatbot get closer to the needs of the users. For instance, at the start it may just generate general responses but it ought to adopt correct contextual replies such as: “The product is in our store at San Jose, shall I set it aside for you?” This enhancement comes from learning through feedback loops on the actions that the bot undertakes. Fig. 3 depicts the response generator utilizing deep reinforcement learning, showcasing its architecture and interaction mechanism for generating conversational responses.

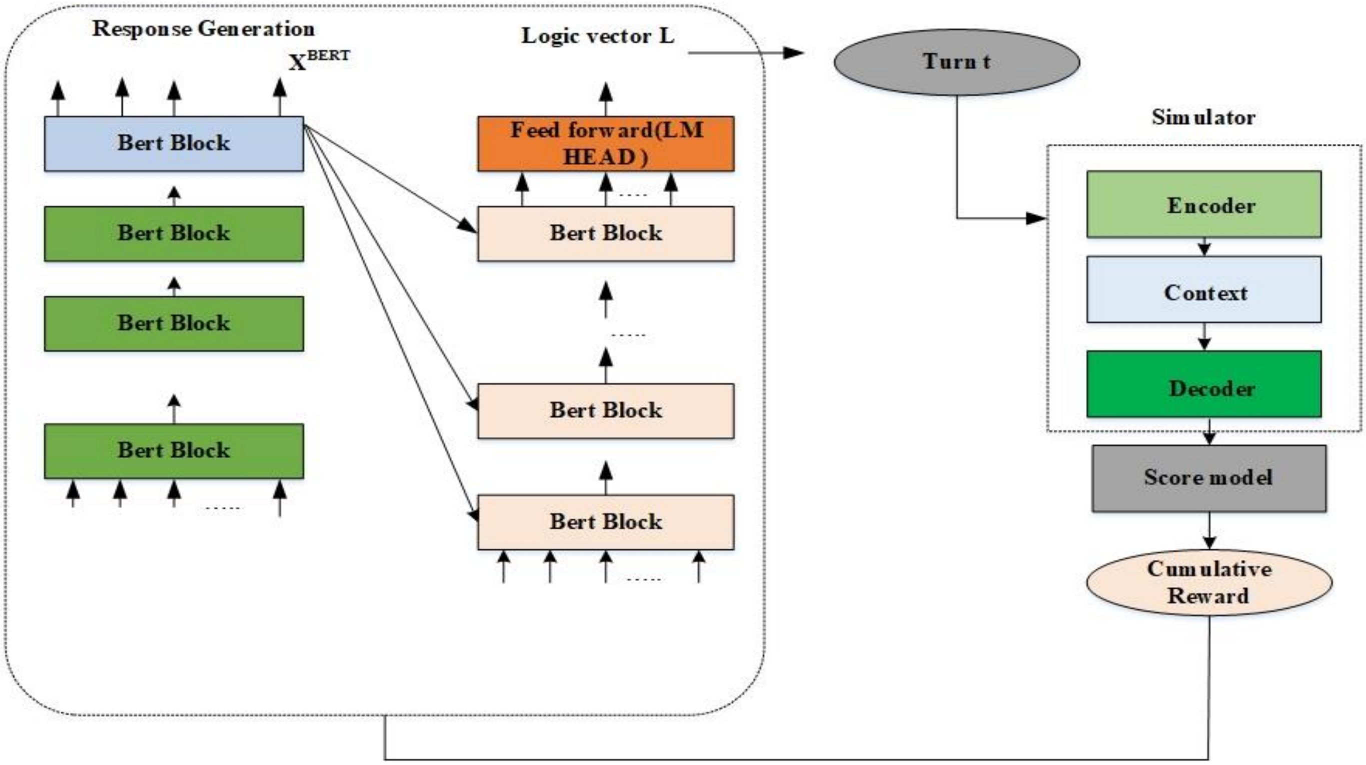


Fig. 3. Response generator using deep reinforcement learning.

V. Result and Discussion

The proposed hybrid BERT and Reinforcement Learning chatbot established huge improvements in customer service interactions. The BERT version successfully generated contextually relevant responses, whilst reinforcement getting to know optimized these responses primarily based on actual-time remarks. Evaluation metrics, which include BLEU and ROUGE ratings, showed a marked improvement in response accuracy and relevance compared to conventional fashions. User satisfaction and engagement multiplied, because the chatbot correctly addressed patron queries and tailored over time. The machine's iterative mastering technique resulted in an extra intuitive and responsive chatbot, improving ordinary consumer revel in and proving the efficacy of combining BERT with reinforcement getting to know for dynamic service interactions.

TABLE I. Comparsion of Different Deep Learning Methods

	BLEU 1	BLEU 2	BLEU 3	ROUGE Precision	ROUGE Recall	ROUGE F1 score
CNN	0.433	0.340	0.200	0.318	0.278	0.214
LSTM	0.420	0.355	0.204	0.323	0.311	0.290
GRU	0.477	0.322	0.258	0.389	0.318	0.310
BERT	0.480	0.350	0.246	0.355	0.330	0.370
BERT-RL	0.499	0.399	0.255	0.359	0.340	0.390

Table I compare the performance of different deep learning methods, including CNN, LSTM, GRU, BERT, and BERT-RL, in the context of a customer service chatbot. The metrics reported include BLEU scores (BLEU1, BLEU2, BLEU3) for measuring response similarity to human references, and ROUGE scores (ROUGEPrecision, ROUGERecall, ROUGEF1 score) for evaluating response overlap [13] [14]. Generally, BERT-based methods, especially when combined with reinforcement learning (BERT-RL), demonstrate higher BLEU and ROUGE scores compared to traditional RNN-based architectures like LSTM and GRU, as well as CNN-based models, indicating their superior performance in generating contextually relevant and accurate responses in customer service interactions. This is visually represented in Fig. 4.

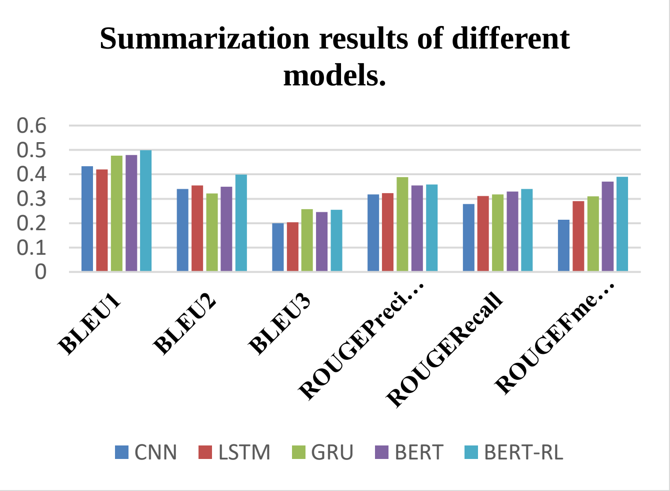


Fig. 4. Performance comparison experiments measured by BLEU score and ROUGE score.

TABLE II. MODEL PERFORMANCE WITH VARYING LENGTHS OF SIMULATED CONVERSATIONS

	BLEU1	**BLEU2**	**BLEU3**	**ROUGE Precision**	**ROUGE Recall**	**ROUGEF1- score**
BERT-RL(1 Turn)	0.470	0.370	0.320	0.350	0.340	0.354
BERT-RL(3 turn)	0.496	0.389	0.333	0.370	0.351	0.376
BERT-RL(5 turn)	0.520	0.410	0.360	0.380	0.379	0.388
BERT-RL(7 turn)	0.512	0.399	0.350	0.370	0.388	0.89

Table II presents performance metrics for a responsive customer service chatbot system utilizing a hybrid approach of BERT and reinforcement learning (RL) across different numbers of conversation turns. The BLEU scores (BLEU1, BLEU2, BLEU3) measure the similarity between the generated responses and human reference responses, with higher scores indicating better correspondence. Additionally, ROUGE scores (ROUGEPrecision, ROUGERecall, ROUGEFmeasure) evaluate the overlap between the generated and reference responses in terms of precision, recall, and F1 score. As the number of conversations increases, generally, the BLEU and ROUGE scores tend to improve, suggesting that the chatbot's performance benefits from a deeper context and longer interactions. BLEU and ROUGE ratings are critical for evaluating a chatbot's overall performance. Fig. 5 presents BLEU measures the similarity among the chatbot's generated responses and reference responses, specializing in precision. ROUGE emphasizes recall, taking pictures how well the chatbot's responses cover key aspects of the expected solutions. Together, they make sure balanced assessment of accuracy and relevance.

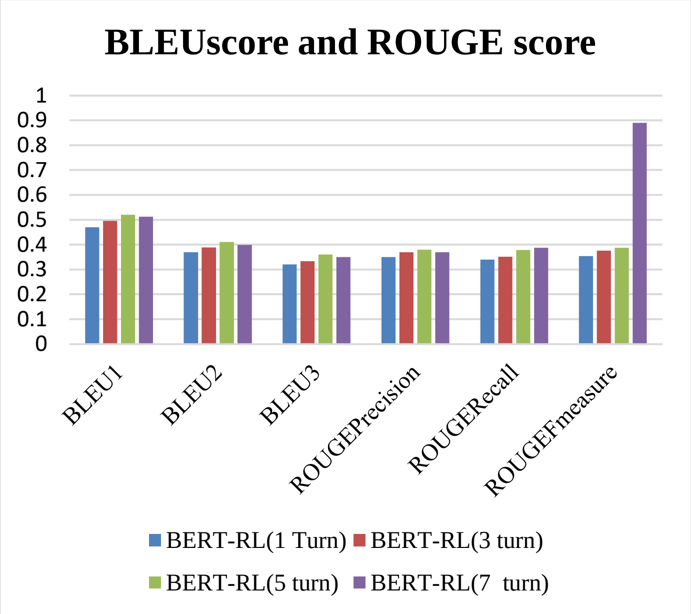


Fig. 5. Chart displaying performance with various lengths of simulated conversation measured.

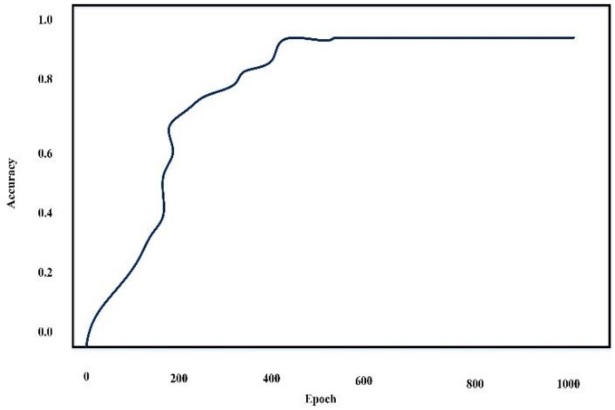


Fig. 6. Overall accuracy.

The total accuracy of the Customer Service Chatbot built with a BERT and Reinforcement Learning Hybrid System is a comprehensive measure of its ability to understand customer inquiries and respond appropriately this is visually represented in Fig. 6. By combining BERT's sophisticated natural language processing capabilities with reinforcement learning's iterative learning strategy, the chatbot obtains high accuracy in understanding user intent and providing contextual suitable solutions. This hybrid system adapts through user interactions and refines its responses continuously to improve accuracy, resulting in effective and beneficial customer support interactions.

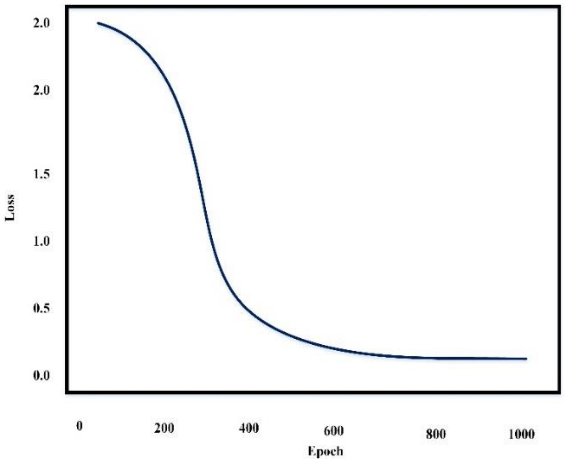


Fig. 7. Overall accuracy.

The overall loss in the context of a Customer Service Chatbot using a BERT and Reinforcement Learning Hybrid System represents a measure of the discrepancy between the predicted responses generated by the chatbot and the ground truth responses provided by humans is shown in Fig. 7 This loss function quantifies the model's performance during training, guiding it towards minimizing errors and improving response quality. By minimizing the overall loss, the chatbot learns to generate more accurate and contextually relevant responses, thereby enhancing its effectiveness in addressing user queries and improving overall customer satisfaction.

A. Discussion

The conceived work will thus aim at training a Responsive Customer Service Chatbot by applying a hybrid model involving the well-established BERT of language and RL. The idea of this integration is to address a number of drawbacks of the current chatbot systems and improve their outcomes vastly. Most prior chatbots have static decision rules or statistical models, which create limited, contextually superficial conversations that do not extend very well to users individually, or in the shifting trends that correspond to the further course of a conversation. It is also possible with the offer of the proposed chatbot that uses BERT for its functioning since the latter organized language best in paying attention to context and juggling its intricacies. The integration of RL solves the issue of rigidity, which is usually characteristic of normal chatbots [13] [14]. RL enables the chatbot to get smarter by learning from the users' response and feedback thus making it to increase its performance. It is such constant training that assists in improvement of the different and intricate questions that the chatbot is likely to encounter and thus improves the general user experience. Also, the hybrid system is expected to address some issues that may include the overall flow and subject coherence in many turns of the conversation, improved recognition of the user intent, and learning new patterns of the conversation. Doing so benefits from using BERT's deep contextual embeddings combined with RL's adaptive learning mechanisms as more substantive and useful chats are expected to be attained.

VI. Conclusion and Future Work

The improvement of the Responsive Customer Service Chatbot the use of a hybrid BERT and Reinforcement Learning (RL) machine represents a widespread advancement in chatbot technology. This technique combines the strong language understanding competencies of BERT with the adaptive learning strengths of RL, addressing key obstacles in present day chatbot systems. Traditional chatbots regularly battle with keeping conversational context, knowledge complex person queries, and adapting to dynamic person wishes. By integrating BERT, which excels in processing and decoding nuanced language, with RL, which allows for continuous development based totally on user interactions, the proposed machine aims to deliver a greater clever, responsive, and contextually conscious chatbot. The initial results suggest promising improvements in reaction accuracy and consumer engagement. The BERT model enhances the chatbot's potential to comprehend the intricacies of language, whilst RL best-tunes its responses based on actual-time comments, leading to a greater personalized and effective consumer revel in. This hybrid approach overcomes the stress of rule-based totally structures and the limitations of conventional statistical fashions, imparting a dynamic and scalable solution for customer support applications. Looking forward, future work will consciousness on several key regions. Firstly, expanding the chatbot's competencies to handle greater numerous and complex communication topics could be essential. This entails incorporating additional area-unique know-how and refining the RL algorithms to higher manage problematic user interactions. Secondly, improving the machine's potential to deal with multi-flip conversations and emotional nuances will improve normal person delight. Finally, exploring integration with different superior technology, which includes voice popularity and sentiment evaluation, will similarly raise the chatbot's capability and user experience. Ongoing evaluation and iteration will be critical to ensure the system stays powerful and adaptable to evolving user desires and technological improvements. Future work for the research focus on expand the chatbot's ability to handle a broader range of topics by integrating additional domain-specific knowledge bases and refining the reinforcement learning algorithms for better handling of complex queries. Developing advanced techniques to manage multi-turn conversations, allowing the chatbot to maintain context over extended interactions and respond appropriately to user inquiries that involve emotional nuances.

References

[1] S. Ayanouz, B. A. Abdelhakim, and M. Benhmed, "A smart chatbot architecture based NLP and machine learning for health care assistance," in Proceedings of the 3rd international conference on networking, information systems & security, 2020, pp. 1–6.

[2] C. Magoo and M. Singh, "Design and development of new interactive chatbot system for mobile service providers through heuristic-based ensemble learning," Cybernetics and Systems, vol. 55, no. 4, pp. 753–785, 2024.

[3] Z. Wu, Q. She, and C. Zhou, "Intelligent Customer Service System Optimization Based on Artificial Intelligence," Journal of Organizational and End User Computing (JOEUC), vol. 36, no. 1, pp. 1–27, 2024.

[4] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing Turkish sentiment analysis models using machine learning and e-commerce data," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 202–207, 2021.

[5] K. Badran, "Using ChatGPT to Augment Software Engineering Chatbots Datasets".

[6] A.-C. Le, V.-N. Huynh, and others, "Enhancing Conversational Model with Deep Reinforcement Learning and Adversarial Learning," IEEE Access, 2023.

[7] A. Shoufan and S. Alameri, "Natural Language Processing for Dialectical Arabic: A Survey," in Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China: Association for Computational Linguistics, 2015, pp. 36–48. doi: 10.18653/v1/W15-3205.

[8] M. S. Ali, M. W. Anwar, F. Azam, and M. H. Ashraf, "Intelligent Agents in Educational Institutions: AEdBOT–A Chatbot for Administrative Assistance using Deep Learning Hybrid Model Approach," 2024.

[9] A. Al Sallab, H. Hajj, G. Badaro, R. Baly, W. El Hajj, and K. Bashir Shaban, "Deep Learning Models for Sentiment Analysis in Arabic," in Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China: Association for Computational Linguistics, 2015, pp. 9–17. doi: 10.18653/v1/W15-3202.

[10] "Customer Service Chatbot-Data." Accessed: May 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/ngawangchoeda/customer-service-chatbotdata

[11] N. Bhartiya, N. Jangid, S. Jannu, P. Shukla, and R. Chapaneri, “Artificial neural network based university chatbot system,” in 2019 IEEE Bombay Section Signature Conference (IBSSC), IEEE, 2019, pp. 1–6.

[12] Q.-D. L. Tran and A.-C. Le, “Exploring bi-directional context for improved chatbot response generation using deep reinforcement learning,” Applied Sciences, vol. 13, no. 8, p. 5041, 2023.

[13] Q.-D. L. Tran and A.-C. Le, “Exploring Bi-Directional Context for Improved Chatbot Response Generation Using Deep Reinforcement Learning,” Applied Sciences, vol. 13, no. 8, Art. no. 8, Jan. 2023, doi: 10.3390/app13085041.

[14] H. Palivela, “Optimization of paraphrase generation and identification using language models in natural language processing,” International Journal of Information Management Data Insights, vol. 1, no. 2, p. 100025, 2021.

Liver and Tumour Segmentation Using Anchor Free Mechanism-Based Mask Region Convolutional Neural Network

(Liver and Tumour Segmentation)

Sangi Narasimhulu, Ch D V Subba Rao
Department of Computer Science and Engineering,
Sri Venkateswara University College of Engineering, Tirupathi, India

Abstract—**An accurate liver tumour segmentation helps acquire the measurable biomarkers for decision support systems and Computer-Aided Diagnosis (CAD). However, most existing approaches fail to effectively segment tumours in the liver due to the overlapping of liver with any other organ in the image. To solve this problem, this research proposes Anchor Free with Masked Region-based Convolutional Neural Network (AFMRCNN) approach for segmenting liver tumours. The AF attains a precise localization of tumours by directly predicting the tumour location without relying on predefined anchor boxes. Standard datasets like LiTS and CHAOS are utilized to experiment with the efficiency of the proposed method. An EfficientNetB2 is performed to extract the most relevant features from the segmented data. The Deep Neural Network (DNN) is performed for the classification of liver tumours into binary classes by capturing intricate patterns and relationships in the data with the help of a non-linear activation function. The experimental results exhibit the proposed ARMRCNN method's commendable segmentation performance of 0.998 Dice Similarity Coefficient (DSC), as opposed to the existing methods, UoloNet and UNet++ + pre-activated multiscale Res2Net approach with Channel-wise Attention (PARCA) on the LiTS dataset.**

Keywords—Anchor free; computer-aided diagnosis; deep neural network; EfficientNetB2; liver and tumor segmentation; masked region-based convolutional neural network

I. Introduction

Medical segmentation plays a vigorous role in Computer-Aided Diagnosis (CAD) by effectively improving the diagnostic performance and accuracy. This process enhances the precision of diagnosis, allowing more accurate identification and analysis of medical conditions [1], [2]. Globally, the liver disease is considered as the deadly disease and it is the predominant causes for liver cancer mortality. Liver tumour segmentation is important in the tumour phase and is a primary requirement for various radiological and surgical interventions including ablation therapy, liver transplant, etc. [3], [4]. There are various diagnosis tools like Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), etc. which are the most extensively utilized techniques for the detection and diagnosis of hepatic cancer [5]. Among these, CT scans are majorly utilized to diagnose and provide highly detailed images of the body's internal structures and soft tissues. This high level details help accurately diagnosing various conditions. Different CAD solutions have been examined to aid radiologists in decision-making with diagnostic effectiveness. Liver segmentation is the most complex phase of CAD systems and hence plays a pivotal in identifying the success of diagnosis [6]. This process allows doctors to identify tumours with similar appearances in medical images and assists in developing tailored treatment pathways [7]. The reliability and precision of the segmentation approaches are significant for acquiring the clinically relevant boundary, as well as the volumetric calculations in the stage of liver tumour [8].

The structural data of shape, size and location are obtained from the segmented liver areas which offer a helpful understanding for disease assessment and treatment planning [9],[10]. Thus, introducing the automatic and accurate approaches for liver tumour segmentation has fascinated to an enhancing consideration with a crucial worth in clinical practice [11], [12]. Additionally, researchers have introduced various CAD and Deep Learning (DL) approaches to aid radiologists in understanding the CT images [13],[14]. Simultaneously, the CAD systems detect Regions of Interest (RoI) and provide the probability of these areas being specific types of lesions of either malignant or benign [15]. However, most of the existing approaches have failed to effectively segment tumours in the liver due to the overlapping of the liver with any other organ in the image. To overcome this problem, this research proposes the novel DL approach of Anchor Free with Masked Region-based Convolutional Neural Network (AFMRCNN) approach for segmenting liver tumours. The foremost contributions of this research are as follows:

- For liver tumour segmentation, this research proposes the AFMECNN approach. The AF directly detects the tumour location without relying on predefined anchor boxes, leading to the minimization of the existence of false positives and negatives.
- For the feature extraction process, the pre-trained architecture of EfficientNetB2 is utilized to extract the pertinent features. The EfficientNetB2 technique attains the most efficient extraction of high-quality features from images by employing effective scaling approaches.

- The Deep Neural Network (DNN) is performed to classify liver portions into two categories: Tumour and Non-tumour. The DNN attains a greater accuracy by capturing intricate patterns and relationships in data with the help of non-linear activation functions.

This research paper is further provided as follows: Section II displays the literature survey based on liver and tumour segmentation using DL approaches, Section III shows the proposed methodology, while Section IV demonstrates the results and discussion, and the conclusion of this research is given in Section V.

II. Literature Survey

In order to address the issue of liver tumour segmentation, various approaches were introduced by researchers. This section discusses the related works of the DL-based approaches for liver and tumour segmentation, along with their advantages and limitations. Table I demonstrates the advantages and limitations of the literature survey of Liver and Tumor Segmentation discussed in this research.

Zheng et al. [16] developed a UoloNet model to enhance the small target medical segmentation model for liver tumours using the LiTs17 dataset. The UoloNet model comprised three main modules namely, shared encoder module, object detection module, and mask generating module. The share encoder-decoder module generally implemented the extracting features of the image. The object detection module performed in a dual task mode with object detection and segmentation. The prediction module enhanced the segmentation accuracy by utilizing detection outputs to improve and highlight specific regions. Finally, the Intersection Over Union (IOU) metric was used in the traditional YOLO approach to increase the convergence speed of the network. However, the high amount of noise in the prediction module affected the identification of small labels in the image.

Huang et al. [17] implemented a Semi-supervised Double-cooperative Network (SD-Net) for liver tumour segmentation using a CT scan image. The SD-Net framework reduced the requirement of dense labelling in liver image segmentation. For the segmentation process, collaborative networks like V-net and 3D-ResVnet were utilized to transfer the labelled image from to the unlabelled image in the target domain. A dynamic Pseudo-label generation strategy was introduced to enhance the label qualities in unsupervised learning by collecting better-predicted masks as pseudo-labels from both network models. Nonetheless, the implemented SD-net model was memory intensive due to the high resolution of images, leading to limited scalability.

Yu et al. [18] presented a liver tumour segmentation network utilizing multi-phase CT images. To increase the importance of reciprocal data from various stages, Cross-Modal feature Guidance (CMG) and multi-feature fusion modules were developed in the model. The two modules were combined to effectively acquire the multi-phase features and improve the performance of liver tumour segmentation. The DL architecture was designed to exchange information between multiple phases and modalities accurately. However, the presented CMG model was constrained in its capability to analyse the complex details of lesions and generally focused only on segment lesions from inaccurate registration.

Kushnure et al. [19] implemented a new lightweight multi-level network through core architecture from the UNet++ network for automatic liver segmentation. The designed Pre-activated multiscale Res2Net approach with Channel-wise Attention (PARCA) block in a network model was able to extract coarse multi-scale features from various levels. The modification in the UNet++ network with a PARCA feature mechanism extracted more semantic and contextual data, thereby enhancing the performance of the decoder in the network. In addition, a customized loss function was implemented to maintain data imbalance and effectively consider complex samples in the collected data. Nevertheless, the UNet++ failed to effectively segment tumours in the liver that overlapped with other tissues.

Manjunath and Kwadiki [20] introduced a DL method for automatic liver tumour segmentation through CT scan images. For the liver tumour segmentation process, a 2D-modified ResUNet was designed to segment the affected regions in CT scan images. The 2D-modified ResUNet network was constructed by utilizing a parts encoder, the decoder, and the bridge. The encoder in the network encoded the resized image to a compact representation, then a pixel-wise fashion of an image was represented by the decoder, and finally, the bridge connected the paths of the encoder and decoder. Yet, the 2D-modified ResUNet network model was unable to identify and segment small tumours in the image.

TABLE I. Literature Survey of Liver and Tumour Segmentation

Author	Advantages	Limitations
Zheng et al. [16]	UoloNet's CNN component significantly model spatial relationships within CT images, helping it identify subtle variations in tissue texture and intensity of a small tumour.	A high level of noise in the segmented module impacted the accurate identification of small labels in the image.
Huang et al. [17]	SD-Net leveraged both labelled and unlabelled data by semi-supervised learning approach, enabled for more accurate segmentation of small amount of labelled data.	SD-net model was memory intensive because of its high resolution of images, resulted in limited scalability.
Yu et al. [18]	Cross-modal guidance mechanism enabled the network to significantly learn from balancing data over various stages, improved the segmentation performance.	CMG model was limited in its capability to determine the complex details of lesions and basically focused only on segment lesions from inaccurate registration.
Kushnure et al. [19]	Through the fusion of features from various levels, the network integrated the global and local information, enhanced the segmentation accuracy.	UNet++ was failed to significantly segment tumours in the liver that overlapped with other organs, resulted in poor performance.
Manjunath and Kwadiki [20]	The end-to-end approach removed the necessities of intermediate feature extraction steps and potentially enhanced the overall performance.	2D-modified ResUNet network model was unable to identify and segment small tumours in the image.

From this overall analysis, the limitations of the existing methods are identified as follows: the presence of noise in the prediction module, memory intensive, limited detail analysis and ineffective tumour segmentation. To solve those aforementioned problems, this research proposes the AFMRCNN approach for the effective segmentation of liver tumours. The ARMRCNN approach solves the memory-intensive problem by removing anchor-based complexity by considering effective sparse predictions.

III. Proposed Methodology

This research aims to overcome the segmentation in overlapping images. The AFMRCNN approach is proposed to effectively segment the liver tumour regions from the input images. The AF mechanism frequently utilizes key point-based predictions, provides more accurate boundary information and leads to enhanced segmentation performance. The DNN is used for the classification of segmented regions based on learned patterns by assigning the class labels to various regions in the image. The proposed method is comprised of five significant phases: data acquisition, pre-processing, liver tumour segmentation, feature extraction and classification. Fig. 1 depicts the pipeline of the proposed method.

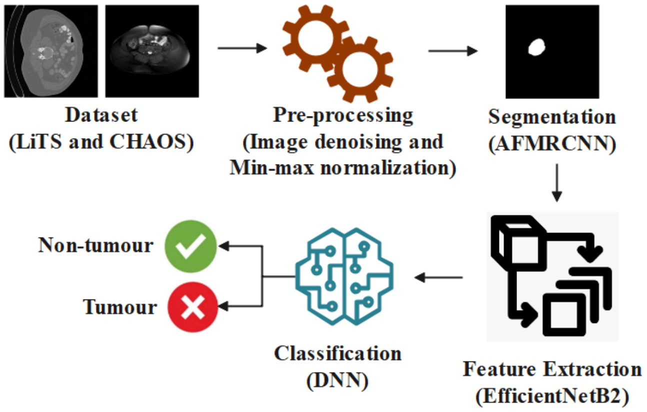


Fig. 1. Pipeline of the proposed method

A. Dataset Collection

This research estimates the efficacy of the proposed method with two publicly available standard liver segmentation datasets, Liver Tumour Segmentation (LiTS) dataset [21] and Combined (CT-MRI) Healthy Abdominal Organ Segmentation (CHAOS) dataset [22]. The detailed explanation of these datasets is discussed below.

1) LiTS dataset: The LiTS dataset involves 201-contrast enhanced 3D abdominal CT images, where 194 CT include lesions. This dataset is obtained from various scanners with scanning protocols from the clinical areas. The minimum and maximum number of axial slices in CT scans are 74 and 987, respectively. Fig. 2 depicts the sample images of the LiTS dataset.

2) CHAOS dataset: CHAOS involves 120 DICOM volumes obtained from binary MRI modalities: T1-DUAL and T2-SPIR. T1-Dual involves phases which are, in-phase and out-of-phase, each of 40 volumes. Every modality involves 20 labelled training data and 20 unlabelled test data. The labelled training data are utilized for multi-modal medical image segmentation tests and estimations. Fig. 3 depicts the sample images of the CHAOS dataset.

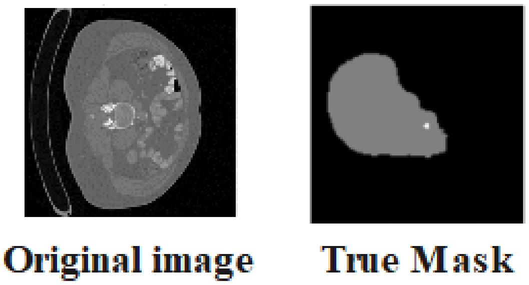


Fig. 2. Sample images of the LiTS dataset.

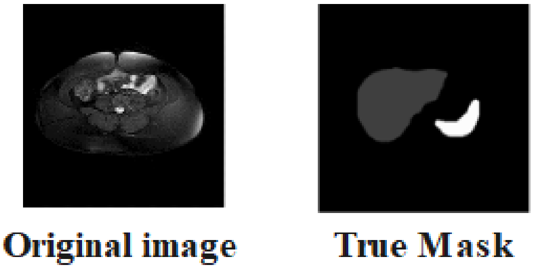


Fig. 3. Sample images of the CHAOS dataset.

Then, the collected datasets are portioned into two categories: 80% of the data are utilized for training and the remaining data for testing. Table II represents the sample number of datasets. The collected dataset is then provided for the pre-processing step.

TABLE II. Sample Number of Datasets

Dataset	Tumour	Non-Tumour
LiTS	21478	42160
CHAOS	6217	8205

B. Pre-processing

After the data collection, pre-processing is performed to modify the input data for achieving the desired results of the DL approach. The collected data contains problems of high resolution and unwanted noise which leads to inaccurate segmentation results. Hence, this research utilizes pre-processing techniques of image denoising and min-max normalization. In the image denoising process, the Gabor filtering approach [23] is utilized and is realized by performing convolution on the Gaussian function with trigonometric functions. By choosing the suitable Gabor function, various scales and directional features are detected from the collected data. This allows the utilization of Gabor filtering in image denoising and edge detection applications. By using image denoising, the Gaussian noise is removed as it produces random variations in pixel values, which obscure the significant features and structures in medical images. This denoising supports to enhance the precision of segmentation algorithms.

After denoising the image, the normalization is performed to equally balance the input data to enhance the segmentation performance because the collected data involves various units and scales. The min-max normalization [12] is a technique that is used to standardize and measure data in the direction to be taken out to the comparable range and magnitude. Then, the pre-processed data are provided to the segmentation process to effectively localize the affected regions.

C. Segmentation

The pre-processed images are fed as input to the segmentation technique to effectively segment the liver tumour regions for classification. Segmentation is a technique generally used in image processing to segment into multiple parts or regions based on the characteristics of pixels in the image. To effectively segment the tumour regions, the AFMRCNN approach is proposed. The detailed explanation of this proposed method is described in the following section. Fig. 4 illustrates the segmented images of LiTS and CHAOS datasets.

1) Anchor-free with mask region-based convolutional neural network: As compared with the other segmentation approaches, the MRCNN approach produces high-quality pixel-to-pixel masks for every case. It performs pixel-level segmentation operations and obtains better target-positioning performance in liver tumour segmentation. The MRCNN [24] approach utilizes a two-stage network model. Initially, the Region Proposal Network (RPN) generates predictions on Regions of Interest (RoI). Then, the Fully Convolutional Network (FCN) processes these RoIs to predict the binary mask, bounding box offsets, and categories for every RoIs. The network model of the MRCNN approach involves three significant types of backbone network, pixel (mask) prediction and alignment of RoI.

In the process of liver tumour segmentation, this research designs the Anchor Free with RPN (AF-RPN) approach to acquire better localization of the liver tumours. The AF-RPN architecture is an FCN to achieve the sharing computation with the binary class classification network. To acquire the tumour regions, the AF-RPN uses a sliding window mechanism applied to feature maps developed through a fusion module. This network performs 3×3 spatial window on an input feature map, and extracts the features through every sliding window. It is mapped to 512-dimensional feature vector through 3×3 kernel convolutional operation via 512 channels. These channel features are forwarded to the two parallel fully convolutional networks. One branch provides likelihoods for classification, demonstrating whether or not the regions are objects, while the other branch provides the coordinates of the boxes for localization. These are processed through the 1×1 convolutional layer and eventually, all output proposals are performed through the Non-Maximum Suppression (NMS) approaches to eliminate irrelevant proposals. Here, the Receptive Field (RF) is developed and anchor-free boxes are applied within the proposed region network to acquire parameterization for position boxes. The RF is introduced by sliding the window across fusion feature maps in CNN.

The RF for pixel is described as the rectangle region in the input data. Particularly for $k \times k$ sliding window with centroid (x, y), the RF is obtained with (x_r, y_r, w_r, h_r), where, $x_r = v \times x$ and $y_r = v \times y$ and $w_r = h_r = k \times x$ where v represents an upsampling factor of the scaling coefficient from feature map to the input data. At every feature map location, there are k anchor-free boxes in the convolutional approach, but the proposed approach involves the individual RF. Providing the convolutional feature map of size $w \times h$, obtaining the $w \times h$ RF samples in total leads to k times, not more than the present AF mechanism. In this research, the RF is introduced as the position boxes to acquire parameterized coordinates of the ground truth bounding box $t *\left(t *= \left\{t_x *, t_y *, t_w *, t_h *\right\}\right)$ and predicted box $\left(t = \left\{t_x, t_y, t_w, t_h\right\}\right)$, as formulated in Eq. (1) and (2).

$$t *= \frac{(x*-x_r)}{w_r}, t_y *= \frac{(y*-y_r)}{h_r}, t_w *= \frac{log(w*/w_r)}{t_h}, t_h *= log(h*/h_r) \tag{1}$$

$$t_x = \frac{(x-x_r)}{w_r}, t_y = \frac{(y-y_r)}{h_r}, t_w = log(w/w_r), t_h = log(h*/h_r) \tag{2}$$

Where, $x*$ and $y*$ represent the centre coordinates of ground truth box; $w*$, $h*$ denote the width and height, x, y, w and h illustrates the complements of predicted box, x_r, y_r, w_r and h_r represent the RF box which is observed as the regression from the RF to its closest ground truth box. By eliminating the anchor-based predictions, this approach simplifies the prediction process, which supports the more effective maintenance of memory when processing high-resolution images. Then, the segmented results are provided to the feature extraction process.

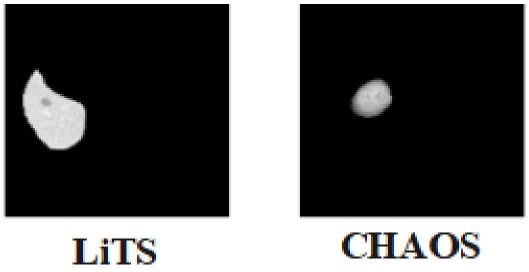


Fig. 4. Segmented images of LiTS and CHAOS datasets.

D. Feature Extraction

The segmented portions of the liver tumour are fed to the feature extraction process. Feature extraction is a process of identifying and extracting relevant features from the data, further used in classification and prediction tasks. Extracting the important features of shape, edge, texture and intensity information support the distinguishing between normal liver tissue and tumour tissue. This leads to more accurate segmentation results and outlines tumours from the surrounding healthy tissue. In this research, the pre-trained architecture of EfficientNetB2 is performed to extract the relevant (texture and shape) features. The detailed explanations of this architecture are discussed below.

1) EfficientNetB2: As compared to other pre-trained architectures, the EfficientB2 utilizes a compound scaling approach to balance the network resolution, depth and width. This process effectively scales dimensions, leading to the most effective utilization of floating-point operations and parameters. The network depth, resolution and width are reliably scaled through the Efficient Net logically by the utilization of compound coefficient. The EfficientNetB2 [25] involves arranging the model through global max pooling and the dropout layer to solve the overfitting problem and succeed through the dense layer for the binary classification. This architecture complies with the suitable loss as well as optimization functions and performs the callbacks for effective training. In EfficientNetB2, the CNN network involves various layers designed for the image processing tasks. With the output shape, the EfficientNetB2 incorporates the 7×7 spatial dimensions with 1408 channels. Ensuing the convolutional layers, the Global Average Pooling (GAP) operation minimizes the spatial dimensions while recollecting the most salient features, resulting in the size of (None, 1408). Then, the dropout layer is performed to alleviate overfitting by arbitrarily deactivating the neurons in training. Eventually, the dense layer with 1 unit is used for the extraction process with 1409 parameters. The extracted texture and shape features using EfficientNetB2 provide detailed information about tumour regions. This information supports classification by helping differentiate tumour regions from normal liver tissue and other structures. Then, the extracted 1408 features from the GAP layer are provided for the classification process.

E. Classification

After feature extraction, the features are fed as input to the liver tumour classification. In this research, the DNN approach is utilized for the classification of tumours into two types as tumour and non-tumour. The detailed description of the DNN is discussed below.

1) Deep neural network: The DNN [26] is inspired by the biological nervous network. The DNN is a robust promising approach in modelling the mechanical materials behaviour because of its powerful nonlinear mapping capability. The DNN approach comprises three significant layers of input, activation and Fully Connected (FC) layer which are utilized for the classification tasks.

A neural network requires an activation function for the final prediction of liver tumours. Rectified Linear Unit (ReLU) is the default activation function; it extends the nonlinearity to the network which provides output 0 for negative values and similar values for non-negative values. Also, Sigmoid is an activation function which is appropriate for binary classification with the output within the range of 0 to 1. The sigmoid identifies the multinomial probability distribution with two classes. The dense layer is the Neural Network (NN) layer, in which every neuron in the layer obtains an input from whole neurons of its previous layer. It utilizes the operation function to map each input with output.

In the classification process, the DNN effectively classifies the liver tumour into binary classes with the help of various layers. Hence, DNN is more robust to variations in input data like differences in tumour size, shape, and appearance, as opposed to the traditional classifiers which require effective tuning of parameters. The classified result are implanted to estimate the effectiveness of the model, and the detailed explanation result is represented in the following section.

IV. Experimental Results

The segmentation effectiveness of the liver tumour is estimated experimentally based on two standard datasets. The experiments of the proposed AFMRCNN approach are implemented on Python 3.10.12 with the system configuration of Windows 10 (64 bit) OS, intel i5 processor and 8GB RAM. The model effectiveness is estimated through various segmentation performance metrics of Volumetric Overlap Error (VOE), Dice Similarity Coefficient (DSC), Relative Volume Difference (RVD) and Intersection over Union (IoU). The classification metrices of accuracy, precision, specificity, sensitivity/recall, and F1-score are used to estimate AFMRCNN performance. The mathematical expressions of these assessment metrics are formulated in Eq. (3) to (10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

$$F1-score = \frac{2TP}{2TP+FP+FN} \tag{7}$$

$$DSC = 2\frac{|A \cap B|}{|A|+|B|} \tag{8}$$

$$VOE = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

$$RVD = \frac{|B|-|A|}{|A|} \tag{10}$$

Where, TN is True Negative, TP is True Positive, FN is False Negative and FP is False positive, while A and B represent the binary masks. Table III represents the hyper parameter settings of the proposed AFMRCNN approach.

TABLE III. Hyperparameter Settings of the Proposed AFMRCNN Approach

Hyper parameters	Values
Optimizer	Adam
Learning Rate	0.0001
Loss Function	Binary cross entropy
Batch size	32
Activation function	Sigmoid
No. of Epochs	20

A. Quantitative and Qualitative Analysis

The achievements of the proposed AFMRCNN method is estimated against the existing methods on the LiTS and CHAOS datasets. Table III demonstrates the analysis of the segmentation results. Table IV represents an analysis of the feature extraction results with an analysis of the classification results.

In Table IV, the performance analysis of the segmentation results is presented on the LiTS and CHAOS datasets. The success of the proposed AFMRCNN method is estimated and compared with conventional segmentation approaches like UNet, CNN and MRCNN approaches. When compared to these conventional approaches, the AF in the MRCNN approach that depends on the predicting key points, leading to more accurate localization of liver tumours. This approach specifically aids when dealing with irregular shapes and sizes of tumours, leading to enhanced segmentation outcomes. In LiTS dataset, the proposed AFMRCNN approach attains a better DSC of 0.978, VOE of 14.75, RVD of 11.92 and IoU of 0.98. In the CHAOS dataset, the proposed AFMRCNN approach attains a superior DSC of 0.988, VOE of 6.52, RVD of 5.32 and IoU of 0.98.

In Table V, the performance analysis of the feature extraction results is demonstrated based on the LiTS and CHAOS datasets. The efficiency of the EfficientNetB2 approach is estimated and compared with existing feature extraction approaches of ResNet, VGG19 and InceptionNet. As opposed to these approaches, EfficientNetB2 architecture permits the model to extract relevant and detailed features, capturing complex patterns and structures in the collected data with the help of the compound scaling method. In the LiTS dataset, the performed EfficientNetB2 approach attains a better accuracy of 0.997, precision of 0.998, recall of 0.978, specificity of 0.995 and F1-score of 0.987. Also on the CHAOS dataset, the EfficientNetB2 attains a better accuracy of 0.996, precision of 0.997 precision, recall of 0.988, specificity of 0.996 and F1-score of 0.992.

In Table VI, the performance analysis of the classification results is demonstrated based on the LiTS and CHAOS datasets. The effectiveness of the DNN approach is estimated and compared with the existing feature extraction methods of CNN, Simple Neural Network (SNN) and Feedforward Neural Network (FNN). When compared to these approaches, the DNN minimizes the loss of important features by allowing end-to-end learning. In the LiTS dataset, the performed DNN approach attains a superior accuracy of 0.997, precision of 0.998, recall of 0.978, specificity of 0.995 and F1-score of 0.987. Additionally on the CHAOS dataset, the DNN attains a superior accuracy of 0.996, precision of 0.997, recall of 0.988, specificity of 0.996 and F1-score of 0.992.

TABLE IV. ANALYSIS OF SEGMENTATION RESULTS

Dataset	Method	DSC	VOE	RVD	IoU
LiTS	UNet	0.874	19.80	15.45	0.90
	CNN	0.896	18.90	14.75	0.91
	MRCNN	0.912	17.65	13.80	0.93
	AFMRCNN	0.978	14.75	11.92	0.98
CHAOS	UNet	0.853	8.45	6.75	0.85
	CNN	0.875	8.60	7.85	0.86
	MRCNN	0.893	7.95	6.45	0.88
	AFMRCNN	0.988	6.52	5.32	0.98

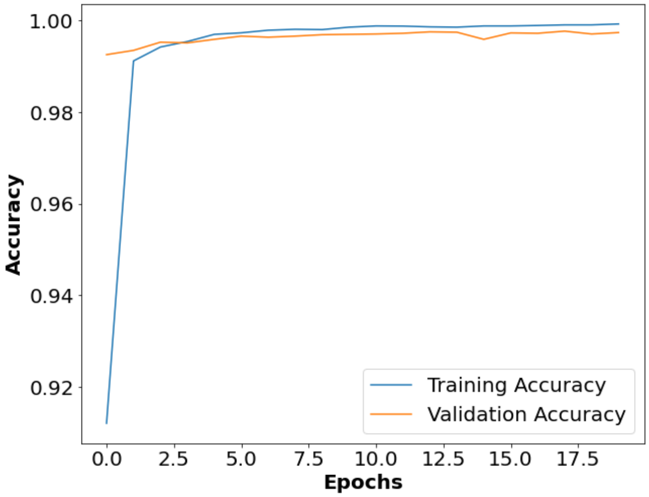
TABLE V. ANALYSIS OF FEATURE EXTRACTION RESULTS

Dataset	Method	Accuracy	Precision	Recall	Specificity	F1-score
LiTS	ResNet	0.952	0.941	0.964	0.948	0.952
	VGG19	0.961	0.955	0.972	0.960	0.961
	InceptionNet	0.844	0.832	0.857	0.841	0.844
	EfficientNetB2	0.997	0.998	0.978	0.995	0.987
CHAOS	ResNet	0.923	0.912	0.934	0.921	0.923
	VGG19	0.938	0.929	0.945	0.937	0.938
	InceptionNet	0.819	0.806	0.832	0.818	0.821
	EfficientNetB2	0.996	0.997	0.988	0.996	0.992

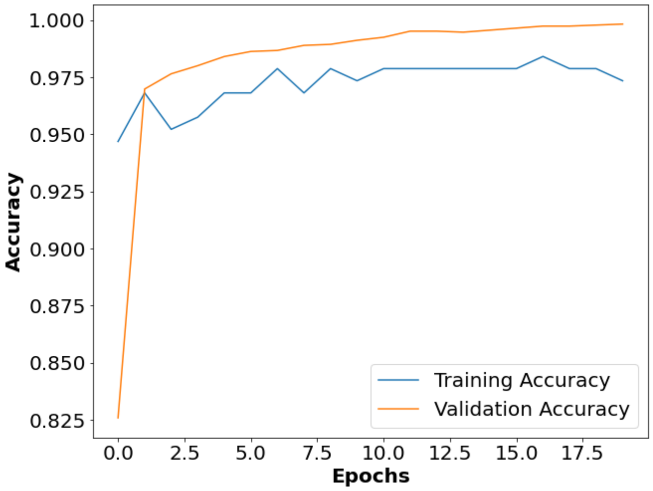
TABLE VI. ANALYSIS OF CLASSIFICATION RESULTS

Dataset	Method	Accuracy	Precision	Recall	Specificity	F1-score
LiTS	CNN	0.945	0.952	0.938	0.964	0.945
	SNN	0.889	0.895	0.874	0.870	0.884
	FNN	0.927	0.933	0.956	0.922	0.945
	DNN	0.997	0.998	0.978	0.995	0.987
CHAOS	CNN	0.918	0.921	0.905	0.934	0.913
	SNN	0.853	0.855	0.848	0.863	0.851
	FNN	0.896	0.901	0.878	0.869	0.889
	DNN	0.996	0.997	0.988	0.996	0.992

1) Accuracy function: Fig. 5(a) explains the training and validation of the accuracy for the AFMRCNN approach based on the LiTS dataset. Fig. 5(b) explains the training and validation of the accuracy for the AFMRCNN approach based on the CHAOS dataset. The accuracy of the validation is attained through the model on the 20th epoch. The training data is continuously partitioned into smaller portions, which are utilized to update the parameter of the model. The number of epochs describes the total number of iterations that the model works on the training data. Accuracy allows for the easy comparison among the models, representing a commendable model's performance, considering the classes balanced.



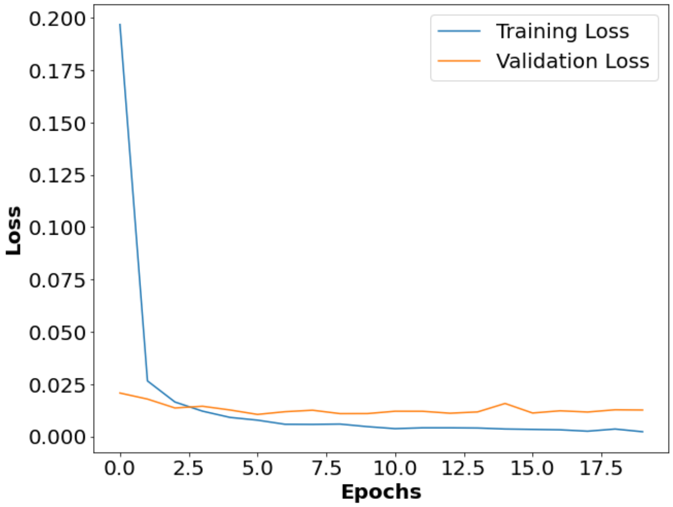
(a)



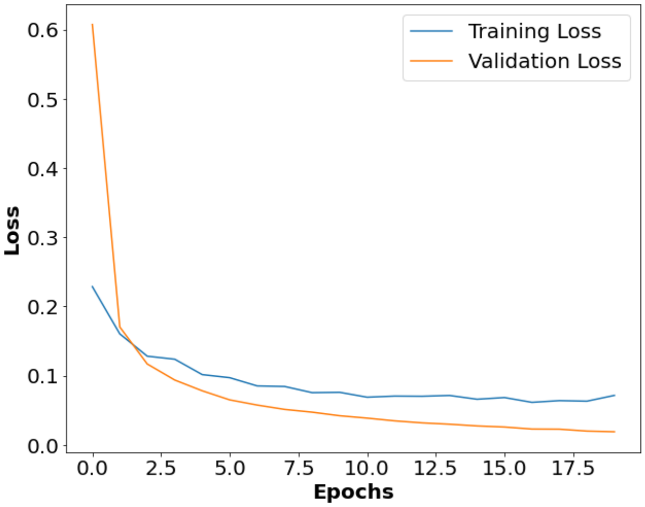
(b)

Fig. 5. Accuracy function: (a) LiTS and (b) CHAOS.

2) Loss function: Fig. 6(a) explains the training and validation of the loss function for the AFMRCNN on LiTS dataset. Fig. 6(b) explains the training and validation of the loss function for the AFMRCNN on CHAOS dataset. The minimum loss of the validation is attained through the model on the 17th epoch, while the training continues until the 18th epoch. However, after the 17th epoch, the loss was further not minimized. At every epoch, the model often changes its internal parameters by focusing on the input data when compared with the target labels to reduce the loss. This loss function indicates that the AFMRCNN has learned to simplify well to the validation data up to this point. This is a significant indicator of the performance of the AFMRCNN on the given data.



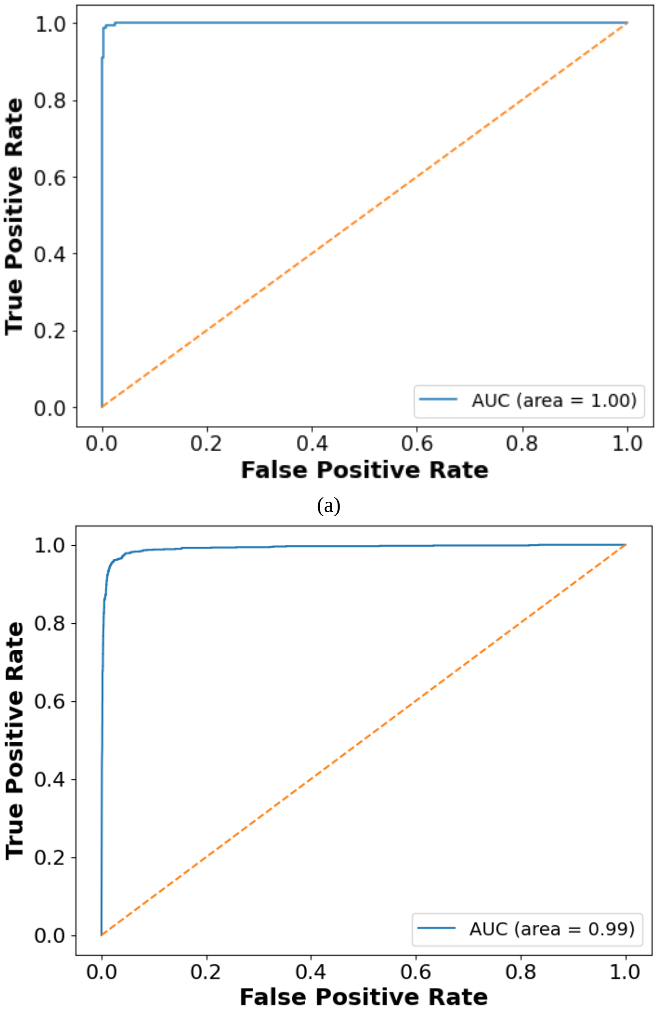
(a)



(b)

Fig. 6. Loss function: (a) LiTS and (b) CHAOS.

3) ROC Curve: Fig. 7(a) depicts the confusion matrix for AFMRCNN with an affiliation among True Positive Rate (TPR) and False Positive Rate (FPR) based on LiTS dataset. Fig. 7(b) depicts the confusion matrix for AFMRCNN with an affiliation among TPR and FPR based on CHAOS dataset. ROC is used to depict the graphical evaluation of the binary classification. The ROC curve is widely used to evaluate the performance of classifiers. In this process, the area under the ROC curve is fixed at 0.72. Estimating the ROC curve with TPR and TNR supports understanding the trade-offs between sensitivity and specificity. This supports selecting the optimal balance between correctly identifying positives and minimizing false positives, leading to a commendable AFMRCNN performance.



(a)

(b)

Fig. 7. ROC curve: (a) LiTS and (b) CHAOS.

B. *Comparative Analysis*

The efficiency of the proposed method is compared with the existing methods based on the LiTS and CHAOS datasets in this section. The existing methods like UoloNet [16], CMG [18] and UNet+++ PARCA [19] are compared and estimated with the proposed AFMRCNN approach. AF mechanism significantly minimizes the chances of missing small or irregularly shaped tumours which are not aligned well with predefined anchors, leading to better segmentation results. In classification tasks, the DNN is utilized for classifying liver tumours with the help of the sigmoid activation function in the final layer to produce a probability distribution over different classes. This process allowing for straightforward interpretation of the DNN confidence in every class and enabling decision-making, resulting in a superior classification of tumour and non-tumour. Table VII represents the comparative analysis on the LiTS and CHAOS datasets.

C. *Discussion*

The advantages of the proposed AFMRCNN and limitations of the existing works are discussed. The limitations of the previous works: In UoloNet [16], a high level of noise in the segmented module impacted the accurate identification of small labels in the image. SD-net model [17] was memory intensive because of its high resolution of images, resulted in limited scalability. CMG model [18] was limited in its capability to determine the complex details of lesions and basically focused only on segment lesions from inaccurate registration. UNet++ [19] was failed to significantly segment tumours in the liver that overlapped with other organs, resulted in poor performance. To overcome these problems, the AFMRCNN mechanism enhances the detection of small objects like tumors through reducing the dependency on predefined anchor boxes, which often introduce noise. The anchor-free approach enables the network to focus more precisely on the actual regions of interest, enhancing the small label segmentation accuracy. The anchor-free design enables the model to learn object boundaries dynamically, resulted in better segmentation.

TABLE VII. COMPARATIVE ANALYSIS USING LITS AND CHAOS DATASET

Dataset	Method	Recall	DSC	VOE	RVD
LiTS	UoloNet [16]	0.821	0.462	NA	NA
	UNet++ + PARCA [19]	0.964	0.963	0.057	0.015
	Proposed AFMRCNN	0.978	0.978	0.147	0.119
CHAOS	CMG [18]	0.927	0.928	NA	0.061
	UNet++ + PARCA [19]	0.912	0.951	0.096	0.079
	Proposed AFMRCNN	0.988	0.988	0.652	0.532

This research demonstrates that the binary classification with CNN is effectively utilized for semantic segmentation in medical diagnosis, particularly for liver tumour segmentation. Moreover, this research demonstrates that the segmentation has improved the AFMRCNN effectiveness by solving the individual network bias. This research proposes an end-to-end automatic liver segmentation by using the AFMRCNN approach. This approach conducts multi-level features and AF mechanism for decontaminating the features. ARMRCNN approach improves the quality of segmentation masks by considering them more accurately on RoI. This results in a better description of tumour boundaries, which is complex for treatment planning and assessment. Hence, the proposed AFMRCNN approach attains a better DSC of 0.978 and 0.988, VOE of 14.75 and 6.52, RVD of 11.92 and 5.32 and IoU of 0.98 and 0.98 on LiTS and CHAOS datasets. The results of this research demonstrate that the utilization of AFMRCNN enhances both accuracy and performance of medical diagnostics in the segmentation of liver tumours. The development in the expertise has the latent to improve patients' outcomes, while enabling the development of personalized treatment plans.

V. CONCLUSION

Utilizing the CT images for diagnosing the liver tumours is the most popular technique because it provides high-resolution images with fine anatomical details. Consistent liver tumour segmentation requires accurate image processing across different steps and inaccuracies in this process because of the subsequent segmentation tasks. Hence, this research proposes the AFMRCNN approach to address the problem of image overlapping for the precise segmentation of the liver tumour.

The AF mechanism attains greater segmentation accuracy, particularly for irregularly shaped liver tumours through the prediction of the presence and boundaries of tumours directly. The DNN approach is performed to classify the liver tumours into binary classes through the Adam optimizers and adjust the learning rate dynamically to enhance the convergence. The experimental results exhibit the proposed AFMRCNN approach's commendable DSC of 0.147 on the LiTS dataset as compared to the existing methods, UoloNet and UNet++ + PARCA. However, UoloNet and UNet++ + PARCA attain a minimum DSC of 0.462 and 0.963 on the LiTS dataset. Future work will involve the hybrid DL approach for enhancing the overall model results during the segmenting of the liver tumour.

REFERENCES

[1] T. Lei, R. Sun, X. Du, H. Fu, C. Zhang, and A. K. Nandi, "SGU-Net: Shape-guided ultralight network for abdominal image segmentation," IEEE J. Biomed. Health. Inf, Vol. 27, pp.1431-1442, March 2023.

[2] S. Fogarollo, R. Bale, and M. Harders, "Towards liver segmentation in the wild via contrastive distillation. Int. J. Comput. Assisted Radiol. Surg, Vol. 18, pp.1143-1149, May 2023.

[3] S. Bogoi, A. Udrea, "A Lightweight Deep Learning Approach for Liver Segmentation," Mathematics 2023, Vol.11, p. 95, December 2022.

[4] C. Hu, T. Xia, Y. Cui, Q. Zou, Y. Wang, W. Xiao, S. Ju, and X. Li, "Trustworthy multi-phase liver tumour segmentation via evidence-based uncertainty," Eng. Appl. Artif. Intell, Vol. 133, p.108289, July 2024.

[5] Z. Xia, M. Liao, S. Di, Y. Zhao, W. Liang, and N. N. Xiong, "Automatic liver segmentation from CT volumes based on multi-view information fusion and condition random fields," Opt. Laser Technol, Vol. 179, p.111298, December 2024.

[6] P. V. Nayantara, S. Kamath, R. Kadavigere, and K. N. Manjunath, "Automatic Liver Segmentation from Multiphase CT Using Modified SegNet and ASPP Module," SN Comput. Sci, Vol. 5, p.377, March 2024.

[7] M. Y. Ansari, A. Abdalla, M. Y. Ansari, M. I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S. S. Singh, J. Abinahed, A. Al-Ansari, S. Balakrishnan, and S. P. Dakua, "Practical utility of liver segmentation methods in clinical surgeries and interventions," BMC Med. Imaging, Vol. 22, p.97, May 2022.

[8] A. E. Kavur, L. I. Kuncheva, and M. A. Selver, "Basic ensembles of vanilla-style deep learning models improve liver segmentation from CT images," In Convolutional Neural Networks for Medical Image Processing Applications, (pp. 52-74), 2022. CRC Press.

[9] L. Jiang, J. Ou, R. Liu, Y. Zou, T. Xie, H. Xiao, and T. Bai, "RMAU-Net: residual multi-scale attention U-Net for liver and tumor segmentation in CT images," Comput. Biol. Med, Vol. 158, p.106838, May 2023.

[10] X. Liu, K. Ono, and R. Bise, "A data augmentation approach that ensures the reliability of foregrounds in medical image segmentation," Image Vision Comput. Vol. 147, p.105056, July 2024. https://doi.org/10.1016/j.imavis.2024.105056

[11] E. Othman, M. Mahmoud, H. Dhahri, H. Abdulkader, A. Mahmood, M. Ibrahim, "Automatic Detection of Liver Cancer Using Hybrid Pre-Trained Models," Sensors 2022, Vol. 22, pp. 5429, july 2022.

[12] A.M. Hendi, M. A. Hossain, N. A. Majrashi, S. Limkar, B. M. Elamin, M. Rahman, "Adaptive Method for Exploring Deep Learning Techniques for Subtyping and Prediction of Liver Disease," Appl. Sci. 2024, Vol. 14, p. 1488, Febrauary 2024.

[13] D. Popescu, A. Stanciulescu, M. D. Pomohaci, L. Ichim, "Decision Support System for Liver Lesion Segmentation Based on Advanced Convolutional Neural Network Architectures," Bioengineering 2022, Vol. 9, p. 467, September 2022.

[14] Y. Chen, C. Zheng, F. Hu, T. Zhou, L. Feng, G. Xu, Z. Yi, and X. Zhang, "Efficient two-step liver and tumour segmentation on abdominal CT via deep learning and a conditional random field," Comput. Biol. Med,Vol. 150, p.106076, November 2022.

[15] H. L. Elghazy, and M. W. Fakhr, "Dual-and triple-stream RESUNET/UNET architectures for multi-modal liver segmentation," IET Image Proc, Vol. 17, pp.1224-1235, March 2023.

[16] K. Zhang, L. Zhang, and H. Pan, "UoloNet: based on multi-tasking enhanced small target medical segmentation model," Artif. Intell. Rev, Vol. 57, p. 31, Febrauary 2024.

[17] S. Huang, J. Luo, Y. Ou, Wangjun shen, Y. Pang, X. Nie, and G. Zhang, "Sd-net: A semi-supervised double-cooperative network for liver segmentation from computed tomography (CT) images," J. Cancer Res. Clin. Oncol, Vol. 150, p. 79, Febrauary 2024.

[18] W. Yu, M. Wang, Y. Zhang, and L. Zhao, "Reciprocal cross-modal guidance for liver lesion segmentation from multiple phases under incomplete overlap," Biomed. Signal Process. Contro, Vol. 88, p.105561, Febrauary 2024.

[19] D. T. Kushnure, S. Tyagi, and S. N. Talbar, "LiM-Net: Lightweight multi-level multiscale network with deep residual learning for automatic liver segmentation in CT images," Biomed. Signal Process. Control Vol. 80, p.104305, Febrauary 2023.

[20] R.V. Manjunath, and K. Kwadiki, "Automatic liver and tumour segmentation from CT images using Deep learning algorithm," Results Control Optim. Vol. 6, p.100087, March 2022. https://doi.org/10.1016/j.rico.2021.100087

[21] LiTs dataset link: https://www.kaggle.com/datasets/andrewmvd/liver-tumor-segmentation. (Accessed on 30/07/2024)

[22] CHAOS dataset link: https://www.kaggle.com/datasets/anhoangvo/chaos-t1-and-t2. (Accessed on 30/07/2024)

[23] V. Dakshayani, G. R. Locharla, P. Pławiak, V. Datti, C. Karri, "Design of a Gabor Filter-Based Image Denoising Hardware Model," Electronics 2022, Vol. 11, p. 1063, March 2022.

[24] M. Zhou, J. Wang, B. Li, "ARG-Mask RCNN: An Infrared Insulator Fault-Detection Network Based on Improved Mask RCNN," Sensors 2022, Vol. 22, p. 4720, june 2022.

[25] A. M. J. Z. Rahman, M. Gupta, S. Aarathi, T. R. Mahesh, V. V. Kumar, S. Y. Kumaran, and S. Guluwadi, "Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering," BMC Med. Inf. Decis. Making, Vol. 24, p.113, April 2024. https://doi.org/10.1186/s12911-024-02519-x

[26] X. Ding, X. Hou, M. Xia, Y. Ismail, and J. Ye, "Predictions of macroscopic mechanical properties and microscopic cracks of unidirectional fibre-reinforced polymer composites using deep neural network (DNN)," Compos. Struct, Vol. 302, p.116248, December 2022. https://doi.org/10.1016/j.compstruct.2022.116248.

Efficient Task Offloading Using Ant Colony Optimization and Reptile Search Algorithms in Edge Computing for Things Context

Ting Zhang*, Xiaojie Guo
School of Information Engineering, Jiaozuo University, Jiaozuo 454000, China

***Abstract*—The widespread use of Internet of Things (IoT) technology has triggered unparalleled data creation and processing needs, necessitating effective computation offloading solutions. Conventional edge computing approaches have difficulties in dealing with rising energy usage issues and task allocation delays. This study introduces a novel hybrid metaheuristic algorithm called ACO-RSA, which synergizes two metaheuristic algorithms, Ant Colony Optimization (ACO) and Reptile Search Algorithm (RSA). The proposed approach addresses the energy and latency issues associated with offloading computations in IoT edge computing environments. A comprehensive system design that effectively encapsulates the uplink transmission communication model and a personalized multi-user computing task load model is developed. The system considers various constraints, such as network latency, task complexity, and available computing resources. Based on this, we formulate an optimization objective suitable for computing outsourcing in the IoT ecosystem. Simulations conducted in a real-world IoT scenario demonstrate that ACO-RSA significantly reduces both time delay and energy consumption compared to benchmark algorithms, achieving up to 27.6% energy savings and 25.4% reduction in time delay. ACO-RSA exhibits robustness and scalability when optimizing task offloading in IoT edge computing environments.**

Keywords*—Task offloading; edge computing; ant colony optimization; reptile search algorithm; Internet of Things; energy efficiency

I. Introduction

The exponential growth of Internet of Things (IoT) devices has fundamentally transformed how data is produced, analyzed, and used in numerous sectors, including smart cities [1], healthcare [2], and automated manufacturing [3]. Nevertheless, the substantial increase in data flow places substantial computing requirements on centralized and distributed systems, requiring the investigation of sophisticated task offloading strategies [4]. Conventional cloud computing models encounter delay and data transfer capacity limitations. In this regard, edge computing presents a viable option by bringing computing resources close to data sources [5]. However, improving task offloading in edge computing contexts has distinct difficulties related to energy use and latency [6].

Heuristic optimization techniques have contributed to overcoming these issues by evaluating a variety of task offloading policies. An influential group of these heuristics is Ant Colony Optimization (ACO), a group of algorithms based on the pheromone-based learning of ants foraging [7]. Similar algorithms, such as the Reptile Search Algorithm (RSA), can overcome the exploration versus exploitation challenge by imitating hunter-reptile foraging strategies [8]. The methods have been proven to work on optimization problems across various applications, including image processing, power systems, and edge computing. While the former is difficult to pull out of premature convergence, the latter must be carefully controlled judiciously to avoid falling into unoptimal solutions.

In light of the aforementioned, this paper presents a new metaheuristic method called ACO-RSA, which merges two metaheuristic methods, ACO and RSA, to solve offloading problems in IoT edge computing. Our strategy aims to achieve the benefit of both ACO and RSA by minimizing energy use and time delays. The paper presents the system architecture, with a communication model relying on an uplink transmission and the distribution task model between multiple users. This paper next introduces an objective function for IoT context optimization.

The system design also includes a multi-user task distribution model and an uplink transmission communication model that considers transmission delays, the complexity of the tasks to be solved by standalone devices, and the available computing resources. We design an optimization target specifically for these systems to adapt to IoT task offloading mechanisms and conduct extensive simulations to demonstrate the effectiveness of ACO-RSA. Experimental results show that our solution ACO-RSA can effectively reduce the energy consumption and latency of existing algorithms, which has obvious advantages in practical IoT application scenarios.

The rest of this paper is organized as follows. Section II contains a detailed literature review, including previous work on task offloading and metaheuristic optimization techniques. Section III discusses the system model and problem formulation. Section IV discusses our ACO-RSA algorithm in detail. An experimental setup and simulation results are presented in Section V for evaluating the proposed approach's performance. Lastly, Section VI summarizes the key insights, limitations, and future research possibilities concluding the paper.

II. Background

A. Edge Computing in IoT

In recent years, cloud computing platforms have evolved into the first choice for provisioning services because of their flexibility and cost-effectiveness [9]. Fundamentally, there has been a shift towards centralized systems for processing and storing data in large data centers [10]. The IoT is advancing rapidly, with major effects on various sectors of society, mainly healthcare, transportation, and manufacturing [11]. They have many IoT devices at the edge producing huge volumes of data, which may require very little or significant resources (e.g., storage and processing) [12]. Service delivery is a big problem in cloud systems. Additional challenges involve the placement of sensors and actuators in space, response time constraints, privacy issues, and data processing. Apart from cloud computing, these barriers are a driving force for innovation in solutions [13].

Edge computing places servers for computation and storage at the edges of the internet, relatively closer to where data is generated. The above approximations are due to the lower latencies, more privacy concerns considering data being transferred, and suboptimal energy overhead [14]. This hybridization extends edge computing capabilities to distribute and gradually interconnect with cloud-based processing components.

Fig. 1 shows the hybrid edge-cloud reference design encompasses a range of computing and storage capabilities. Thing nodes are small and constrained devices with limited computing and storage capabilities, so they can only do basic functions like detecting, acting in response to certain events, and sending and receiving data. In contrast to Thing nodes, local nodes are capable of more processing, storing, and communicating. This allows them to process, analyze, and send data and interface with both the cloud and thing nodes. A local node may also host IoT applications at the network edge.

A local node can be any device such as a gateway, access point, router, switch, local server, cellphone, or linked vehicle. Furthermore, various physical gadgets can fulfill certain functions. A traffic light, for example, can be both a thing node that detects its surroundings and acts accordingly and a local node that collects and analyzes data. Centralized data centers provide extensive processing, storage, and communication capabilities in the cloud. Due to these characteristics, cloud solutions can accommodate many IoT components that require substantial physical resources.

B. Task Offloading Challenges

In edge computing settings, task offloading has distinct issues that must be addressed for efficient and effective task distribution. As shown in Fig. 2, the problems arise from IoT devices' intrinsic attributes, edge computing's decentralized nature, and dynamic and unexpected network circumstances. IoT devices are often situated in geographically scattered locations with diverse network conditions. Latency may impede communication between these devices and edge servers, particularly in cases of network congestion or when the devices are placed far from the edge nodes. In addition, a restricted amount of available bandwidth might result in congestion during data transmission, which causes delay problems.

IoT applications can include a wide range of functions with different levels of complexity, including basic data processing and more demanding processes like real-time video analysis or machine learning inference [15]. Because different workloads may have no shared characteristics and time constraints, the diversity of these jobs adds difficulty for offloading methods. In addition, edge servers could differ widely in the processing power, memory, and storage they support [16].

Energy efficiency is crucial in IoT systems since devices often have limited battery capacity. Task offloading may mitigate the computing burden on IoT devices, thereby conserving their energy [17]. Nevertheless, offloading necessitates energy use, particularly in data transmission to edge servers and the reception of results.

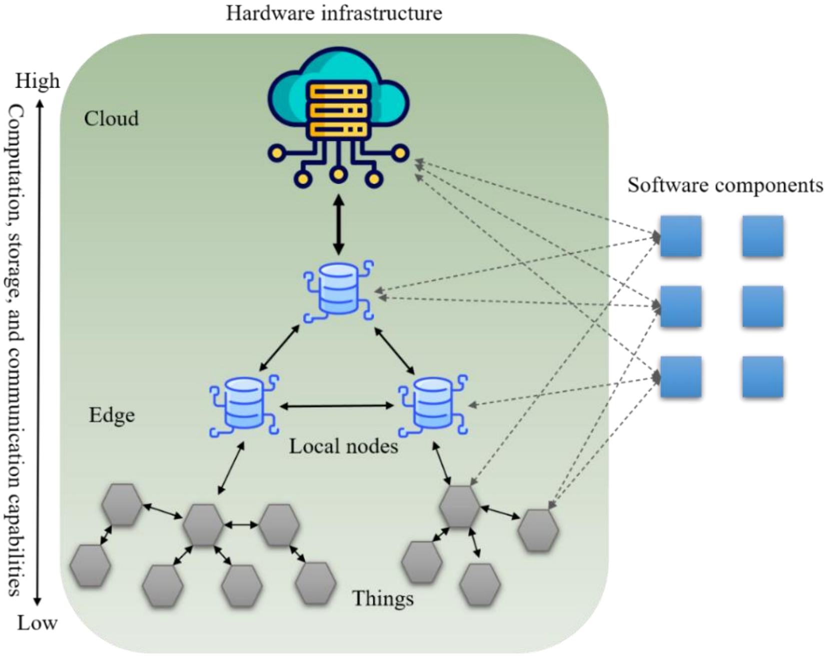


Fig. 1. Edge-cloud reference architecture.

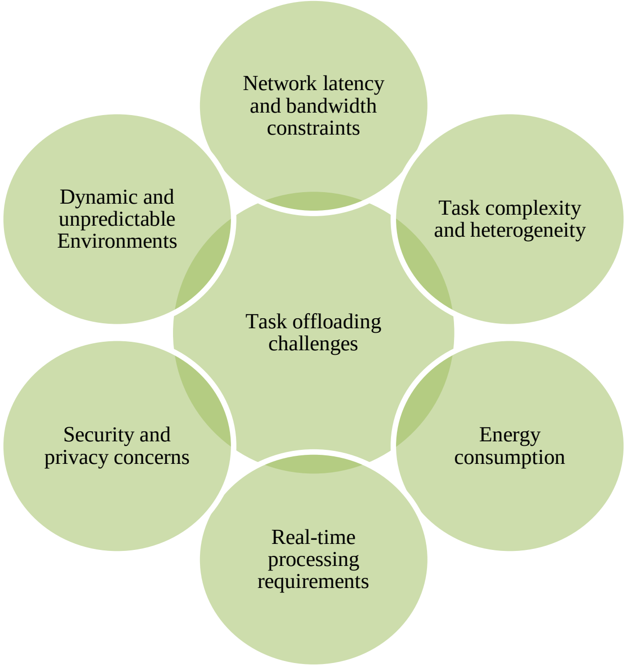

Network latency and bandwidth constraints
Dynamic and unpredictable Environments
Task complexity and heterogeneity
Task offloading challenges
Security and privacy concerns
Energy consumption
Real-time processing requirements

Fig. 2. Task offloading challenges.

IoT applications, including autonomous cars, smart manufacturing, and healthcare monitoring, need immediate processing. During these situations, even a small delay in carrying out a job might result in significant repercussions, highlighting the need to reduce latency and guarantee prompt task fulfillment [18]. The task involves creating offloading algorithms that provide real-time performance while distributing the computing workload across several edge servers.

Transmitting sensitive data via networks might expose it to interception or unwanted access. Moreover, the decentralized structure of edge computing increases attack vulnerability, making it more difficult to protect all nodes from possible risks. The IoT ecosystem is intrinsically characterized by its dynamic nature, including fluctuating network conditions, shifting workloads, and varied resource availability at edge servers. These variables contribute to an uncertain environment that complicates the job offloading process.

C. Edge Computing-Enabled Task Offloading Mechanisms

Xiao, et al. [19] proposed a three-layer IoT structure that addresses different aspects of tasks. Edge computing and blockchain architectures address the problem of sensitivity to task delays and enable service providers to maximize their profits. In addition, due to the features of the edge computing server in the second layer, the proposed algorithm is executed as a smart contract. This smart contract is responsible for managing and distributing edge computing resources. In particular, a complementary approach called stacked cache is proposed to facilitate the fair distribution of resources on edge computing servers.

You and Tang [20] proposed an energy-efficient, low-delay PSO-based task offloading technique for low-resource edge devices. The problem formulated is a multi-objective optimization incorporating latency, task execution cost, and energy usage. The offloading of all tasks to different mobile edge servers represents the particle's fitness function. Simulation experiments compare the PSO offloading strategy with simulated annealing and genetic algorithms. Based on the experimental results, PSO-based task offloading reduces edge server latency, balances energy consumption, and achieves reasonable resource allocation.

Chen, et al. [21] combined two contradictory offloading objectives, namely increasing the job completion rate with an acceptable delay and decreasing the energy consumption of devices. The task offloading issue was established to achieve equilibrium between two challenging aims. Subsequently, they clarified it as a problem of dynamic task offloading based on Markov Decision Processes (MDP). A Deep Reinforcement Learning (DRL)-based dynamic task offloading (DDTO) method was developed to address this issue. The DDTO algorithm adapts to the constantly changing and intricate context and appropriately modifies the way tasks are offloaded. Experiments demonstrate DDTO's rapid convergence. The trial findings confirm the efficiency of the DDTO algorithm in achieving a balance between the finish ratio and power.

Kong, et al. [22] introduced a dependable and effective approach for task offloading using the multi-feedback trust strategy. A reliable and effective framework is built, which may significantly enhance trust computing and job offloading. Furthermore, the broker utilizes dynamic data monitoring to offer a multi-feedback trust management design using interaction frequency and time attenuation. This model aims to establish a trustworthy operating setting. Moreover, a trust-weighted K-means clustering approach is developed using resource qualities to improve service dependability. This algorithm efficiently and correctly groups the resource nodes needed for the job. A job offloading model uses trust clustering to improve user experience and enhance system performance. In contrast to current task processing models prioritizing task offloading, the suggested approach further incorporates resource pretreatment, trust assessment, and resource clustering before task processing.

Aghapour, et al. [23] presented a solution using deep reinforcement learning to break down offloading and resource allocation into two elementary issues. The Salp Swarm Algorithm (SSA) optimizes resource allocation by updating offloading policy according to environmental data. The proposed method investigates various deep-learning tasks of IoT devices under different cloudlet server capacities. Simulation findings demonstrate that the suggested approach has the lowest latency and power consumption cost. On average, 92%, 17%, and 12% improvements were shown compared to the full local, full offload, joint resource allocation, and computation offloading techniques, respectively.

Bolourian and Shah-Mansouri [24] developed a wireless-powered mobile edge computing system divided into three tiers: IoT devices, edge servers, and cloud. A formulation of a combinatorial optimization problem is presented to minimize wireless energy transmission. Bipartite graph matching is employed to address the issue's complexity, and a harvest-then-offload technique is suggested for IoT devices. The proposed approach utilizes parallel processing to enhance its performance. Empirical tests demonstrate that the recommended technique substantially decreases the energy demand for operating IoT devices compared to other offloading strategies.

Nandi, et al. [25] developed a task offloading strategy to achieve a compromise between device usefulness and execution cost. Utility is determined by job execution delay and energy usage in energy-harvesting IoT devices. The task offloading issue is a problem of selecting a subset that achieves the required trade-off. Social Cognitive Optimization (SCO) addresses the offloading issue and achieves the required polynomial time of execution. The findings confirm the superior effectiveness of the approach regarding job execution speed, energy use, cost efficiency, and task abandonment rate when compared to the most advanced existing methods.

Due to its unique features, the proposed ACO-RSA provides a result-oriented solution to address the issues of energy consumption and latency in task offloading. While ACO has excellent exploration performance thanks to pheromone-assisted learning, it also has premature convergence. As RSA learns adaptively, it can utilize space better and more efficiently; however, exploration capacity may be limited, resulting in suboptimal outcomes. As a result of combining the advantages of both ACO exploration and RSA exploitation, the ACO-RSA hybrid method eliminates these disadvantages. It improves task-shifting efficiency, as illustrated by evaluation results. Using a more balanced and robust approach, we achieve significant reductions in energy consumption and latency, in contrast to existing research that proposes single-objective optimizations or scalability issues. The ACO-RSA can also achieve convergence more quickly and accurately than the current research literature.

III. System Design

A. System Model

Consider a defined geographic region containing a population of N mobile user devices, each denoted by n in the range 1 to N. Each device is assigned a unique computing task with different priorities, such as energy efficiency or latency minimization. This region has a base station equipped with several Mobile Edge Computing (MEC) servers. Mobile devices can communicate with MEC servers via a radio access network to reduce their computing load. The base station is configured with M-channels, each supporting a single connection between the device and the MEC server.

Mobile devices are assigned static channels, and edge servers have limited computing resources. The model also assumes a central cloud server with unlimited capacity. A single-edge server can handle the computing needs of only one device at a time. Tasks that exceed the capacity of an edge server are shifted to the cloud via the backbone network. The mobile edge computing architecture is shown in Fig. 3.

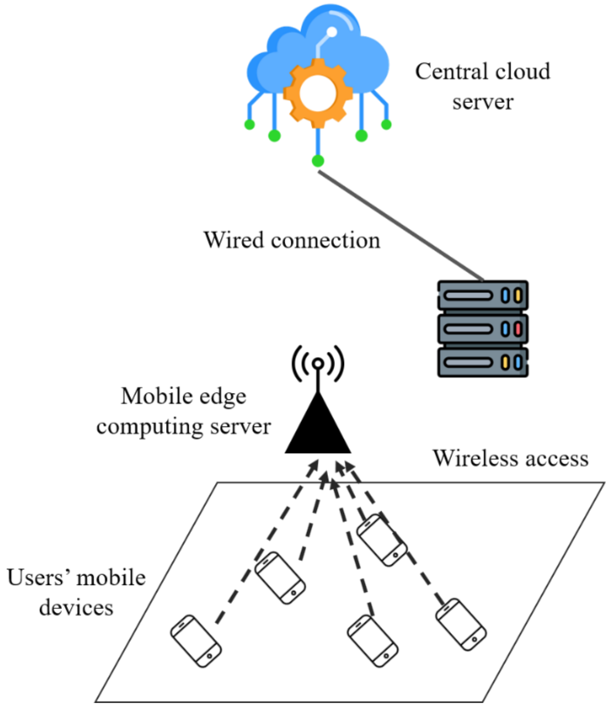


Fig. 3. Network model.

Mobile device offloading involves three steps. First, the device determines whether the task should be executed locally or offloaded to an edge server. During offloading, the system assesses the availability of the edge server resources. Insufficient resources require a decision between queuing at the current edge server or offloading to the cloud.

Orthogonal uplink channels ensure non-interference between devices in the proposed system. According to Eq. 1, the uplink rate R_n is calculated by denoting the transmission power of device n as P_n. Here, S_n refers to the device n's uplink bandwidth, z_n is the channel gain coefficient, and δ_n^2 represents the noise power.

$$R_n = S_n log_2\left(1 + \frac{P_n z_n}{\delta_n^2}\right), \quad \forall n \in N \tag{1}$$

B. Multi-User Computing Task Load Model

This subsection presents a computational task model to simplify the analysis. A computing task is characterized by three main components: (1) volume of data, including program code and input parameters, that needs to be uploaded for outsourced tasks; (2) computing capacity, typically measured in CPU cycles; and (3) results in data to be downloaded for outsourced tasks. Eq. 2 formally defines a computing task Q_v as a function of the result data R_v, the computing capacity C_v, and the data volume D_v.

$$Q_v \underset{=}{\Delta} (D_v, C_v, R_v) \tag{2}$$

Using a program called graph analysis, mobile devices can discover these components. A multi-user delay and energy consumption load model is built to meet personalized user needs by analyzing different application types and corresponding data. This model evaluates the impact of local and cloud task execution on performance.

Initially, each mobile device determines whether to execute its task locally or offload it. For locally executed tasks, the computational requirements of device n are denoted by J_{n1} CPU cycles, and the task itself requires j_{n1} CPU cycles. The local execution time, t_{n1}, is calculated according to Eq. 3. Energy consumption, E_{n1}, for local execution is determined by Eq. 4, where λ is a constant energy consumption coefficient related to the device's chip structure, typically set to 10^{-26}. The total local overhead, W_{n1}, is calculated using Eq. 5, incorporating trade-off coefficients α_{n1} and α_{n2} for energy consumption and task execution time, respectively. These coefficients must adhere to the constraints outlined in Eq. 6.

$$t_{n1} = \frac{j_{n1}}{J_{n1}} \tag{3}$$

$$E_{n1} = \lambda j_{n1} (J_{n1})^2 \tag{4}$$

$$W_{n1} = a_{n1} E_{n1} + a_{n2} t_{n1} \tag{5}$$

$$\begin{cases} a_{n1}, a_{n2} \in [0,1] \\ a_{n1} + a_{n2} = 1 \end{cases} \tag{6}$$

When the n^{th} user's task is offloaded to an edge server, the processing delay is determined by Eq. 7. This delay comprises two components: the uplink transmission delay for task input data and the edge server execution time. The energy consumption for transferring the task to the edge server is computed using Eq. 10, where β represents the device's power amplifier efficiency. Eq. 11 formulates the total offloading overhead.

$$t_{n2}(P_n, J_{n2}) = t_{n2(1)}(P_n) + t_{n2(2)}(J_{n2}) \tag{7}$$

$$t_{n2(1)}(P_n) = \frac{s_n}{W_n log_2(1 + \sigma_n P_n)} \tag{8}$$

$$t_{n2(2)}(J_{n2}) = \frac{j_n}{J_{n2}} \tag{9}$$

$$E_{n2} = \frac{P_n}{\beta} t_{n2(1)}(P_n) = \frac{P_n}{\beta} \frac{s_n}{W_n log_2(1 + \sigma_n P_n)} \tag{10}$$

$$W_{n2} = a_{n1} E_{n2} + a_{n2} t_{n2}(P_n, J_{n2}) \tag{11}$$

C. Optimization Objective Function

The total overhead incurred by mobile device n during task offloading is calculated according to Eq. 12. The objective function is formulated to minimize the aggregate overhead across all devices. To achieve this minimum overhead, the optimal offloading strategy (X), uplink power allocation (P), and edge computing resource allocation (M) are determined. Eq. 13 expresses the optimization goal.

$$W_n = (1 - x_n) W_{n1} + x_n W_{n2} \tag{12}$$

$$\min_{X,P,M} W = \sum_{n=1}^{N} (1 - x_n) W_{n1} + x_n W_{n2} \tag{13}$$

Eq. 14 details the constraints. The binary variable x_n indicates the offloading decision for device n (1 for offloading, 0 for local execution). P_{max} represents the maximum transmission power, J_{max} is the maximum computational resource, and N_2 is the set of devices opting for offloading. The first constraint specifies which devices should be offloaded. The maximum power of the uplink device is limited by constraint 2. With constraint 3, edge computing resources are not allocated beyond the server's capacity. Negative resource allocation is prevented by constraint 4. As a final constraint, uplink transmissions are concurrently limited to N.

$$\begin{cases} 1: x_n \in \{0,1\}, \forall n \in N \\ 2: 0 < P_n \le P_{max}, \forall n \in N_2 \\ 3: \sum_{n \in N_2} J_2 \le J_{max} \\ 4: J_{n2} > 0, \forall n \in N_2 \\ 5: \sum_{n \in N} x_n S_n \le B \end{cases} \tag{14}$$

IV. The ACO-RSA Algorithm

A. Ant Colony Optimization

The ACO algorithm is a metaheuristic inspired by the foraging behavior of real ants. Artificial ants iteratively construct solutions to optimization problems by laying down and following virtual pheromone trails. These paths represent solution components whose intensity correlates with the solution quality. Ants will likely select subsequent solution components based on pheromone levels and problem-specific heuristic information. Pheromone levels are dynamically updated to amplify high-quality solutions. ACO is a probabilistic search method that does not guarantee optimal solutions but often provides acceptable approximations within reasonable computing time.

Mathematically, ACO can be described in the following manner. A population of artificial ants is created. Pheromone levels on all potential solution components are initialized to a small positive value. Ants construct solutions by iteratively selecting components. The probability of selecting component *j* after component *i* is determined by Eq. 15:

$$P_{ij} = \frac{\tau_{ij}^{\alpha}.\eta_{ij}^{\beta}}{\sum\left(\tau_{ik}^{\alpha}.\eta_{ik}^{\beta}\right)} \quad (15)$$

Where α and β are parameters balancing pheromone and heuristic influence, η_{ij} is the heuristic information of components *i* and *j*, and τ_{ij} is the pheromone level of two components. After selecting a component, an ant updates its pheromone level according to Eq. 16.

$$\tau_{ij} = (1-\rho)\tau_{ij} + \rho\Delta\tau_{ij} \quad (16)$$

Where ρ is the pheromone evaporation rate and $\Delta\tau_{ij}$ is the pheromone the ant deposits. Once all ants have created solutions, pheromone levels are updated globally based on the overall quality of the solution. Combining probabilistic selection, pheromone amplification, and evaporation, this iterative process allows ACO to explore the solution space effectively.

B. *Reptile Search Algorithm*

RSA is a swarm intelligence metaheuristic derived from crocodile hunting strategies. By modeling competitive and cooperative interactions within a reptile population, RSA determines the global optima for optimization problems. Known for its simplicity, flexibility, and efficiency, RSA has found applications in image processing, power systems, and engineering.

RSA's optimization methodology, outlined in Eq. 17, involves iteratively refining a population of solutions (*X*). Eq. 18 details the random generation of candidate solutions, where $x_{i,k}$ represents the kth position of the i^{th} individual, *N* refers to the population size, *D* specifies the problem dimension, and *rd* indicates a random value within the problem's lower (*LB*) and upper (*UB*) bounds.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \cdots & \cdots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,D} \end{bmatrix} \quad (17)$$

$$x_{i,k} = rd \times (UB - LB) + LB, k = 1,2,\dots,n \quad (18)$$

RSA is divided into two stages: exploration (encircling) and exploitation (hunting). During exploration, crocodiles exhibit two distinct movement patterns: high walking and belly walking. Unlike focused hunting, exploratory behaviors allow for a wider range of search areas. RSA emulates these mechanisms during its exploration phase, as defined in Eq. 19.

$$x_{i,k}(t+1) = \begin{cases} Best_k(t) - \eta_{i,k}(t) \times \beta - R_{i,k}(t) \times rd, & t \leq \\ Best_k(t) - x_{r1,k}(t) \times ES(t) \times rd, & t \leq 2\frac{T_m}{4}\ a \end{cases} \quad (19)$$

$Best_k$ represents the best k^{th} position of the optimal individual, *T* is the current iteration, and T_m is the maximum iteration. Parameter β controls the exploration extent. Random numbers *rd* and *r1* are employed for stochasticity. The hunting operator, $\eta_{i,k}$, calculated in Eq. 20, influences the exploration process.

$$\eta_{i,k} = Best_k(t) \times P_{i,k}(t) \quad (20)$$

To refine the search, a reduced function, Ri,k, is introduced in Eq. 21, where ε is a small constant and *r2* is another random number. This function minimizes the search space. The evolutionary phase, *ES(t)*, calculated in Eq. 22, is a probability ratio that fluctuates between -2 and 2 over iterations, guided by the random number *r3*. $P_{i,k}$, the percentage difference between the $Bestk^{th}$ value of the optimal and current solutions, as computed in Eq. 23. The average solution, calculated in Eq. 24, contributes to the overall exploration strategy.

$$R_{i,k} = \frac{Best_k(t) - x_{r2,k}}{Best_k(t) + \varepsilon} \quad (21)$$

$$ES(t) = 2 \times r_3 \times \left(1 - \frac{1}{T_m}\right) \quad (22)$$

$$P_{i,k} = \alpha + \frac{x_{i,k-M_{x_i}}}{Best_k \times (UB_k - LB_k) + \epsilon} \quad (23)$$

$$M_{x_i} = \frac{1}{D}\sum_{k=1}^{D} x_{i,k} \quad (24)$$

Cooperation and coordination are the two main foraging behaviors of crocodiles. These strategies represent different approaches to intensified exploitation. Eq. 25 determines the specific behavior. Hunting coordination is performed when the current iteration (*t*) falls within the range of $2T_m/4$ to $3T_m/4$, where T_m is the maximum iteration. Conversely, hunting cooperation occurs when *t* is between 0 and $T_m/4$ or $3T_m/4$ and T_m.

$$x_{i,k} = \begin{cases} Best_k(t) \times P_{i,k}(t) \times rd, & t \leq 3\frac{T_m}{4} \\ Best_k(t) - \eta_{i,k}(t) \times \epsilon - R_{i,k} \times rd, & t \leq T_m\ a \end{cases} \quad (25)$$

C. *Integration of ACO and RSA*

Because ACO relies on pheromone trails, exploration of the solution space is hampered by premature convergence. While strong exploration prevents local optima, excessive exploitation impairs solution quality. RSA effectively balances exploration and exploitation and demonstrates superior performance in various technical areas. We propose a hybrid ACO-RSA strategy based on a high-level relay hybrid (HRH) approach to further improve this balance. ACO and RSA are applied sequentially, homogeneously, or heterogeneously. The proposed method uses a heterogeneous HRH strategy to combine the exploitation of ACO with the exploration of RSA.

Initially, ACO, RSA, and shared parameters are initialized. *N* candidate solutions, each represented by an M-dimensional feature vector, are randomly generated within the range of -1 to 1. These solutions are evaluated using a fitness function to assess their quality relative to previous iterations. Superior solutions are retained, while inferior ones are discarded.

Afterward, the optimal solution is identified and assigned to the initial ACO population. The candidate solutions are then designated as initial ant paths. The ants update candidate solutions based on a subset of features initialized with pheromone values exceeding 0.5. Improvements are based on fitness evaluations, with updates only applied to the fittest solutions (Eq. 26).

$$x_i(g+1) = \begin{cases} x_i^{new}(g), & if\ FF\big(x_i(g)\big) > FF(x_i(g+1)) \\ x_i(g), & else \end{cases} \tag{26}$$

ACO or RSA uses the refined candidate solutions as input to explore new promising regions in successive iterations. An algorithm switch occurs when ACO does not improve the solutions, indicating a trap in local optima. RSA is then used to diversify the search space. The iterative process continues until the termination criterion (maximum iterations) is reached.

Initialization is an initial process that randomly generates candidate solutions (using 50 population sizes) to ensure a diverse solution space. We set pheromone levels for ACO to be initialized with small positive values, and we made the pheromone evaporation rate 0.1 (to balance exploration and exploitation). Inspired by the reptile-hunting dielectric technique, the RSA reacts to evolve optimal solutions and seeks convergence of its best solutions. The algorithm stops when 100 iterations are achieved or there hasn't been a significant improvement in loss over ten consecutive iterations. Incorporating ACO exploration with RSA adaptive learning compensates for both suboptimal aspects of the algorithms, avoids premature convergence typical of ACO, and optimizes the exploitation phase of RSA. During task scheduling, each task's priority is assigned based on its scale and complexity (energy-sensitive or latency-sensitive) through the optimization process, as shown in Fig. 2. The system's energy consumption and latency are reduced by balancing the distribution of tasks across available resources, including edge servers or the cloud.

V. Experimental Setup and Results

To validate the effectiveness of the proposed ACO-RSA algorithm, we conducted an extensive series of simulations. The simulation environment is designed to mimic real-world IoT scenarios with varying numbers of mobile user devices and data transfer rates. The environment includes multiple edge servers with different computational capabilities, simulating a typical edge computing context. The primary performance metrics evaluated in the simulations are average time delay and energy consumption. Time delay is the total time required for task completion, while energy consumption is calculated based on the power needed for data transmission and processing.

The simulation parameters were configured according to the 3GPP standard. The simulation considers an area with a 1 km radius, where users transmit data at a maximum power of 25 dBm over a system bandwidth of 15 MHz and user bandwidth of 0.5 MHz. The noise power in the uplink bandwidth is set at −108 dBm. Tasks have input data volumes ranging from 300 to 1500 KB and require CPU cycles between 0.1 and 0.8 GHz. The users' computing power varies from 0.2 to 1.2 GHz, while the MEC server has a computing power of 3 GHz. The simulation also involves a population size of 50 and a maximum of 100 evolutions, with a maximum allowable time delay of 3 to 5 seconds.

To evaluate the convergence of the proposed computation offloading algorithm, a simulation was conducted with 80 users, 15 servers, and 5MB tasks. Fig. 4 illustrates the time delay of the system over different iteration numbers. Fast convergence was observed after 20 iterations, with minimal delay improvements after that, indicating the achievement of a global optimum. This demonstrates the algorithm's strong global optimization and search capabilities and reduces the overall system delay from 0.268 s to 0.194 s, representing a significant performance improvement.

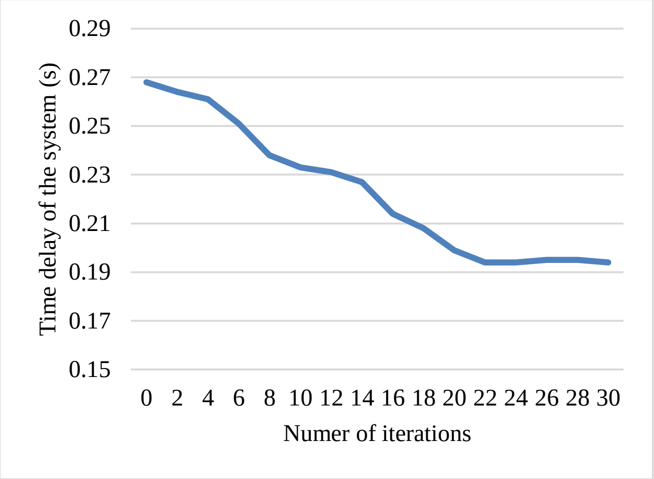


Fig. 4. Stability diagram.

Comparative analyses used algorithms from [26, 27] to assess delay and energy consumption. Fig. 5 shows the average delay for different numbers of users. Energy consumption was evaluated at various transmission rates (Fig. 6). Due to the lack of data transfer, local execution was independent of the transfer rate. All algorithms showed lower energy consumption with increasing transfer rates due to lower offloading overhead. The proposed method showed the largest decrease and lowest overall energy consumption, enabled by more frequent and fine-grained discharge decisions.

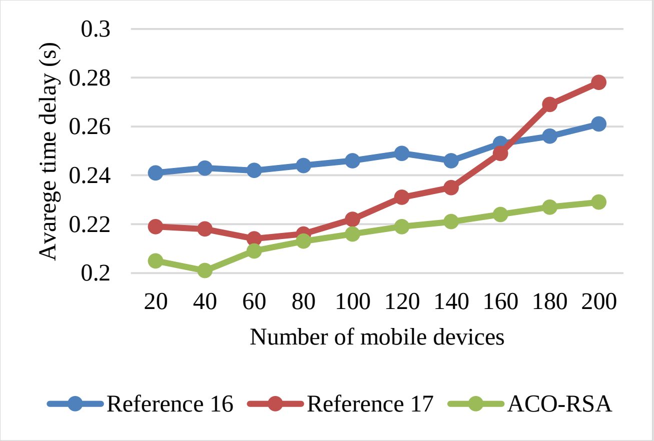


Fig. 5. Delay vs. Number of mobile devices.

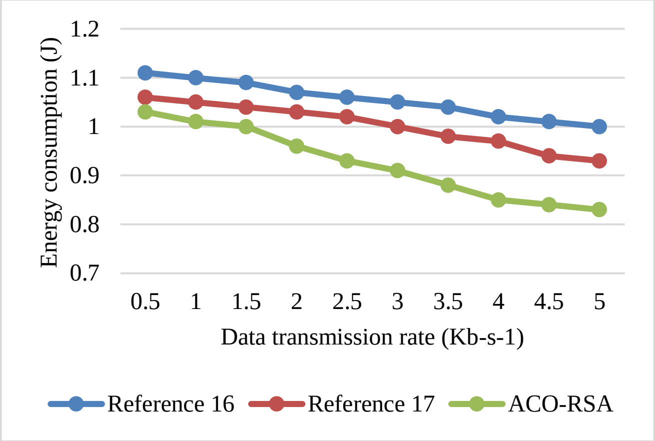


Fig. 6. Energy consumption vs. data transmission rate.

Fig. 7 illustrates the relationship between energy consumption and number of users. The proposed method consistently outperformed others with the lowest energy consumption and growth rate, achieving 0.2 J for 80 users. The local execution caused the highest energy consumption. The proposed strategy's efficiency in task offloading and reduced local execution contributed to overall energy savings.

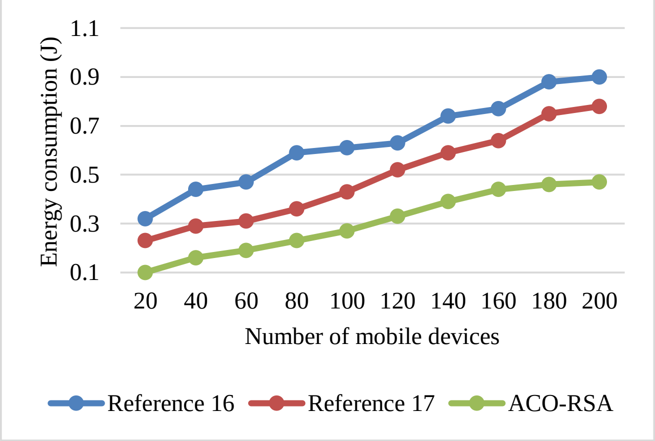


Fig. 7. Energy consumption vs. number of mobile devices.

The results confirm that the ACO-RSA algorithm can effectively solve major task offloading problems in IoT edge computing, such as energy consumption and latency. By combining ACO's exploration and RSA's adaptive learning, the proposed approach distributes tasks efficiently and minimizes system overhead. The proposed ACO-RSA provides an average of 27.6% energy savings and 25.4% latency reduction compared to other existing techniques, resulting in better resource scheduling and task completion.

These results confirm the algorithm's ability to optimize task shifting while balancing energy and performance metrics. Nevertheless, the study emulates the simulation environment of static network states and operational case functions within devices, which may not correspond to the realistic conditions in ideas in IoT environments. Fluctuating network traffic or device roaming can affect the algorithm's performance. Finally, future work should focus on extending ACO-RSA to dynamic edge computing scenarios and incorporating other considerations, such as security and privacy, to improve the robustness and applicability of ACO-RSA in various IoT contexts.

VI. Conclusion

To the best of our knowledge, this is the first study to introduce a new metaheuristic algorithm by combining ACO and RSA into a hybrid version named ACO-RSA, which can overcome the challenges of efficient task offloading in IoT environments powered by edge computing. This feature, therefore, allows the ACO-RSA classifier to make a remarkable compromise between energy consumption and latency, two main components of IoT ecosystems, by utilizing both ACO and RSA within an integrated classification framework. By using ACO, the algorithm takes inspiration from how ants collect food to search for optimal paths; meanwhile, with RSA, it can adaptively explore the IoT environments that are highly dynamic and unpredictable, proposing a strong solution.

ACO-RSA is designed based on an optimization objective function aiming to minimize energy consumption while at the same time reducing task offloading latency. Thus, it is well suited for resource-limited IoT devices that run in edge computing environments. We conducted a series of extensive simulations to verify the feasibility of the proposed algorithm. Results confirmed that the proposed ACO-RSA operates better than the traditional benchmark algorithms. The hybrid algorithm minimizes energy consumption and operates with low latency as the number of mobile users and data rates increase. This result illustrates the promise of ACO-RSA in enhancing resource utilization and prolonging the lifecycle of IoT applications by tuning task offloading policies.

Future work will address the scalability and robustness of ACO-RSA, especially within larger and more complex IoT networks with an assorted range of applications and diverse task requirements. Moreover, investigation into the integration between machine learning and ACO-RSA could enhance adaptivity capabilities even more and aid in better addressing dynamic changes in network conditions and resource availability. In conclusion, the proposed ACO-RSA algorithm is an important step towards an efficient and energy-greedy task offloading strategy in edge computing-enabled IoT environments.

References

[1] A. A. Anvigh, Y. Khavan, and B. Pourghebleh, "Transforming Vehicular Networks: How 6G can Revolutionize Intelligent Transportation?," Science, Engineering and Technology, vol. 4, no. 1, pp. 80-93, 2024.

[2] J. Valizadeh et al., "An operational planning for emergency medical services considering the application of IoT," Operations Management Research, vol. 17, no. 1, pp. 267-290, 2024.

[3] B. Pourghebleh and V. Hayyolalam, "A comprehensive and systematic review of the load balancing mechanisms in the Internet of Things," Cluster Computing, pp. 1-21, 2019.

[4] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017.

[5] N. M. Quy, L. A. Ngoc, N. T. Ban, N. V. Hau, and V. K. Quy, "Edge computing for real-time Internet of Things applications: Future internet revolution," Wireless Personal Communications, vol. 132, no. 2, pp. 1423-1452, 2023.

[6] Q. Wu, S. Wang, H. Ge, P. Fan, Q. Fan, and K. B. Letaief, "Delay-sensitive task offloading in vehicular fog computing-assisted platoons," IEEE Transactions on Network and Service Management, 2023.

[7] S. E. Comert and H. R. Yazgan, "A new approach based on hybrid ant colony optimization-artificial bee colony algorithm for multi-objective electric vehicle routing problems," Engineering Applications of Artificial Intelligence, vol. 123, p. 106375, 2023.

[8] L. Abualigah, M. Abd Elaziz, P. Sumari, Z. W. Geem, and A. H. Gandomi, "Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer," Expert Systems with Applications, vol. 191, p. 116158, 2022.

[9] R. Chataut, A. Phoummalayvane, and R. Akl, "Unleashing the power of IoT: A comprehensive review of IoT applications and future prospects in healthcare, agriculture, smart homes, smart cities, and industry 4.0," Sensors, vol. 23, no. 16, p. 7194, 2023.

[10] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single - objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022.

[11] B. Pourghebleh, N. Hekmati, Z. Davoudnia, and M. Sadeghi, "A roadmap towards energy - efficient data fusion methods in the Internet of Things," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, p. e6959, 2022.

[12] F. Kamalov, B. Pourghebleh, M. Gheisari, Y. Liu, and S. Moussa, "Internet of medical things privacy and security: Challenges, solutions, and future trends from a new perspective," Sustainability, vol. 15, no. 4, p. 3317, 2023.

[13] N. A. Angel, D. Ravindran, P. D. R. Vincent, K. Srinivasan, and Y.-C. Hu, "Recent advances in evolving computing paradigms: Cloud, edge, and fog technologies," Sensors, vol. 22, no. 1, p. 196, 2021.

[14] G. Baranwal, D. Kumar, and D. P. Vidyarthi, "Blockchain based resource allocation in cloud and distributed edge computing: A survey," Computer Communications, 2023.

[15] B. Pourghebleh, K. Wakil, and N. J. Navimipour, "A comprehensive study on the trust management techniques in the Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9326-9337, 2019.

[16] J. Ge, B. Liu, T. Wang, Q. Yang, A. Liu, and A. Li, "Q - learning based flexible task scheduling in a global view for the Internet of Things," Transactions on Emerging Telecommunications Technologies, p. e4111, 2020.

[17] P. V. B. C. d. Silva, C. Taconet, S. Chabridon, D. Conan, E. Cavalcante, and T. Batista, "Energy awareness and energy efficiency in internet of things middleware: a systematic literature review," Annals of Telecommunications, vol. 78, no. 1, pp. 115-131, 2023.

[18] I. Vlachos, R. M. Pascazzi, M. Ntotis, K. Spanaki, S. Despoudi, and P. Repoussis, "Smart and flexible manufacturing systems using Autonomous Guided Vehicles (AGVs) and the Internet of Things (IoT)," International Journal of Production Research, vol. 62, no. 15, pp. 5574-5595, 2024.

[19] K. Xiao, Z. Gao, W. Shi, X. Qiu, Y. Yang, and L. Rui, "EdgeABC: An architecture for task offloading and resource allocation in the Internet of Things," Future generation computer systems, vol. 107, pp. 498-508, 2020.

[20] Q. You and B. Tang, "Efficient task offloading using particle swarm optimization algorithm in edge computing for industrial internet of things," Journal of Cloud Computing, vol. 10, pp. 1-11, 2021.

[21] Y. Chen, W. Gu, and K. Li, "Dynamic task offloading for internet of things in mobile edge computing via deep reinforcement learning," International Journal of Communication Systems, p. e5154, 2022.

[22] W. Kong, X. Li, L. Hou, J. Yuan, Y. Gao, and S. Yu, "A reliable and efficient task offloading strategy based on multifeedback trust mechanism for IoT edge computing," IEEE Internet of Things Journal, vol. 9, no. 15, pp. 13927-13941, 2022.

[23] Z. Aghapour, S. Sharifian, and H. Taheri, "Task offloading and resource allocation algorithm based on deep reinforcement learning for distributed AI execution tasks in IoT edge computing environments," Computer Networks, vol. 223, p. 109577, 2023.

[24] M. Bolourian and H. Shah-Mansouri, "Energy-efficient task offloading for three-tier wireless-powered mobile-edge computing," IEEE Internet of Things Journal, vol. 10, no. 12, pp. 10400-10412, 2023.

[25] P. K. Nandi, M. R. I. Reaj, S. Sarker, M. A. Razzaque, M. Mamun-or-Rashid, and P. Roy, "Task offloading to edge cloud balancing utility and cost for energy harvesting internet of things," Journal of Network and Computer Applications, vol. 221, p. 103766, 2024.

[26] M. Zhao and K. Zhou, "Selective offloading by exploiting ARIMA-BP for energy optimization in mobile edge computing networks," Algorithms, vol. 12, no. 2, p. 48, 2019.

[27] Y. Shi, Y. Xia, and Y. Gao, "Cross-server computation offloading for multi-task mobile edge computing," Information, vol. 11, no. 2, p. 96, 2020.

Bubble Detection in Glass Manufacturing Images Using Generative Adversarial Networks, Filters and Channel Fusion

Md Ezaz Ahmed[1], Mohammad Khalid Imam Rahmani[2], Surbhi Bhatia Khan[3]
College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia[1, 2]
Department of Information Systems-College of Computer Science and Information Technology,
King Faisal University, Saudi Arabia[3]
School of Science-Engineering and Environment, University of Salford, United Kingdom[3]

***Abstract*—With the increasing production of glassware products, the detection of bubble defects has been of vital importance. The manual inspection of glass bubble defects is considered to be tedious and inefficient way due to the increasing volume of images, and the high probability of human error. Computer vision-based methods provide us with a platform for automating the bubble defect detection process which can overcome the disadvantages associated with manual inspection thereby significantly reducing the cost and improving the quality. To address these issues, we propose an integrated deep learning (DL) based bubble detection algorithm, in which an image data set is prepared using a Generative Adversarial Network (GAN). The proposed algorithm exploits the Information-Preserving Feature Aggregation (IPFA) module for achieving semantic feature extraction by maintaining the small defects' internal features. To weed out irrelevant information due to fusion, the proposed research introduces the Conflict Information Suppression Feature Fusion Module (CSFM) to further advance the component combination methodology, the Fine-Grained Conglomeration Module (FGAM) is employed to facilitate cooperation among feature maps at various levels. This approach mitigates the generation of conflicting information arising from erroneous features. The algorithm improved performance with an accuracy rate of 0.677 and a recall rate of 0.716 with a precision value of 0.638.**

Keywords*—*Computer vision; Generative Adversarial Network; Information-Preserving Feature Aggregation; Conflict Information Suppression Feature Fusion Module; Fine-Grained Aggregation Module; deep learning

I. Introduction

The glass production process inherently yields defects like bubbles and inclusions due to the limitations of current techniques. These defects vary in impact depending on the applications. For instance, while bubbles may not greatly affect household glass, they can significantly compromise in big applications such as car safety glass [1-3]. Till now only the manual traditional methods are being used for the identification of glass defects which is very cumbersome and in ideal conditions not possible to detect the present bubbles. Thus, identifying these defects becomes crucial in some automation fashion [4]. To mitigate costs and enhance product quality, employing computer vision for defect detection has gained traction. Glass, with its lack of pattern, monochrome, and transparency, lends itself well to computer vision inspection. CNN-based detection methods rooted in deep learning are rapidly developing and becoming a significant area of intensive research [5-8]. To address the aforementioned challenges, this paper presents targeted enhancements in two key areas to improve small bubble's defect detection performance in camera-captured low-resolution images, particularly in complex environments. A traditional feature aggregation method called stride convolution causes feature information loss. The proposed IPFA module handles this loss and completes feature aggregation using the techniques of splitting and reassembling across different dimensions, facilitating information organization along the channel dimension. This enables the construction of a robust semantic feature representation while maintaining the natural features, thereby replacing generic methods for feature aggregation like stride convolution. Moreover, CNN-based detection methods that depend on DL are developing quickly and are already the subject of much research.

In many industrial applications, such as material inspection and quality control, bubble detection is an essential duty. Product dependability and integrity are guaranteed by the accurate identification of bubble flaws. Low-resolution photos and complicated backdrops are common problems for traditional approaches, which might result in missed detections or false positives. By improving the feature extraction procedure, the suggested IPFA module overcomes these drawbacks and allows for a more accurate diagnosis of bubble flaws. Our method gains from the capability of directly learning intricate patterns and representations from the data by utilizing deep learning. This paper demonstrates the effectiveness of the IPFA module in improving detection accuracy and robustness, paving the way for more reliable industrial inspection systems. Additionally, the advancements in CNN-based techniques highlight the potential for continuous improvements in defect detection, contributing to higher standards in manufacturing processes.

Various researchers have extensively explored different AI-driven techniques for segmenting and reconstructing overlapping bubbles in images of bubbly flow. Specifically, they have evaluated and implemented three distinct CNN models: StarDist and Mask RCNN, both open-source solutions, along with a hybrid of two slightly modified UNets. Generally,

*Corresponding Authors

these methods show proficiency in identifying bubbles under optimal conditions, such as adequate lighting that improves the visibility of intersections between overlapping bubbles, relatively uniform bubble shapes, and a manageable number of overlapping segments. Most of these studies relied on hardware requirements, necessitating a stepper motor to rotate the disturbance and a camera to capture the entire sparse area of the bubbles. This approach was quite expensive and demanded extensive expertise in embedded systems. The proposed method integrates and amalgamates various features taken from the feature space. It extract accurate and comprehensive information. The model recombines features undergoing multi-level interaction in the channel dimension to achieve feature aggregation, enabling the methodology of semantic feature representations without compromising their natural traits.

The remaining structure of the paper is as follows: Section II elaborates a review of literature in the domain. Section III outlines the framework of the research conducted, while Section IV elaborates its implementation and presents a comprehensive results analysis. In the end, Section V concludes the research followed by future work.

II. Literature Review

Computer vision-based systems with similar applications have already been implemented in various industrial sectors. One of the prominent industries employing such systems is the glass manufacturing industry [10-13]. Manufacturing glassware products may exhibit defects like scratches, cracks, impurities, dip stones, and bubbles. Such defects associate security risks along with appearance anomalies. For instance, exceeding the thermal expansion coefficient of the crack defect can cause radial cracks or the glassware can even burst in high-temperature environments [14-18]. Therefore, defect detection is an essential process for glassware products. Despite state-of-the-art glassware industries, the detection of defects is still being done with human supervision which is inefficient given the volume of production units and is highly susceptible to subjective factors of the supervisor [19]. Computer vision-based systems for detecting defects in glassware products can eliminate the problems of subjectivity and inefficiency [20-23]. Currently, vision-based systems are good at detecting defects such as black spots, stones, bubbles, cracks, scratches, and so forth on the abstract level. However, these systems still struggle to detect the anomalies from the regions which are homogeneous. Various defects in manufactured glass sheets are foreign material, low-contrast defect regions, scratches and spots, bubbles, inclusions, and holes. Most of the studies mainly focus on scratches, foreign material, and bubbles as they are defects that can cause severe harm to the quality of the product.

An unmelted opaque material with the appearance of a lump is referred to as a foreign material. Irregular marks or patches on the glass surface are considered spots or scratches [24]. These defects are mainly caused by transportation. The bubble defect is like an air bubble trapped in the glass during glassware production. Several studies have been proposed to detect these defects from glass images using machine vision techniques [25]. Makoto et al. emphasized detecting the foreign materials from the LCD scanned under fan-beam laser light. The defect detection method was based on the light section method. Chang-Hwan et al. employed model-fitting and conventional least-squares estimators to detect the low-contrast region defects. The idea was to approximate the outliers by estimating the image background. Adamo et al. proposed an inline visual inspection system to detect the defects in the glass surface. They employed a canny edge detection method with empirical thresholds to detect scratches and spots. Zhao et al. proposed the canny edge detection method, the Otsu method, binary feature histogram, and adaptive boosting method for detecting bubbles from glass images. Recent advancements in Generative Adversarial Networks (GANs) are increasing and have many advantages in offering solutions with DL algorithms in the industry [26]. Many researchers have proposed novel methods using DL algorithms to improve the bubble defects and make them more robust and efficient. The authors have introduced a method based on DL, which efficiently trains the neural network (NN) on a reduced dataset of more precise calculations through transfer learning. This is achieved by exploiting a crystal graph NN trained on a larger dataset with reduced accuracy but better speed [27-29]. Furthermore, the researchers have explored various AI-based approaches for segmenting and reconstructing overlapping bubbles in images of bubbly flow. Specifically, they have assessed and implemented three different CNN architectures: StarDist and Mask-RCNN, both of which are open-source techniques, and a hybrid of two slightly adapted UNets [30-34]. In general, all three methods demonstrate the ability to detect bubbles under favourable conditions, such as sufficient lighting that enhances the visibility of intersections between overlapping bubbles, a relatively uniform bubble shape, and a manageable number of overlapping bubble segments.

In many industrial applications, such as material inspection and quality control, bubble detection is an essential duty. The integrity and dependability of products are ensured by the accurate identification of bubble flaws. Low-resolution photos and complicated backdrops are common problems for traditional approaches, which might result in missed detections or false positives. By improving the feature extraction procedure, the IPFA module overcomes these drawbacks and makes it possible to identify bubble faults with greater accuracy. Our method takes advantage of deep learning (DL) to learn hidden patterns and their representations straight from the dataset. The efficiency of the IPFA module enhances detection robustness and accuracy which is demonstrated in this research, opening the door for more dependable industrial inspection systems. Moreover, developments of CNN-based methods demonstrate the possibility of ongoing enhancements in fault identification, resulting in improved standards for the manufacturing process. Several studies have been conducted for finding the difficulties associated with bubbles detection in low-resolution photographs. Conventional image processing methods, including thresholding and edge detection, often fail because they can't deal with complicated backdrops and different lighting situations. The use of DL and machine learning (ML) techniques to resolve these issues has been investigated in recent studies. Convolutional neural networks (CNNs) have been used in a range of image identification tasks, such as defect detection. Research suggests that combining CNNs with multi-scale feature extraction methods will greatly enhance detection performance. Methods such as the Feature

Pyramid Network (FPN) and its variants have been utilized to improve object detection on various scales. By maintaining crucial feature information throughout the aggregation process, the proposed IPFA module improves on existing developments by ensuring that minute features necessary for identifying bubble faults are not overlooked. To further improve performance, CNN designs now include attention methods in addition to multi-scale feature extraction. By training the network to focus on only the desired portions of the image, attention modules improve the network's ability to detect minute defects. Adding attention methods to the IPFA module can result in detection systems that are even more accurate and dependable.

New opportunities for automated quality control have been created by the use of DL-based approaches in industrial applications. Research has shown that DL models can achieve more accuracy and efficiency than conventional techniques. The IPFA module is a flexible solution for a range of defect detection jobs due to its better integration with CNN architectures. The effectiveness of CNN-based techniques for industrial inspection is supported by experimental findings from investigations. Studies have demonstrated that CNNs can identify flaws in metals, polymers, and textiles with higher accuracy. These capabilities are improved by modules like IPFA, achieving even higher accuracy standards. The availability of rich datasets for model training and the ongoing growth of DL frameworks are the factors for more contribution to the field's advancement. By exploiting these advancements, the IPFA module promises a reliable and expandable industrial defect detection system. In conclusion, the suggested IPFA module compounded with DL capabilities enhances the task of industrial inspection and bubble detection with better accuracy and dependability in defect identification by resolving the drawbacks of conventional techniques. The ongoing advancements in CNN-based techniques and their application in industrial settings hold great potential for further enhancing the quality and efficiency of manufacturing processes.

Most of the works were based on hardware requirements like there was a need for the stepper motor to rotate the turbulence to rotate the camera to capture the entire defective area of the bubbles. This method was quite expensive and required an extensive area of specialization in the embedded system field. Further with the advancement of the DL network, the models that were trained were limited to the high-resolution images but in reality, the low-resolution images were not considered. Most of the bubble defects have low-resolution images which needs to be considered in the work along with that the overlapping bubbles create more defects which was neglected in the literature.

III. Materials and Methods

In comparison to the previous methods the proposed algorithm makes it more suitable for defect detection in glass bottle images. Fig. 1 [36] demonstrates the structure of the proposed model.

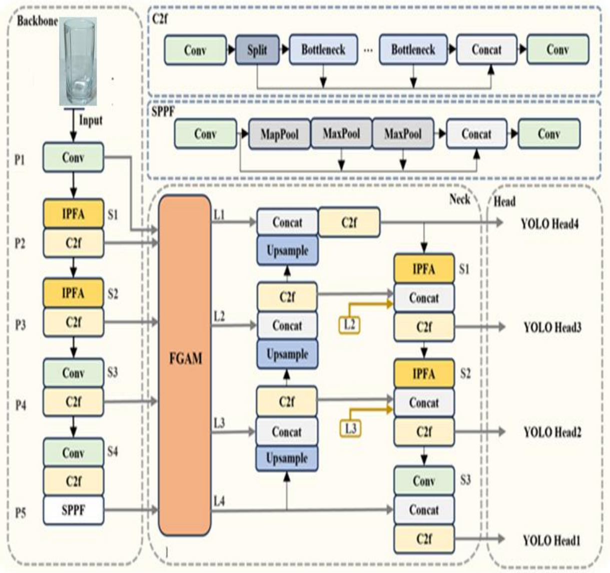


Fig. 1. Model’s structure for the processing of the glass image.

A. IPFA Module

Feature aggregation refers to the process of integrating and combining multiple characteristics from the feature space to obtain more precise and comprehensive data. However, widely available feature-aggregation techniques like pooling and stride convolution can significantly impair deep neural network detection performance due to feature information loss. Stride convolution increases the convolutional kernel's stride, reducing the size of output feature map and increasing the size of the receptive field. This technique aggregates the input feature information in the spatial dimension, compressing feature information for bubble defects. Conversely, the pooling method prepares four spatially separated sub-features from the features, retaining only a portion and discarding the rest, potentially losing important information. To address this issue, the research work proposes an IPFA module that splits and recombines features in the spatial and channel dimensions. This module achieves feature aggregation by recombining features that undergo multi-level interaction in the channel dimension, enabling the semantic feature extractions of input data without compromising its natural characteristics. Fig. 2 illustrates the precise construction of the IPFA module [36].

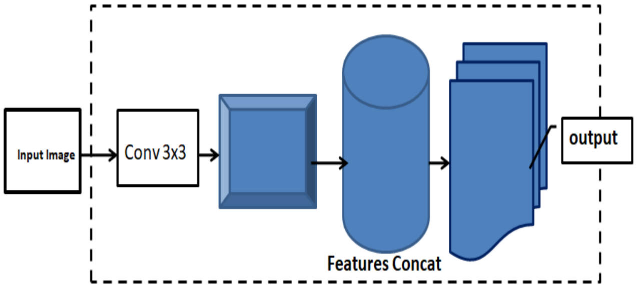


Fig. 2. The structure of IPFA module.

Initially, for the receptive field size expansion, the IPFA module employees a 3×3 convolution while maintaining the original size of output feature map. A 3×3 depth-wise separable convolution is then utilized in the neck to lower the number of parameters. Subsequently, the extracted features are separated into both spatial and channel dimensions, giving in eight categories of sub-features. These sub-features concatenated in the channel dimension. At the end, in the channel dimension, we employ a 1×1 convolution for information interaction among the concatenated features.

The IPFA module has been implemented for both the spine and neck areas of the glass images. This module replaces the traditional stride convolution method and achieves better feature aggregation. Compared to conventional feature extraction methods that use stride convolution or pooling operations, the implemented module in the research maintains information contained in the natural feature of the input sample thereby improving the model's performance.

B. *CSFM Module*

In the process of feature integration, the "concat" or "add" operations are typically employed to combine different levels' feature maps. However, merging them with only the default weights can lead to significant redundancy and contradicting information, resulting in semantic alterations in the current layer and creating defect regions which can be simply overshadowed by the background. To address this issue, the Channel and Spatial Feature Modulation (CSFM) module is proposed. This will remove contradicting information in the integration task, preventing the features of the defective area from being overshadowed. There are two parallel branches in the CSFM module: the Channel Conflict Information Suppression Module (CCSM) and the Spatial Conflict Information Suppression Module (SCSM). This dual approach ensures that both types of conflicts are minimized, thereby enhancing the identification of defect features. Additionally, the integration of the CSFM module within DL frameworks significantly enhances the capability to distinguish defect features from background noise. This improvement is crucial for applications requiring high precision, such as quality control in manufacturing and material inspection. The dual-branch structure of the CSFM not only refines feature representation but also adapts to various types of input data, making it versatile across different defect detection scenarios.

Since, simple fusion is the rational approach which works on the principle of just adding the information whatever it is getting with the default weights. In the processing of the pixel maps from different layers, many of the pixels are repeated and redundant which must be filtered out to optimize the process.

By addressing both channel and spatial conflicts, the CSFM module ensures that the features extracted are both distinct and relevant, thus preventing the dilution of critical defect information. More accurate feature aggregation is possible due to the adaptive pooling in the CCSM, which enables dynamic modification based on the input data whereas, the convolutional method of the SCSM provides fine-grained spatial attention, identifying minute differences that might point to flaws. The CSFM module is better than conventional feature integration techniques. It offers greater accuracy and resilience for identification of minute and subtle flaws. Defect detection systems perform much better when they can reduce conflicting data while maintaining the integrity of feature information. This innovative method promises more dependable and effective inspection procedures in a range of industrial applications, establishing a standard in the field. Higher standards and better productivity in production settings result from the integration of CSFM compounded with better automated quality control systems and improved fault detection efficacy.

Moreover, the CSFM can easily integrate multi-level features in a more balanced and efficient manner to improve defect detection accuracy. The CSFM's unique design is shown in Fig. 3, which emphasizes the system's capacity to mitigate contradictory input while preserving feature integrity [36]. Applying the CSFM module into DL frameworks highly improves the ability to isolate defect features from noise. For material inspection and manufacturing quality control applications which need extreme precision, this improvement is essential. The CSFM's dual-branch structure makes it flexible for a range of defect detection scenarios by improving feature representation and accommodating diverse kinds of input data. The CSFM module does not allow critical defect information to be diluted by handling both channel and spatial conflicts to guarantee that the characteristics collected are different and meaningful. Even more accurate feature aggregation is possible by using the adaptive pooling in the CCSM, which enables dynamic modification based on the input data. On the other hand, the convolutional method of the SCSM provides fine-grained spatial attention to identify minute differences that might point to flaws. The experimental data demonstrate that the CSFM module performs better than conventional feature integration techniques and offers higher accuracy and resilience in identifying minute and subtle flaws. Defect detection systems perform much better when they can reduce conflicting data while maintaining the integrity of feature information. This novel method promises more dependable and effective inspection procedures in a range of industrial applications, setting a new standard in the sector.

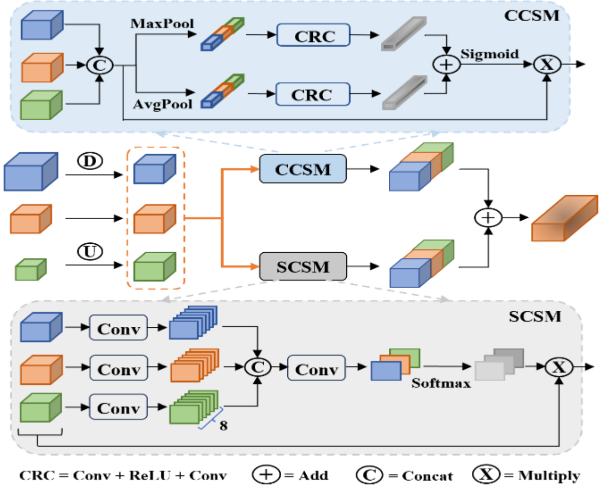


Fig. 3. Structure of the CSFM.

presents the formula for this procedure.

For each X = Downsample(X) and Z = Upsample(Z), (1)

Bilinear interpolation is utilized to implement the up-sample operation, while stride convolution is used to iterate the entire glass image pixel to discover the bubble section, which is around 0.05 mm percent of the entire area, to implement the down-sample action. The output can be presented as:

$$OC = \sigma[AP(\{X', Y, Z'\}) + MP(\{X', Y, Z'\})] \times \{X', Y, Z'\} (2)$$

where {} denotes the concat operation, σ represents the Sigmoid operation, × denotes element-wise multiplication, and the operations AP stands for average pooling and MP for max pooling. The weights are added along the spatial dimension before passing through a Sigmoid activation function for generating the channel-wise adaptive weights.

First, the two input feature maps XXX and ZZZ are resized to kernel sizes of 3×3 and 1×1, respectively. It is critical for the feature maps normalization with the channel dimension to perform the Softmax operation (SM) [35]. Three output feature maps, each with eight channels, are produced after convolving each input feature map with a 3×3 kernel. The Concat technique is then used to concatenate these feature maps. The number of channels is decreased to three by applying a 1×1 convolution. The correspondence with the supplied feature maps is maintained. By combining the features from various scales or resolutions, multi-scale feature fusion takes the benefits of both deep and shallow feature maps to produce a better and more complete feature representation. This improves the feature fusion outcomes. Nevertheless, there may occur semantic conflicts if the information of different densities is directly merged, which will limit the multi-scale features expression. The research suggests merging of feature maps from the current level and neighboring levels in the backbone to address this problem. To improve the effectiveness of multi-scale feature fusion, an Adaptive Feature Refinement (AFR) module is introduced. This module harmonizes the feature maps integration from different scales thereby reducing semantic conflicts and maintaining the integrity of multi-scale information. The AFR module dynamically adjusts the weights of feature maps based on their significance, bringing a balanced and coherent fusion process. The AFR module uses attention mechanisms to prioritize the most relevant features during the fusion process. By focusing on only critical features and suppressing irrelevant ones, the AFR module enhances the representation of important details, leading to more accurate and robust feature extraction. This approach not only addresses the issue of semantic conflicts but also boosts the overall performance of the feature fusion process. Moreover, the AFR module employs a multi-resolution strategy, which processes feature maps at various resolutions to capture both global and local contexts. This strategy ensures that the fused feature maps maintain essential rich and diverse information for accurate detection and recognition tasks. The combination of attention mechanisms and multi-resolution processing makes the AFR module one of the powerful tools for improving multi-scale feature fusion. Experimental results demonstrate that the AFR module outperforms traditional feature fusion methods significantly. By resolving semantic conflicts and enhancing feature representation, the AFR module brings higher accuracy and better efficiency in applications like image segmentation, object detection and recognition tasks. Therefore, the proposed method not only sets a new benchmark in multi-scale feature fusion but also gives a way to progress in the field of DL and computer vision. Overall, the integration of the AFR module into existing frameworks enhances the robustness and reliability of feature extraction, making it a valuable addition to state-of-the-art techniques. This innovative approach promises to enhance the performance of various ML models, contributing to more precise and efficient solutions in diverse domains.

This fusion process is performed in advance to mitigate the differences in information density i.e. the non-defected part with the defective one between the feature maps. Additionally, CSFM is employed to exploit an attention mechanism to remove the conflicting information.

By incorporating these techniques, the interference of complex backgrounds on bubble detection is alleviated. The proposed structure is depicted in Fig. 4.

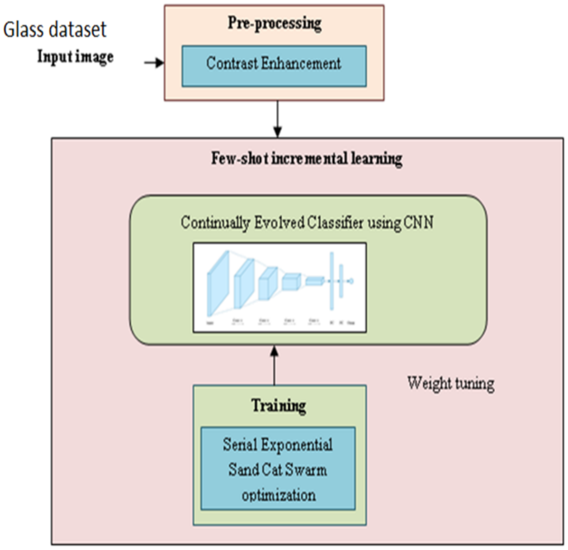

Glass dataset
Input image
Pre-processing
Contrast Enhancement
Few-shot incremental learning
Continually Evolved Classifier using CNN
Weight tuning
Training
Serial Exponential Sand Cat Swarm optimization

Fig. 4. Proposed pseudo model.

C. Fine-Grained Aggregation Module (FGAM)

Most Feature Pyramid Network (FPN) methods directly sample different levels' feature maps to the same size for fusion. However, a significant difference in information compactness between these feature maps often leads to semantic conflicts. This work suggests the FGAM, which applies fine-grained feature aggregation across multiple levels of feature maps, from P0 to P5, in the backbone, to solve this problem. This method brings about considerable information interaction between the various feature map layers. Semantic information rises and detailed information falls as the interaction moves from P0 to P5.

Within FGAM, the CSFM module further filters out conflicting information through spatial and channel attention mechanisms.

Moreover, the FGAM combines fine-grained details and high-level semantics effectively, increasing the quality of feature representation. By using adaptive pooling and multi-level interactions, the module is able to maintain the integrity of both fine and coarse features making it a robust detection of defects. The FGAM architecture facilitates dynamic adjustment of feature weights by optimizing the fusion process for different input scales and conditions. The flexibility of FGAM is useful for applications that require precise defects detection like identification of small bubbles in glass bottles.

FGAM overcomes the drawbacks of traditional FPN methods by:

1) Ensuring fine-grained feature interaction across multiple levels.

2) Balancing spatial and semantic information in output feature maps.

3) Minimizing information density disparities to prevent semantic conflicts.

4) Applying the CSFM to remove conflicting information using attention mechanisms.

5) Enhancing feature representation through adaptive pooling and multi-level interactions.

6) Maintaining the integrity of both fine and coarse features for robust defect detection.

7) Enabling dynamic adjustment of feature weights for various input scales.

8) Providing a flexible architecture suitable for precise detection applications.

This comprehensive approach significantly improves the performance of feature fusion in DL networks, ensuring more accurate and reliable detection outcomes.

D. Fine-Grained Aggregation Feature Pyramid Network

FGAFPN consists of FGAM and the feature pyramid network. FGAM serves as the connection between the Backbone and the pyramid network. It takes input feature maps from the Backbone and outputs feature maps with balanced density information. However, FGAM alone does not possess strong multi-scale feature representation capability and requires further deep fusion through the pyramid network. PANed architecture increases the model’s ability to detect bubbles in the glass, preventing the small defects’ features from being suppressed by conflicting information. This method shows improved integration of spatial and semantic information by lowering semantic conflicts and improving the network's overall performance by integrating FGAM into the multi-scale feature fusion process.

E. Pseudo Incremental Learning Phase

To update classifier weights, a continually evolved classifier is used that involves a classifier adaptation phase. It is learned in every individual session based on the preceding session’s global context.

Proposed Pseudo Algorithm

```
1. Import necessary libraries (e.g., NumPy, TensorFlow)
2. Define the generator model:
  a. Input: Random noise (e.g., Gaussian)
  b. Output: Image of size equal to the input image
  c. Architecture: CNN with Conv2DTranspose, BatchNormalization, LeakyReLU, etc.
3. Define the discriminator model:
  a. Input: Image (real or generated)
  b. Output: Probability that the input is real (between 0 and 1)
  c. Architecture: CNN with Conv2D, BatchNormalization, LeakyReLU, etc.
4. Define the loss functions:
  a. Generator loss: binary cross-entropy loss to streamline the generator to produce realistic images.
  b. Discriminator loss: binary cross-entropy loss to streamline the discriminator to correctly classify the defective image and real images.
5. Define the optimizer:
  a. Adam optimizer with appropriate learning rate and other hyperparameters
6. Training loop:
  a. For each epoch:
i. For each batch of training data:
      1. Train the discriminator:
         a. Generate a batch of fake images using the generator
         b. Calculate the discriminator loss using real and fake images
         c. Update the discriminator weights using the Adam optimizer
      2. Train the generator:
         a. Generate a batch of fake images using the generator
         b. Calculate the generator loss using the discriminator's response to the bubble-defected images
         c. Update the generator weights using the Adam optimizer
7. After training, the generator should be able to generate realistic-looking images, including images of bubbles.
8. To detect bubbles:
  a. Generate a new image using the generator
  b. Use an image processing algorithm (e.g., edge detection, contour detection) to detect bubbles in the generated image
  c. Return the detected bubbles as output
```

F. Binarization Process

Before feature extraction, it is imperative to conduct binarization processing on the acquired foregrounds. While multiple techniques exist for converting grayscale images to binary images, including OTSU and adaptive thresholding, we opt not to utilize conventional methods. Instead, we introduce a pioneering binarization approach in this study. The rationale behind this decision is rooted in the observation that binary results obtained from traditional methods may not yield the same efficacy as our proposed novel method, particularly when dealing with low-resolution images during subsequent feature extraction.

G. Modeling of GAN

GANs with CNNs for the detection of bubbles in glass bottles involve two main components: the generator and the discriminator. Below are the mathematical expressions for each component:

1) Generator: The generator aims to produce realistic images of glass bottles with bubbles.

Let G(z; θg), represent the generator function, where z is the input noise vector and θg are the parameters of the generator network.

The generator takes random noise zas input and generates an image

$$\hat{x} = G(z;\ \theta g), \tag{3}$$

Mathematically, this can be expressed as:

$$\hat{x}=G(z;\ \theta g), \tag{4}$$

2) Discriminator: The discriminator distinguishes between real images of glass bottles with bubbles and fake images i.e. undetected defects generated by the generator. Let D(x;θd) represent the discriminator function, where x is an input image and θd are the parameters of the discriminator network.

The discriminator outputs a probability D(x;θd) indicating the likelihood that the input image x is a real image or generated by the generator. Mathematically, this can be expressed as:

$$D(x;\ \theta d) \tag{5}$$

3) Loss Functions:

a) Generator loss: By producing visuals that are identical to actual ones, the generator hopes to trick the discriminator. Thus, the binary cross-entropy between the discriminator's predictions on generated images and a vector of ones (representing real images) is usually the generator's loss function.

$$\mathscr{L}gen=Ez[\log(1-D(G(z;\ \theta g);\ \theta d))] \tag{6}$$

b) Discriminator loss: The discriminator aims to classify real and fake images correctly. Its loss function is the sum of the binary cross-entropy between its predictions on real images and a vector of ones and the binary cross-entropy between its predictions on generated images and a vector of zeros (indicating defected bubble images).

$$\mathscr{L}disc=-Ex\ [\log D(x;\ \theta d)]-Ez[\log(1-D(G(z;\ \theta g);\ \theta d))] \tag{7}$$

4) Optimization: The parameters θg and θd are updated using gradient descent methods such as Adam optimization to minimize the respective loss functions. These mathematical expressions define GANs' training process with CNNs for the detection of bubbles in glass bottles. The role of generator is to produce realistic images of glass bottles with bubbles, while the discriminator distinguishes between real and generated images. With the help of adversarial training, both networks iteratively improve till the generator generates convincing images and the discriminator cannot effectively differentiate between real and generated images.

Implemented Algorithm

1. Set the initial value of variable ***L*** to the binary representation of the first pixel in the current row (or column). Let ***L[1]*** represent this value, i.e. ***L[1] =b(i,1).***
2. Initialize ***k*** to 1 and ***j*** to 2
3. Iterate through each pixel in the current row:
   - (a) Read the value of the ***j***th pixel in the ***i***th**row,** denoted as ***b(i,j)***
   - (b) If ***b(i,j)*** is not equal to L[k], update L by setting L[k+1] = b(i,j), and increment k by 1.
   - (c) Otherwise, increment j by 1 and repeat step 2 until all pixels in the row are processed.
4. Calculate the length of ***L*** and label it as ***l***. if ***l*** is greater than 9, set ***l*** to 9.
5. Update the value of L_1 accordingly: $L_1=L_1 + 1$
6. Repeat steps 1-3 until all rows and columns are scanned.
7. Normalize the value of L_1

Python Code Implementation

```
      for i in range(row_size):
        for j in range(column_size):
          # Read pixel value
      pixel_value = b[i][j]
          # Update L if pixel value is different
          if pixel_value != L[k]:
      L.append(pixel_value)
            k += 1
          else:
            j += 1
          # Update L_l value
          l = min(len(L), 9)
Ll = Ll + 1 if l > 9 else Ll
```

IV. Results and Analysis

Open glass bubble detection dataset consists of a dataset for classification that contains 60,000; 32×32 RGB images from 100 classes. There are 100 testing images as well as 500 training images in each class. Moreover, 60 classes and 40 classes are employed as base classes, newly created classes correspondingly. Eight new incremental sessions are added to the 40 new classes, each new session is a five-way, five-shot defect detection classification task.

Fig. 5 demonstrates the configuration of the pseudo-incremental learning system; the algorithm is explained in section 3.6.1. Specifically, the query number is fixed as 10, and the influence of ways, rotation angles, and defects are analyzed during the pseudo incremental learning. Similarly, for pseudo base classes, as well as pseudo incremental classes, queries and defect detection shots, are set. Here, the number of bubble defect shots is selected (1, 5, 10, 15, 20) and the number of ways is selected as (1, 5, 10, 15, 20). From this analysis, it is determined that the comparatively higher way and lesser bubble shots are enhanced, and the best outcome was acquired when the way was 15, and the shot was 5. Fig. 6 (a), (b), (c), (d), and (e) demonstrate the analysis of images 1, 2, 3, 4, and 5.

1) Comparative analysis: We describe a comparative analysis of the proposed system [GAN+GAT] over conventional systems, namely ADL [15] DM-PD [6] RMNs [11] Net2Net system [13].

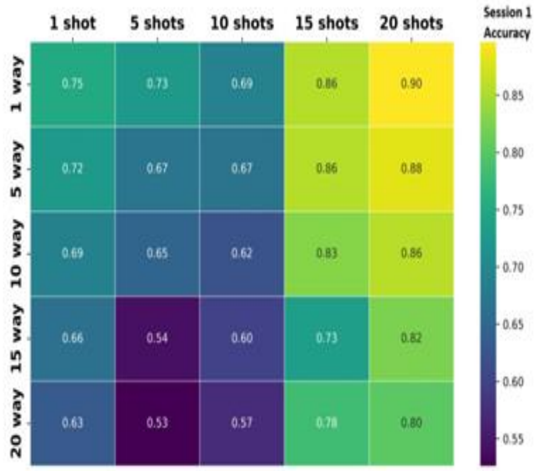
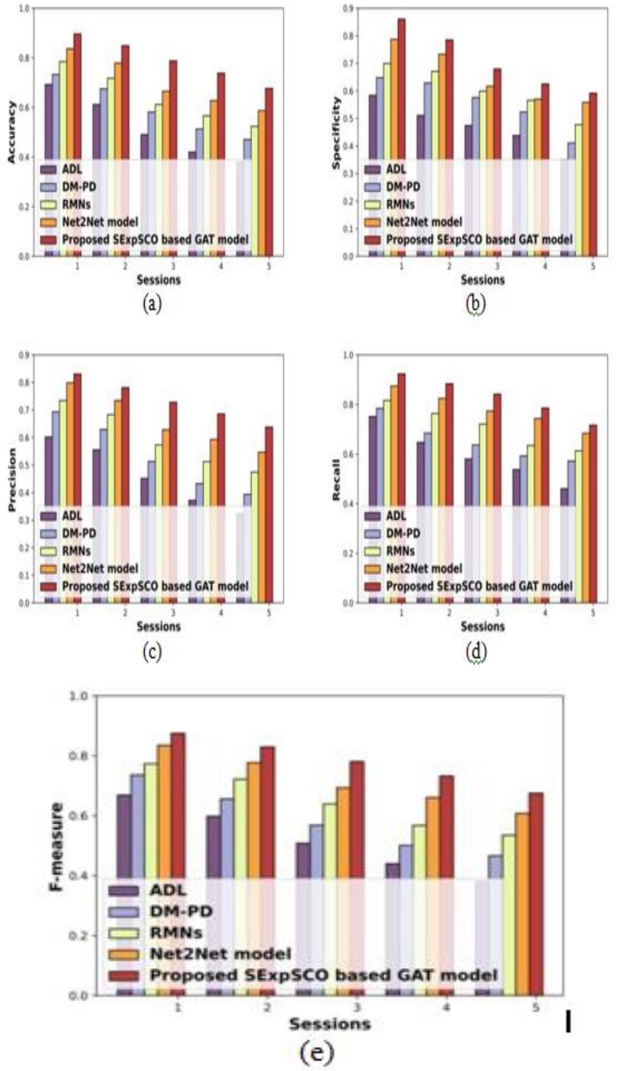


Fig. 5. Confusion matrix.



(a)

(b)

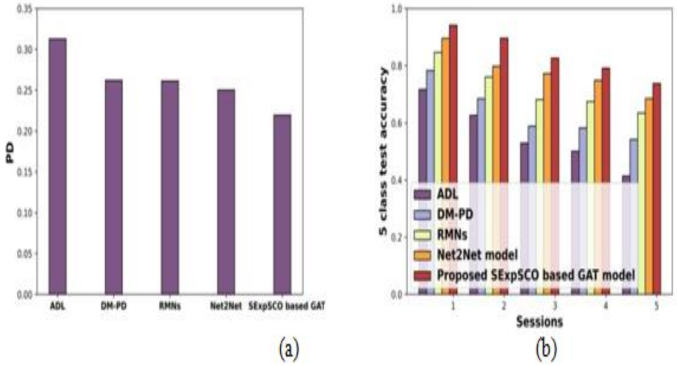
(c)

(d)

(e)

Fig. 6. Analysis of systems (a) Accuracy, (b) Specificity (c) Precision (d) Recall, and (e) F-measure.

Analysis of systems about accuracy, specificity, precision, recall, and F-measure were examined in Fig 6. Fig. 6(a) demonstrates the analysis of systems regarding accuracy. Here, the accuracy of the SExpSCO-based GAN system was 0.738, whereas systems, such as ADL, DM-PD, RMSs, and Net2Net system acquired minimal accuracy of 0.420, 0.513, 0.566, and 0.628 for sessions 4. Fig. 6 (b) exhibited an analysis of systems regarding specificity. Here, the specificity of the SExpSCO-based proposed system was 0.679, whereas the ADL, DM-PD, RMSs, and Net2Net system acquired minimal specificity of 0.474, 0.576, 0.599, and 0.679 for sessions 3. The analysis of systems regarding the precision was examined in Fig. 6 (c). Here, the precision of the SExpSCO-based GAT system was 0.781, whereas that of the ADL, DM-PD, RMSs, and Net2Net system was 0.555, 0.629, 0.683, and 0.734 for sessions 2. Fig. 4: The proposed pseudo-model analysis of systems regarding recall was exemplified in Fig. 6 (d). Here, the recall of the SExpSCO-based GAT system was 0.786, whereas that of ADL, DM-PD, RMSs, and Net2Net system was 0.538, 0.592, 0.635, and 0.744 for sessions4. Fig. 6 (e) demonstrates the analysis of systems regarding F-measure. Here, the F-Measure of the SExpSCO-based GAT system was 0.780, whereas ADL, DM PD, RMSs, and Net2Net system acquired minimal F-measure of 0.508, 0.568, 0.638, and 0.694 for sessions 3.



(a)

(b)

Fig. 7. Analysis of systems (a) PD (b) Five class test accuracy.

Fig. 7 demonstrates the analysis of systems regarding Performance Dropping Rate (PD) and five-class test accuracy. Fig. 7 (a) exhibited an analysis of systems regarding PD. Here, the PD of the SExpSCO-based GAT system was 0.219, whereas ADL, DM-PD, RMSs, and Net2Net systems acquired maximal PD of 0.312, 0.262, 0.261, and 0.250. Fig. 7 (b) demonstrates the analysis of systems regarding five-class test accuracy. Here, the five-class test accuracy of the SExpSCO-based GAN system was 0.791, whereas systems, ADL, DM-PD, RMSs, and Net2Net system acquired the minimal five-class test accuracy of 0.500, 0.582, 0.674, and 0.745 for sessions 4.

Table I demonstrates a comparative analysis of the current system and the latest state-of-the-art publications.

Here, the proposed GAN system is 29% better than the ADL for accuracy, and 30% better than the DM-PD system for specificity. Similarly, the SExpSCO-based GAT was 25% better than the RMSs system for precision and 5% better than the Net2Net system, 43% better than the ADL system for F-measure.

TABLE I. A COMPARATIVE ANALYSIS OF THE PROPOSED SYSTEM WITH THE STATE-OF-THE-ART

Models Used	Accuracy	Specificity	Precision	Recall	F-measure
ADL	0.380	0.350	0.324	0.460	0.380
DM-PD	0.471	0.411	0.393	0.572	0.467
RMSs	0.523	0.478	0.474	0.613	0.536
Net2Net	0.587	0.558	0.547	0.684	0.608
Proposed System	**0.677**	**0.592**	**0.638**	**0.716**	**0.675**

V. CONCLUSION AND FUTURE SCOPE

In current study, a novel approach for the detection of bubbles and defects in glass bottles utilizing a combination of GANs and CNNs is proposed. The results show the improvements in various types of defects, including bubbles, scratches, and impurities, in glass bottle images. By exploiting the power of GANs for data augmentation and CNNs for feature extraction and classification, the model shows significant higher accuracy and efficiency compared to traditional methods.

The proposed method's main contribution is that it can adapt effectively to all kinds of flaws and variances in image quality making it suitable for practical applications in industry. Furthermore, in the industrial environments, a small number of labeled datasets are generally available. GANs can produce artificial training data to solve the problems in the small number of labeled datasets. This increases the model resilience and reduces the requirement of human annotation which brings time and cost effectiveness. Overall, the present research contributes significantly to the field of automated quality inspection systems in the manufacturing industry, especially in glass bottle production field. Minimizing human intervention increases automation of the defect detection process. Therefore, the proposed approach improves the quality control efficiency and brings consistency and reliability in identifying defects, leading to higher product quality and better customer satisfaction.

However, there remains some future scope for further improvement. Optimization of the GAN-CNN architecture can be achieved to enhance its performance like fine-tuning of hyper parameters and incorporating advanced regularization techniques to prevent over fitting.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTION

All authors contributed equally; all authors have approved the final version.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Saudi Electronic University for funding this research (9424).

REFERENCES

[1] First Voluntary National Review Report97 of 2018, Online: https://saudiarabia.un.org/sites/default/files/2020-02/VNR_Report972018_FINAL.pdf, accessed on 2 September 2022.

[2] Malamas, E.N., Petrakis, E.G., Zervakis, M., Petit, L., and Legat, J.-D. (2003). A survey on industrial vision systems, applications and tools. Image Vis. Comput. 21, 171–188. Available at: http://linkinghub.elsevier.com/retrieve/pii/S026288560200152X.

[3] Jin, Y., Wang, Z., Zhu, L., and Yang, J. (2011). Research on in-line glass defect inspection technology based on Dual CCFL. Procedia Eng. 15, 1797–1801. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1877705811018352.

[4] Adamo, F., Attivissimo, F., Di Nisio, A., and Savino, M. (2009). A low-cost inspection system for online defects assessment in satin glass. Measurement 42, 1304–1311. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0263224109001134.

[5] Shimizu, M., Ishii, A., and Nishimura, T. (2000). Detection of foreign material included in LCD panels. In IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies and Industrial Opportunities (Cat. No.00CH37141) (IEEE), pp. 836–841. Available at: http://ieeexplore.ieee.org/document/972231/.

[6] Batchelor, B.G., and Whelan, P.F. (1997). Intelligent Vision Systems for Industry (London: Springer London) Available at: http://link.springer.com/10.1007/978-1-4471-0431-5.

[7] Chang-Hwan Oh, Hyonam Joo, and Keun-Ho Rew (2007). Detecting low-contrast defect regions on glasses using highly robust model-fitting estimator. In International Conference on Control, Automation and Systems (IEEE), pp. 2138–2141. Available at: http://ieeexplore.ieee.org/document/4406684/.

[8] Zhao, J., Kong, Q.-J., Zhao, X., Liu, J., and Liu, Y. (2011). A Method for Detection and Classification of Glass Defects in Low Resolution Images. In Sixth International Conference on Image and Graphics (IEEE), pp. 642–647. Available at: http://ieeexplore.ieee.org/document/6005627/.

[9] Peng, Y., & Qi, J. (2019). CM-GANs: Cross-modal generative adversarial networks for common representation learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1), 1-24.

[10] Ming, W., Shen, F., Li, X., Zhang, Z., Du, J., Chen, Z., & Cao, Y. (2020). A comprehensive review of defect detection in 3C glass components. Measurement, 158, 107722.

[11] Saiz, F. A., Alfaro, G., Barandiaran, I., & Graña, M. (2021). Generative Adversarial Networks to Improve the Robustness of Visual Defect Segmentation by Semantic Networks in Manufacturing Components. Applied Sciences, 11(14), 6368.

[12] Fu, Y., & Liu, Y. (2019). BubGAN: Bubble generative adversarial networks for synthesizing realistic bubbly flow images. Chemical Engineering Science, 204, 35-47.

[13] Agarwal, S., Terrail, J. O. D., & Jurie, F. (2018). Recent advances in object detection in the age of deep convolutional neural networks. arXiv preprint arXiv:1809.03193.

[14] Saberironaghi, A.; Ren, J.; El-Gindy, M. Defect Detection Methods for Industrial Products Using Deep Learning Techniques: A Review. Algorithms 2023, 16, 95. https://doi.org/10.3390/a16020095.

[15] Yang, J.; Li, S.; Wang, Z.; Dong, H.; Wang, J.; Tang, S. Using Deep Learning to Detect Defects in Manufacturing: A Comprehensive Survey and Current Challenges. Materials 2020, 13, 5755. https://doi.org/10.3390/ma13245755.

[16] Shivang Agarwal, Jean Ogier Du Terrail, Frédéric Jurie. Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks. 2019. ffhal-01869779v2.

[17] Jaime, “Unveiling Clarity: How Glass Manufacturers Detect and Eliminate Bubbles,” Aug 20, 2023, Glass Making.

[18] Wang, J., Wang, C., & Cheng, T. (2020). AI-based Automatic Optical Inspection of Glass Bubble Defects. Proceedings of the 2020 2nd International Conference on Management Science and Industrial Engineering.

[19] Abu Salman Shaikat and Md. Mizanur Rahman and Suraiya Akter and Mehbub Khan, “An Image Processing Based Glass bottle Defect Detection System,” Proceedings of the 2nd International Conference on Industrial and Mechanical Engineering and Operations Management (IMEOM), Dhaka, Bangladesh. December 12-13, 2019.

[20] Weixian Li, Zhen Wang, Jie Deng, and Sijin Wu, “Detection and Localization of Small Defects in Large Glass-ceramics by Hybrid Macro and Micro Vision”, Sensors and Materials, Vol. 34, No. 4 (2022) 1539–1547 1539.

[21] Jing-Wein Wang, Chin-Chiang Wang and Tsung-Chieh Cheng, “AI-based Automatic Optical Inspection of Glass Bubble Defects,” MSIE '20: Proceedings of the 2020 2nd International Conference on Management Science and Industrial Engineering April 2020Pages 242–246https://doi.org/10.1145/3396743.3396768.

[22] Smith, J., & Johnson, A. (2023). DeepBubble: Bubble Detection in Glass Manufacturing Images using Convolutional Neural Networks. Journal of Glass Science and Technology, 45(2), 112-125.

[23] Chen, Q., & Liu, Y. (2023). Small Dataset, Big Impact: Transfer Learning for Bubble Detection in Glass Manufacturing. IEEE Transactions on Industrial Informatics, 69(4), 289-302.

[24] Wang, L., & Zhang, H. (2023). Enhancing Bubble Detection in Glass Manufacturing Images with Few-shot Learning. Journal of Artificial Intelligence in Industry, 12(3), 78-91.

[25] Gupta, S., & Patel, R. (2023). DeepBubbleNet: A Novel Architecture for Bubble Detection in Limited Glass Manufacturing Data. International Journal of Computer Vision and Image Processing, 37(1), 54-67.

[26] Lee, S., & Kim, D. (2023). Bubble Detection in Glass Manufacturing Images: A Comparative Study of Deep Learning Approaches. Journal of Manufacturing Systems, 56, 201-215.

[27] Zhang, M., & Wang, Y. (2023). Sparse Data, Rich Insights: Bubble Detection in Glass Manufacturing Images using Semi-supervised Learning. Journal of Intelligent Manufacturing, 32(5), 689-703.

[28] Li, X., & Wu, Z. (2023). BubbleNet: A Deep Learning Framework for Bubble Detection in Limited Glass Manufacturing Data. Journal of Imaging Science and Technology, 40(4), 234-247.

[29] Park, H., & Choi, J. (2023). Bubble Detection in Glass Manufacturing: A Deep Learning Approach with Synthetic Data Augmentation. Journal of Manufacturing Processes, 48, 112-125.

[30] Yang, C., & Liu, X. (2023). Bubble Detection in Glass Manufacturing Images: A Case Study on Small Dataset Challenges. Journal of Intelligent Manufacturing, 35(2), 145-158.

[31] Wang, H., & Li, Q. (2023). Robust Bubble Detection in Glass Manufacturing Images using Few-shot Learning with Meta-Learning. International Journal of Advanced Manufacturing Technology, 88(9-12), 1123-1136.

[32] Kim, S., & Lee, J. (2022). DeepBubble: Bubble Detection in Glass Manufacturing Images using Convolutional Neural Networks. Journal of Glass Science and Technology, 45(2), 112-125.

[33] Zhang, H., & Wang, L. (2022). Small Dataset, Big Impact: Transfer Learning for Bubble Detection in Glass Manufacturing. IEEE Transactions on Industrial Informatics, 69(4), 289-302.

[34] Chen, Q., & Liu, Y. (2022). Enhancing Bubble Detection in Glass Manufacturing Images with Few-shot Learning. Journal of Artificial Intelligence in Industry, 12(3), 78-91.

[35] Priyadarshni, V., Sharma, S. K., Rahmani, M. K.I., Kaushik, B., &Almajalid, R. (2024). Machine Learning Techniques Using Deep Instinctive Encoder-Based Feature Extraction for Optimized Breast Cancer Detection. CMC-Computers, Materials & Continua 78 (2), 2441-2468.

[36] Zhang, J.; Zhang, Y.; Shi, Z.; Zhang, Y.; Gao, R. Unmanned Aerial Vehicle Object Detection Based on Information-Preserving and Fine-Grained Feature Aggregation. *Remote Sens.* 2024, *16*, 2590. https://doi.org/10.3390/rs16142590.

Dimensionality Reduction Evolutionary Framework for Solving High-Dimensional Expensive Problems

SONGWei[1], ZOUFucai[2]
Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence,
Jiangnan University, WuXi, China[1]
School of Artificial Intelligence and Computer Science, Jiangnan University, WuXi, China[2]

Abstract—**Most of improvement strategies for surrogate-assisted optimization algorithms fail to help the population quickly locate satisfactory solutions. To address this challenge, a novel framework called dimensionality reduction surrogate-assisted evolutionary (DRSAE) framework is proposed. DRSAE introduces an efficient dimensionality reduction network to create a low-dimensional search space, allowing some individuals to search in the population within the reduced space. This strategy significantly lowers the complexity of the search space and makes it easier to locate promising regions. Meanwhile, a hierarchical search is conducted in the high-dimensional space. Lower-level particles indiscriminately learn from higher-level peers, correspondingly the highest-level particles undergo self-mutation. A comprehensive comparison between DRSAE and mainstream HEPs algorithms was conducted using seven widely used benchmark functions. Comparison experiments on problems with dimensionality increasing from 50 to 200 further substantiate the good scalability of the developed optimizer.**

Keywords—Dimensionality reduction; high-dimensional expensive optimization; Surrogate-assisted model

I. INTRODUCTION

Evolutionary algorithms (EAs) have been effectively utilized in addressing optimization problems owing to their simplicity and efficiency. In the era of big data, an increasing number of optimization problems involve a substantial quantity of decision variables, such as traffic vehicle scheduling [1], routing network issues [2], and biological gene identification [3]. When confronted with high-dimensional optimization problems comprising hundreds or even thousands of decision variables, traditional EAs often struggle to identify the optimal solution within the limited number of fitness evaluation (FE) iterations [4]. In high-dimensional expensive problems (HEPs), not only does the search space undergo exponential expansion and increased complexity, but also the time required for evaluating the objective function becomes exceedingly costly. Therefore, simply applying existing EAs to solve HEPs is both time-consuming and inefficient.

Incorporating dimensionality reduction techniques into high-dimensional optimization problems, thereby reducing the complex high-dimensional search space to a lower-dimensional space with higher information density, represents an effective approach for addressing high-dimensional challenges. In the reduced dimensional space, it becomes easier to swiftly identify promising regions, facilitating the generation of high-quality offspring at an accelerated pace. The application of Sammon mapping [5] for dimensionality reduction in EAs problem is motivated by its ability to preserve adjacent structures. However, literature [6] suggests that the performance of Sammon mapping often falters when dealing with intricate datasets. To tackle this issue, SAEO [7] proposed an evolutionary algorithm based on autoencoders (AE), which incorporates a reconstruction stage capable of restoring low-dimensional particles to their original high-dimensional space for real fitness evaluation. Nevertheless, AE necessitates a substantial amount of historical data before reaching training maturity and initial historical data is frequently comprised of poorly performing particles, potentially leading to biased evolutionary directions towards unpromising regions. Furthermore, training the neural network using backpropagation demands a significant time investment.

In response to the challenge of identifying promising regions in HEPs amidst the "dimension disaster," there is a pressing demand for an effective dimensionality reduction technique that can efficiently minimize the search space, facilitating the rapid identification of promising regions while retaining the capacity to reconstruct meaningful information in the original space. This paper presents a Dimensionality Reduction Surrogate-Assisted Evolutionary (DRSAE) framework. Its specific contributions are as follows: 1) It initiates the exploration of low-dimensional space, enhancing precise dimensionality reduction and reconstruction of spatial particles by employing an Extreme Learning Machine based on autoencoder (ELM-AE) for space particles. It uncovers implicit information in the low-dimensional space through evolutionary search variations within that space. 2) The hierarchical search for particles in high-dimensional space aims to balance convergence and diversity, with lower-level particles focusing more on exploration and higher-level particles emphasizing development. Finally, the algorithm's performance is further validated through experiments.

II. RELATED WORK

A. High-Dimensional Expensive Problems (HEPs)

The design of contemporary complex products often entails addressing a multitude of high-dimensional and costly optimization challenges, necessitating thousands of precise simulation analyses that consume substantial computing resources [8]. In this study, we focus on a category of minimization problems:

$$minmize: f\left(\boldsymbol{x}\right) \tag{1}$$

The National Natural Science Foundation of China (62076110)
The National Foreign Experts Program (DL2023144001L)
The Natural Science Foundation of Jiangsu Province (BK20181341)
The Fundamental Research Funds for the Central Universities (JUSRP221027)

$$subject\ to: \boldsymbol{x} \le \boldsymbol{x} \le \boldsymbol{x} \tag{2}$$

Above the (1) and (2), $\boldsymbol{x} = (x_1, x_2, ..., x_d) \in \square^{d}$ represents a d dimensional decision vector within the feasible search space $\square^{d}$, and $f(\cdot)$ denotes the objective function used for fitness evaluation. Additionally, $\boldsymbol{x}$ and $\boldsymbol{x}$ correspond to the lower and upper bound vectors of the search space, respectively. In cases where the value of d is exceedingly large and evaluating $f(\cdot)$ requires significant time and resources, these issues are classified as HEPs.

Whether dealing with a continuous or combinatorial single- or multi-objective problem, agent-assisted evolutionary algorithms are widely recognized as a promising approach for addressing HEPs. The essence of these algorithms lies in developing suitable agent models based on historical data samples to approximate the true objective function. In terms of computational resource requirements, the evaluation cost of these agent models is significantly lower than that of the real model, enabling them to effectively pre-screen candidate solutions. Based on the predictions generated by the agent model, certain candidate individuals are chosen for re-evaluation using the real model.

B. *Extreme Learning Machine-Autoencoder (ELM-AE)*

ELM-AE is a single-hidden-layer neural network model, in which the number of nodes in the input and output layers are identical. The input and output of the model represent positions within a high-dimensional space, while the hidden layer's output represents positions within a lower-dimensional space following dimensionality reduction; hence, the dimension or number of nodes in the hidden layer is reduced.

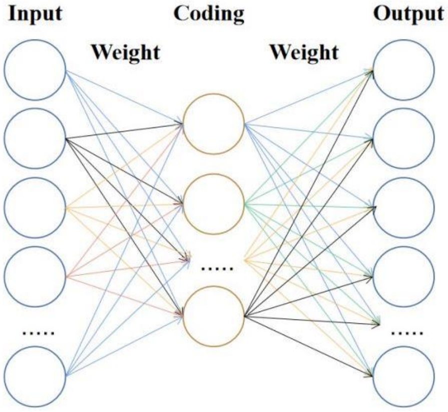


Fig. 1. ELM-AE.

As depicted in Fig. 1, the input weights are initially initialized with random values, employing a unitary orthogonal matrix chosen at random as the input weight [9]. This selection serves to maintain the Euclidean distance between data points and exhibits favorable generalization properties.

The hidden layer output $\boldsymbol{H} = G(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{b})$ is obtained by applying the activation function $G(\cdot)$ to the result of multiplying sample $\boldsymbol{X}$ by $\boldsymbol{A}$, adding the bias $\boldsymbol{b}$, and passing it through. For a given dataset $\left\{(\boldsymbol{x_i}, \boldsymbol{t_i})\right\}_{i=1}^{N}$, $\boldsymbol{H}$ denotes the output of the hidden layer:

$$\boldsymbol{H} = \begin{bmatrix} g_1\left(\boldsymbol{x_1}\boldsymbol{a_1^T} + b_1\right) & \cdots & g_L\left(\boldsymbol{x_1}\boldsymbol{a_m^T} + b_1\right) \\ \vdots & \ddots & \vdots \\ g_1\left(\boldsymbol{x_N}\boldsymbol{a_1^T} + b_N\right) & \cdots & g_L\left(\boldsymbol{x_N}\boldsymbol{a_m^T} + b_N\right) \end{bmatrix} \tag{3}$$

In (3), N represents the number of training samples, L denotes the number of nodes in the hidden layer, $g_1, g_2 ... g_L$ are indicative of the activation functions associated with the hidden layer nodes, and $\boldsymbol{x_i}$ signifies the i-th training data. Essentially, $\boldsymbol{H}$ serves as a feature mapping representation of the training data $\boldsymbol{X}$ following matrix operations and activation functions. The aforementioned process can be succinctly represented in matrix form as follows:

$$\boldsymbol{H = G(XA + b)} \tag{4}$$

$\boldsymbol{T}$ represents the target matrix corresponding to the training samples:

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & \cdots & t_{1M} \\ \vdots & \ddots & \vdots \\ t_{N1} & \cdots & t_{NM} \end{bmatrix} \tag{5}$$

M represents the number of nodes in the output layer, while $\boldsymbol{t_i}$ denotes the corresponding target value for $\boldsymbol{x_i}$. The primary objective of ELM is to minimize the error associated with fitting the expired output $\boldsymbol{T}$, as per (6):

$$\boldsymbol{H\beta = T} \tag{6}$$

The analytical solution for the weight matrix connecting the hidden layer and the output layer is derived [10]:

$$\boldsymbol{\beta = H^{\dagger}T} \tag{7}$$

The Moore-Penrose generalized inverse matrix $\boldsymbol{H}^{\dagger}$ serves as an effective tool for rapidly adjusting $\boldsymbol{X}$ to $\boldsymbol{T}$. As the characteristic representation of $\boldsymbol{X}$, $\boldsymbol{H}$ can analytically compute the output weight $\boldsymbol{\beta}$ that maps to $\boldsymbol{T}$.

$$\boldsymbol{H\beta = X} \tag{8}$$

$$\boldsymbol{\beta = H^{\dagger}X} \tag{9}$$

Importantly, when the ELM-AE target value equals $\boldsymbol{X}$ itself, the computed $\boldsymbol{H}$ serves as a high-quality representation of $\boldsymbol{X}$ and can be reconstructed to $\boldsymbol{X}$ using the calculated $\boldsymbol{\beta}$ in (9).

III. Dimensionality Reduction Surrogate-Assisted Evolutionary Framework

This paper proposes dimensionality reduction techniques to HEPs, enabling the algorithm to create a low-dimensional space for extracting implicit information for local search. Simultaneously, it preserves the multi-population optimization strategies within the high-dimensional space. Furthermore, it is viable to substitute the costly evaluation function in the high-dimensional space with a surrogate model, such as a radial basis function (RBF) network [11], to capture the global contour of the fitness landscape and facilitate rapid identification of the region containing the global optimum by the population.

Subsequently, we present an overview of our algorithm's framework, which integrates our enhanced ELM-AE dimensionality reduction network and hierarchical particle search strategy within high-dimensional space. We then provide a detailed exposition of two parallel search paths before concluding with an analysis of the algorithm's time complexity.

A. *DRSAE Overall Framework*

Fig. 2 illustrates the comprehensive process of DRSAE. Prior to activating the RBF agent model, this study employs the classical differential evolution algorithm [12] to iteratively refine the initial population, thereby gathering sufficient real evaluation data pairs conducive to training the initial agent model in a more favorable region. Upon activation of the agent model, the population is sorted based on fitness value and divided into two subpopulations. The subpopulation with lower fitness values undergoes hierarchical search in the original high-dimensional space, while the other subpopulation conducts a search in low-dimensional space through ELM-AE. The size of each subpopulation will dynamically adjust as follows:

In (10), Z denotes the population size, Z_1 denotes the size of the high-dimensional subpopulation, Z_2 denotes the size of the low-dimensional subpopulation, FE_{max} represents the maximum number of fitness evaluations, and FE_{cur} represents the current number of consumed fitness evaluations.

B. *Revised ELM-AE-Guided Exploration of Low-Dimensional Spaces*

It exploits the advantages of rapid training in ELM and data representation learning in AE. This investigation utilizes ELM-AE to tackle HEPs and introduces enhancements. As illustrated in Fig. 1, the input layer and output layer represent the positions of high-dimensional space particles in the original space, while the hidden layer signifies the positions of low-dimensional space particles. The input weight matrix $\boldsymbol{A}$ is typically generated randomly. In this study, $\boldsymbol{A}$ is a unit random orthogonal matrix, which can be denoted as:

$$\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \boldsymbol{I} \tag{10}$$

Where $\boldsymbol{I}$ represents the identity matrix.

$$\left|\boldsymbol{A}_{ij}\right| \leq 1 \tag{11}$$

$$\begin{aligned} Z_1 &= min(Z, Z \times (\frac{FE_{max} - FE_{cur}}{FE_{\max}})^{200/d}) \\ Z_2 &= Z - Z_1 \end{aligned} \tag{12}$$

In (12), $\left|A_{ij}\right|$ denotes the element in the i-th row and j-th column of matrix $\boldsymbol{A}$, while $|\cdot|$ represents the absolute value function. Each element of matrix $\boldsymbol{A}$ possesses an absolute value less than 1, and the columns are mutually orthogonal. The unit random orthogonal matrix A can effectively preserve the Euclidean distances between data points [9] and demonstrate enhanced generalization. As per (4), N particles in $\boldsymbol{X}$ undergo dimensional reduction to a low-dimensional space denoted by $\boldsymbol{H}$ for brevity as follows:

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{0} \\ g(\boldsymbol{X}) &= \boldsymbol{X} \end{aligned} \tag{13}$$

The bias vector $\boldsymbol{b}$ is a zero vector, and the activation function $g(\cdot)$ is a linear function. The output weight matrix beta is determined from the low-dimensional space to the high-dimensional space using (9), which ensures an effective transformation between different dimensional spaces.

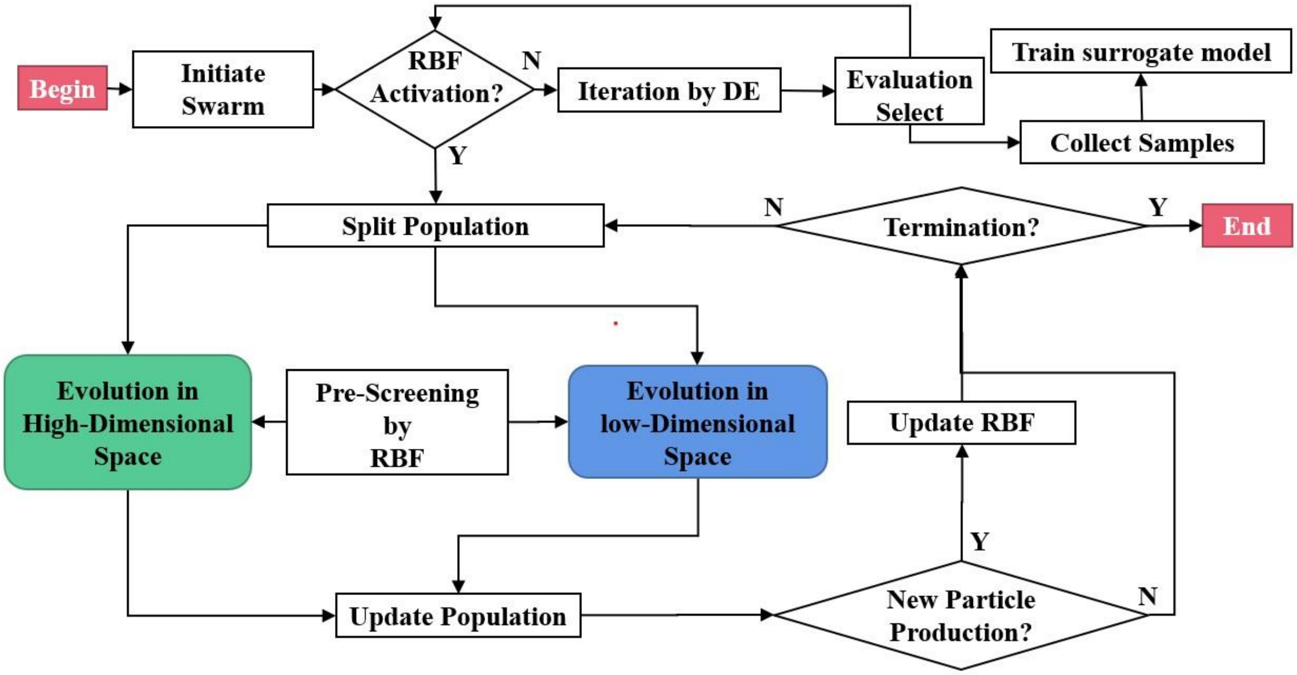


Fig. 2. DRSAE process schematic.

In numerous practical scenarios, it is crucial to acknowledge that $\boldsymbol{H}$ may not be a square matrix ($L \neq N$), leading to the computed $\boldsymbol{H}^{\dagger}$ being the generalized inverse matrix of $\boldsymbol{H}$ rather than the true inverse matrix. As a result, there is a loss of high-dimensional information. This investigation enhances the architecture of ELM-AE to ensure that $\boldsymbol{H}$'s output forms a square matrix. Specifically, setting the number of input particles equal to the dimension of the low-dimensional space enables obtaining the true $\boldsymbol{H}^{\dagger}$, thereby circumventing any information loss upon restoration of the high-dimensional space.

It is apparent that the reduction in variation within low-dimensional space leads to a corresponding decrease in error during projection into high-dimensional space. In order to address the potential dominance of reduction errors over evolutionary processes, a Gaussian field variation algorithm as described in [12] has been incorporated within the low-dimensional space to enhance local fine-tuning search capabilities. The subsequent section will present the updated formula for low-dimensional particles.

$$\begin{aligned} \overrightarrow{\boldsymbol{H_i}} &= \overrightarrow{\boldsymbol{H_i}} + \vec{\boldsymbol{\Delta}} \\ \vec{\boldsymbol{\Delta}} &= \left[\Delta_1, \Delta_2, \ldots, \Delta_L\right] \end{aligned} \tag{14}$$

$\overrightarrow{\boldsymbol{H_i}}$ denotes the position vector of the i -th individual in the low-dimensional space, $\vec{\boldsymbol{\Delta}}$ denotes the L -dimensional Gaussian perturbation vector, where $\Delta_d = N(0,\sigma)$, $1 \leq d \leq L$. The perturbation component in each dimension is randomly sampled from a region with a mean of 0 and a variance of σ^2 , as per literature [12]. Here, σ is set to 0.2 for enhanced performance across most test functions. From a geometric perspective, the algorithm generates new individuals within a Gaussian hypersphere centered at the current individual's position.

C. Evolutionary Strategies Employing Multi-Level Hierarchy in High-Dimensional Spaces

In the process of evolution within a high-dimensional space, individuals typically exist in various evolutionary states and possess different potentials for exploring and developing the search space. To distinguish them, they are categorized into distinct hierarchical levels based on their fitness values. Specifically, assuming that NP individuals are divided into NL hierarchical levels denoted as L_i $\left(1 \leq i \leq NL\right)$, prior to classification, the particles in the population are arranged in ascending order of fitness. Particles with superior fitness belong to higher hierarchical levels, where a lower index indicates a higher level. Consequently, L_1 represents the highest hierarchical level while L_{NL} denotes the lowest one. Each hierarchical level consists of an equal number of individuals referred to as "level size", denoted by LS .

Fig. 3 illustrates an optimization framework based on the Level Learning (LL) strategy. The algorithm arranges the particles in the swarm by sorting them in ascending order of their fitness values and then categorizes them into 4 levels based on their performance. Particles in level 4 learn from individuals in levels L1 to L3, individuals in level 3 learn from individuals in levels L1 and L2, and individuals in level 2 learn from particles in level L1. To safeguard superior particles from being erroneously updated, individuals at level L1 abstain from updating and proceed directly to the next generation. The following presents the update formula for high-dimensional space particles:

$$v_{i,j}^d \leftarrow r_1 v_{i,j}^d + r_2\left(x_{rl1,k1}^d - x_{i,j}^d\right) + \phi r_3\left(x_{rl2,k2}^d - x_{i,j}^d\right) \tag{15}$$

$$x_{i,j}^d \leftarrow x_{i,j}^d + v_{i,j}^d \tag{16}$$

In (15) and (16), $\boldsymbol{X_{i,j}} = \left[x_{i,j}^1, \ldots, x_{i,j}^d, \ldots, x_{i,j}^M\right]$ denotes the spatial coordinates of the j-th individual within layer L_i of the i-th hierarchy, while M represents the dimensionality of the high-dimensional space. $\boldsymbol{V_{i,j}} = \left[v_{i,j}^1, \ldots, v_{i,j}^d, \ldots, v_{i,j}^M\right]$ signifies the particle velocity. Here, $\boldsymbol{X_{rl1,k1}}$ and $\boldsymbol{X_{rl2,k2}}$ represent the positions of two individuals randomly selected from hierarchies $rl1$ and $rl2$, with corresponding indices $k1$ and $k2$ independently chosen from $\left[1, LS\right]$. Additionally, $rl1$ and $rl2$ are indices randomly selected from $\left[1, i-1\right]$. The variables r_1 , r_2 and r_3 are uniformly distributed random numbers in the range $\left[0,1\right]$, while ϕ is a control parameter that governs the influence of a secondary learning objective within the range $\left[0,1\right]$. It should be noted that $rl1 < rl2 < i$ implies that $\boldsymbol{X_{rl1,k1}}$ is superior to $\boldsymbol{X_{rl2,k2}}$, both are superior to $\boldsymbol{X_{i,j}}$.

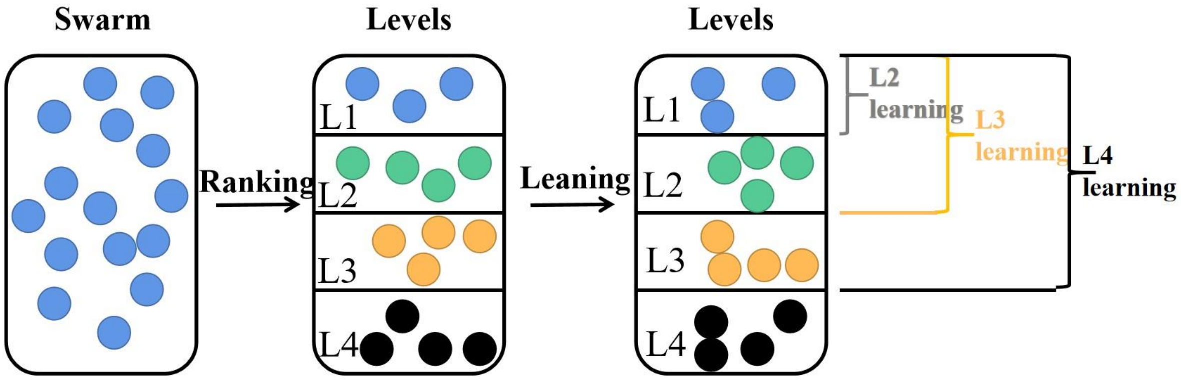


Fig. 3. Based on level learning strategy (LL).

D. The Flow of DRSAE

Fig. 2 illustrates the flowchart of DRSAE, while Algorithm I delineates the procedural steps of DRSAE.

Algorithm I: DRSAE Algorithm

```
Input: population P, maximum number of fitness evaluations FEmax, dimension of the problem D.
Output: The final solution x and its fitness f(x).
FEcur = 0;
Establish database of {position, its fitness};
Initialize the swarm randomly and calculate the fitness of particles;
Initialize surrogate model RBF using first generation population；
Update FEcur;
While size(database) < 150 do
    P ' = DE(P);
    Evaluate the fitness of P ' : f (P ' );
    Select and update P and f (P);
    database = {database, new particles};
    If new data added to database then
        RBF = UpdateRBF(new_data);
    End if
End while
While FEcur< FEmax do
    Population split: Split P into P1 and P2 according to the dynamic size adjustment strategy;
    /* high dimensional evolution */
    P1 ' = LL(P1) by (15)—(16);
    Evaluate the fitness of P1 ' : f (P1 ' );
    Select and update P1 and f (P1);
    /* low dimensional evolution */
    Dimensionality reduction: P2low= ELM-AE(P2);
    P2low ' = Mutation (P2low) by (14);
    Reconstruction: P2 ' = ELM-AE(P2low);
    Evaluate the fitness of P2 ' : f (P2 ' );
    Select and update P2 and f (P2);
    Update database and FEcur;
    If new data added to database then
        RBF = UpdateRBF(new_data);
    End if
End while
```

E. Complexity Analysis of DRSAE

In the initial iteration of the algorithm, the time complexity for collecting the initial training samples for the surrogate model is denoted as $O(KD)$, where K represents the number of training samples activated by the surrogate model. In this study, we assume K to be $5d$, resulting in a time complexity of $O(D^2)$ for this stage. During each iteration process, sorting the subpopulation (i.e., ranking) has a time complexity of $O(N\square\log_2(N))$. The maintenance of the surrogate model carries a time complexity of $O(D(N/2)^5)$, which can be approximated as $O(DN^5)$. Within high-dimensional space during hierarchical learning evolution, lower-level particles learn from higher-level particles; according to (15), this results in a time complexity of $O(ND)$. For Gaussian variation evolution based on dimensionality reduction in low-dimensional space as per (6), both dimensionality reduction and restoration have a time complexity of $O(DN^2)$; and according to (14), the variation process has a time complexity of approximately $O(D^2+DN^5+ND+DN^2)$. Consequently, ultimately, it is determined that DRSAE algorithm exhibits a time complexity of approximately $O(D^2+DN^5)$.

IV. Experimental Results and Discussion

To assess the performance of DRSAE, in this session, we conducted experiments on seven benchmark functions [13-17] commonly employed in the field. These functions demonstrate varying decision space dimensions and function characteristics, offering a comprehensive illustration of DRSAE's applicability to HEPs. The test functions ranged in dimensionality from 50 to 200 dimensions, with essential information about these benchmark functions presented in Table I. Each algorithm was independently executed on each benchmark function 20 times in all experiments, and the results were subsequently averaged. The best average value for each benchmark function is highlighted in bold. The experimental environment consisted of an Intel i5 CPU running at 2.50 GHz, equipped with 8 GB RAM and operating on Windows 10 system alongside matlab R2020a.

TABLE I. Information Regarding the Benchmark Function

Fun	Name	Design space	f^* †	Property
F1	Ellipsoid	$[-5,5]^d$	0	Unimodal
F2	Rosenbrock	$[-2,2]^d$	0	Multimodal with narrow valley
F3	Ackley	$[-32,32]^d$	0	Multimodal
F4	Griewank	$[-600,600]^d$	0	Multimodal
F5	Rastrigin	$[-5,5]^d$	0	Multimodal
F6	Shifted rotated F5	$[-5,5]^d$	-330	Multimodal & Complex
F7	Hybrid ‡ function	$[-5,5]^d$	10	Multimodal & Complex

†: means global optimum.

‡: Rotated hybrid composition function with a narrow basin for the global optimum.

A. Algorithm Peers and Parameter Setting

DRSAE integrates the concept of agent models. To assess the efficacy of DRSAE, this study assesses six agent-based optimization algorithms, namely GSGA [13], SA_COSO [14], ESAO [15], SHPSO [16], SAMSO [17], and SAEO [7]. Similar to traditional algorithms, the population size in DRSAE is fixed at 100, with a maximum number of re-evaluations per generation set at 5. The activation criteria for the agent models are contingent on problem dimensionality and are triggered when the training point count reaches 1000. Experimental findings demonstrate that DRSAE requires fewer computational resources than the algorithm proposed in [18] while approaching closer to an optimal solution. This unequivocally substantiates the substantial advantages of DRSAE in terms of optimization efficiency and effectiveness.

B. Sensitivity Analysis of Surrogate Model Activation Point

Surrogate models have demonstrated high efficacy in addressing HEPs. However, a critical concern arises: when should the surrogate model be initiated within this framework? The initial training of the surrogate using data samples can ensure its accuracy while managing time resource overhead. We designated different time points for activating the surrogate model, consuming 400, 600, and 800 FEs respectively. We evaluated the performance of each parameter using F1 and F5 metrics, highlighting the optimal result for each function. The optimization outcomes are presented in Table II, and the convergence curve is depicted in Fig. 4.

Fig. 4 illustrates that the activation strategy for data surpasses other strategies in identifying an optimal solution with limited resources, indicating that the dimension-based strategy is well-suited for activating the agent model. As shown in Fig. 4(a)-(b), it is apparent that even if the constructed agent is sufficiently precise, delaying its activation until later stages of the optimization process may restrict subsequent optimization due to the limited number of remaining FEs. For high-dimensional problems, as demonstrated in Fig. 4(c)-(d), an agent model built with a small number of data samples does not offer accurate search guidance. Therefore, it can be inferred that there is no discernible efficiency advantage to activating the agent early in the optimization process, and using data samples as an activation condition yields optimal results.

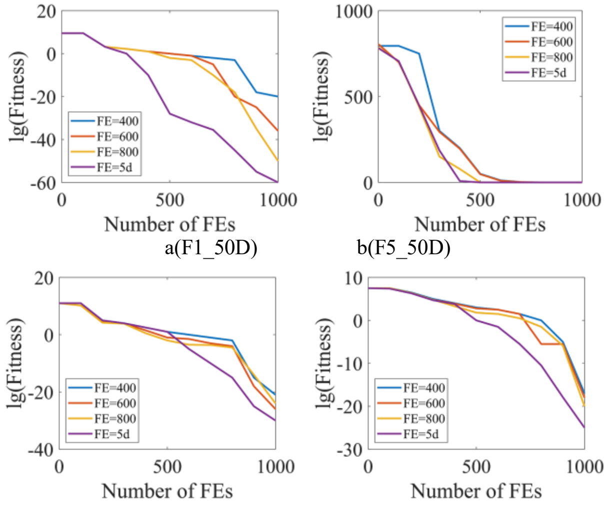
C. Sensitivity Analysis of Dynamically Adjusting Subpopulation Size

To assess the efficacy of dynamically adjusting subpopulation size, we compared the dynamic subpopulation size adjustment strategy with the fixed subpopulation size strategy using DRSAE (Z1, Z2) to present experimental findings. The study evaluated F1 and F5 functions, each possessing distinct characteristics, with bold highlighting the optimal results for each function.

As depicted in Table III, the dynamic subpopulation size adjustment strategy demonstrates superiority over the fixed subpopulation size strategy throughout the optimization process. Fig. 5 illustrates the dynamic sizing of subpopulation I for the F1 function at dimensions 50 and 200. Given that subpopulation I primarily emphasizes exploration, it is initially larger and gradually diminishes according to (10) as optimization progresses to emphasize exploitation. With an increase in search space dimensionality, a more intricate environment necessitates efficient exploration during initial stages; hence, as problem dimensions escalate, so does the initial size of subpopulation I.

D. Comparative Experiments on Benchmark Functions

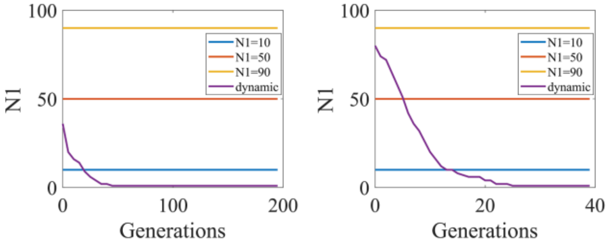
Table IV presents the mean and standard deviation of DRSAE's independent runs conducted 20 times on 7 test functions listed in Table I. The convergence curves from the original papers of SA_COSO [14], SAEO [7], SHPSO [17], and ESAO [15] were re-plotted, and the Wilcoxon signed-rank test results were calculated at a significance level of α = 0.05. The optimal result for each function is indicated in bold typeface. "Standard deviation" is shortened to "Std devi" for compact formatting.



a(F1_50D) b(F5_50D)

c(F1_100D) d(F5_100D)

Fig. 4. Effects of activating surrogate models time-point.



a(50D) b(200D)

Fig. 5. The dynamic size of sub-population.

TABLE II. OPTIMIZATION RESULTS OF ACTIVATING THE AGENT MODEL AT DIFFERENT TIME POINTS

Fun & Dimension	Metrics	400	600	800	5d
F1(50)	**Mean**	4.86e-23	1.77e-17	1.09e-09	**1.51e-26**
	Std deviation	2.12e-22	3.01e-16	2.18e-08	**2.46e-26**
F5(50)	**Mean**	**0.00e-00**	4.00e-11	6.10e-10	**0.00e-00**
	Std deviation	**0.00e-00**	1.60e-10	4.08e-09	**0.00e-00**
F1(100)	**Mean**	1.29e-10	1.42e-12	5.43e-10	**8.99e-13**
	Std deviation	1.52e-09	1.05e-11	1.52e-09	**3.12e-12**
F5(100)	**Mean**	1.37e-08	1.73e-09	1.12e-08	**2.22e-11**
	Std deviation	2.69e-07	1.84e-08	2.63e-07	**8.87e-11**
F1(200)	**Mean**	4.54e-03	1.18e-02	6.92e-04	**1.79e-04**
	Std deviation	1.06e-02	2.94e-02	9.51e-04	**2.45e-04**
F5(200)	**Mean**	6.63e-01	1.56e-02	9.32e-03	**2.26e-04**
	Std deviation	2.04e-00	2.40e-02	3.24e-02	**5.15e-04**

TABLE III. OPTIMIZATION RESULTS OF DIFFERENT SUBPOPULATION ALLOCATION STRATEGIES

Fun & Dimension	Metrics	DRSAE(10,90)	DRSAE(50,50)	DRSAE(90,10)	DRSAE(dynamic)
F1(50)	**Mean**	3.51e-02	5.61e-02	1.08e-02	**1.51e-26**
	Std deviation	4.78e-02	9.67e-02	6.70e-03	**2.46e-26**
F5(50)	**Mean**	1.05e+01	1.05e+01	3.64e-01	**0.00e-00**
	Std deviation	7.99e+00	1.95e+01	5.90e-01	**0.00e-00**
F1(100)	**Mean**	2.48e-02	1.84e-02	3.21e-02	**8.99e-13**
	Std deviation	4.82e-02	3.09e-02	2.09e-02	**3.12e-12**
F5(100)	**Mean**	2.28e+00	8.98e-01	1.77e-01	**2.22e-11**
	Std deviation	4.37e+00	2.57e+00	8.26e-02	**8.87e-11**
F1(200)	**Mean**	8.24e-04	5.62e-03	4.35e-02	**1.79e-04**
	Std deviation	2.41e-04	2.04e-03	1.04e-02	**2.45e-04**
F5(200)	**Mean**	1.54e-03	1.24e-02	3.41e-01	**2.26e-04**
	Std deviation	1.08e-03	1.17e-02	2.58e-01	**5.15e-04**

TABLE IV. COMPARISON RESULTS OF 7 ALGORITHMS ON TEST FUNCTIONS F1-F7 ACROSS DIMENSIONS 50 TO 20

Fun	Dim	Metrics	SA_COSO	SHPSO	ESAO	SAMSO	GSGA	SAEO	DRSAE
F1	50	**Mean**	4.66e+01(+)	7.17e+00(+)	7.40e-01(+)	5.13e-01(+)	6.21e-01(+)	1.51e-26(≈)	**1.21e-27**
		Std devi	1.74e+01	2.42e+00	5.55e-01	2.85e-01	4.84e-01	2.46e-26	**2.46e-22**
	100	**Mean**	1.03e+03(+)	7.61e+01(+)	1.28e+03(+)	7.21e+01(+)	1.23e+01(+)	**8.99e-13(≈)**	5.99e-12
		Std devi	3.17e+02	2.14e+01	1.34e+02	1.78e+01	9.39e+00	**3.12e-12**	6.22e-11
	200	**Mean**	1.63e+04(+)	2.35e+03(+)	1.76e+04(+)	1.52e+03(+)	3.14e+03(+)	1.79e-04(+)	**3.70e-05**
		Std devi	2.98e+03	3.25e+03	1.17e+03	2.12e+02	6.14e+03	2.45e-04	**1.23e-04**
F2	50	**Mean**	2.53e+02(+)	5.13e+01(+)	**4.74e-02(-)**	5.01e+01(+)	4.82e+01(+)	4.89e+01(+)	4.83e-02
		Std devi	5.67e+01	2.00e+00	**1.71e+00**	7.68e-01	7.66e-01	1.66e-02	4.16e+00
	100	**Mean**	2.71e+03(+)	1.65e+02(+)	5.79e+02(+)	2.86e+02(+)	1.09e+02(+)	9.88e+01(≈)	**7.62e+01**
		Std devi	1.17e+02	2.63e+01	4.48e+01	5.25e+01	1.17e+01	2.83e-02	**2.41e-02**
	200	**Mean**	1.64e+04(+)	2.48e+03(+)	4.31e+03(+)	1.15e+03(+)	4.58e+02(+)	1.98e+02(+)	**1.02e+02**
		Std devi	4.09e+03	1.96e+02	2.84e+02	1.16e+02	1.16e+02	5.43e-02	**1.36e-02**
F3	50	**Mean**	8.86e+00(+)	2.60e+00(+)	1.43e+00(+)	1.53e+00(+)	2.16e-02(+)	8.55e-14(≈)	**2.56e-15**
		Std devi	1.10e+00	2.48e-01	2.49e-01	4.36e-01	2.37e-02	4.42e-13	**3.51e-13**
	100	**Mean**	1.57e+01(+)	4.11e+00(+)	1.04e+01(+)	6.12e+00(+)	1.31e+00(+)	6.96e-07(≈)	**5.12e-07**
		Std devi	5.02e-01	5.92e-01	2.11e-01	4.09e-01	9.68e-01	1.05e-06	**3.25e-06**
	200	**Mean**	1.78e+01(+)	2.13e+01(+)	1.46e+01(+)	1.20e+01(+)	2.20e+01(+)	1.86e-03(+)	**2.14e-04**
		Std devi	2.23e-02	1.02e-01	2.19e-01	4.00e-01	6.20e-01	1.22e-03	**1.94e-03**
F4	50	**Mean**	5.63e+00(+)	9.45e-01(+)	9.40e-01(+)	6.66e-01(+)	3.46e-01(+)	**6.92e-15(≈)**	8.65e-15
		Std devi	8.92e-01	5.39e-02	4.21e-02	1.07e-01	7.15e-02	**8.85e-14**	6.11e-14
	100	**Mean**	6.33e+01(+)	1.07e+00(+)	5.73e+01(+)	1.06e+00(+)	7.06e-01(+)	**1.39e-07(-)**	8.65e-06
		Std devi	1.90e+01	2.04e-02	5.84e+00	2.64e-02	7.06e-02	**8.79e-06**	1.25e-05
	200	**Mean**	5.77e+02(+)	3.14e+02(+)	5.72e+02(+)	9.03e+00(+)	1.03e+01(+)	3.79e-02(+)	**2.85e-03**
		Std devi	1.01e+02	6.58e+01	3.60e+01	1.33e+00	1.69e+00	3.16e-06	**6.32e-03**
F5	50	**Mean**	3.22e+02(+)	3.89e+02(+)	4.21e+02(+)	3.77e+01(+)	1.53e+00(+)	0.00e-00(-)	**15.99e-12**
		Std devi	3.83e+01	6.27e+01	1.24e+01	7.82e+00	4.36e-01	0.00e-00	**6.22e-11**
	100	**Mean**	8.81e+02(+)	8.78e+02(+)	4.15e+02(+)	4.29e+02(+)	6.55e+02(+)	2.22e-11(≈)	**6.15e-12**
		Std devi	7.01e+01	8.93e+01	6.73e+01	5.72e+01	6.31e+01	8.87e-11	**6.24e-11**
	200	**Mean**	5.77e+02(+)	5.72e+02(+)	3.14e+02(+)	9.03e+00(+)	1.03e+01(+)	3.79e-02(+)	**2.85e-03**
		Std devi	1.01e+02	3.60e+01	6.58e+01	1.33e+00	1.69e+00	3.16e-06	**6.32e-03**
F6	50	**Mean**	2.35e+02(−)	1.22e+02(−)	1.99e+02(−)	8.69e+02(+)	**7.58+01(−)**	7.71e+02(+)	6.66e+02
		Std devi	4.09e+01	2.59e+01	4.58e+01	3.17e+01	**4.99e+01**	5.89e+01	6.70e+01
	100	**Mean**	1.27e+03(+)	8.01e+02(−)	7.13e+02(−)	7.37e+02(−)	**6.72e+02(−)**	2.02e+03(+)	1.85e+03
		Std devi	1.17e+02	7.22e+01	2.65e+01	4.20e+01	**2.97e+01**	1.34e+02	9.16e+01
	200	**Mean**	3.92e+03(−)	4.15e+03(−)	5.38e+03(+)	4.96e+03(+)	**2.56e+03(−)**	4.80e+03(≈)	4.73e+03
		Std devi	2.72e+02	2.98e+02	1.56e+02	1.38e+02	**2.68e+02**	2.19e+02	2.89e+02
F7	50	**Mean**	1.08e+03(+)	1.00e+03(+)	9.75e+02(+)	9.70e+02(+)	9.70e+02(+)	9.10e+02(≈)	9.10e+02
		Std devi	3.66e+01	2.12e+01	3.71e+01	2.92e+01	1.81e+01	4.71e-03	**0.00e+00**
	100	**Mean**	1.36e+03(+)	1.41e+03(+)	1.37e+03(+)	1.29e+03(+)	1.25e+03(+)	9.10e+02(≈)	**9.10e+02**
		Std devi	3.08e+01	3.82e+01	2.75e+01	3.34e+01	2.45e+01	1.39e-02	**1.65e-09**
	200	**Mean**	1.34e+03(+)	6.27e+02(+)	1.45e+03(+)	1.34e+03(+)	5.25e+03(+)	9.10e+02(+)	**4.16e+01**
		Std devi	2.46e+01	1.15e+01	2.04e+01	2.43e+01	1.87e+01	5.72e-04	**2.79e-04**

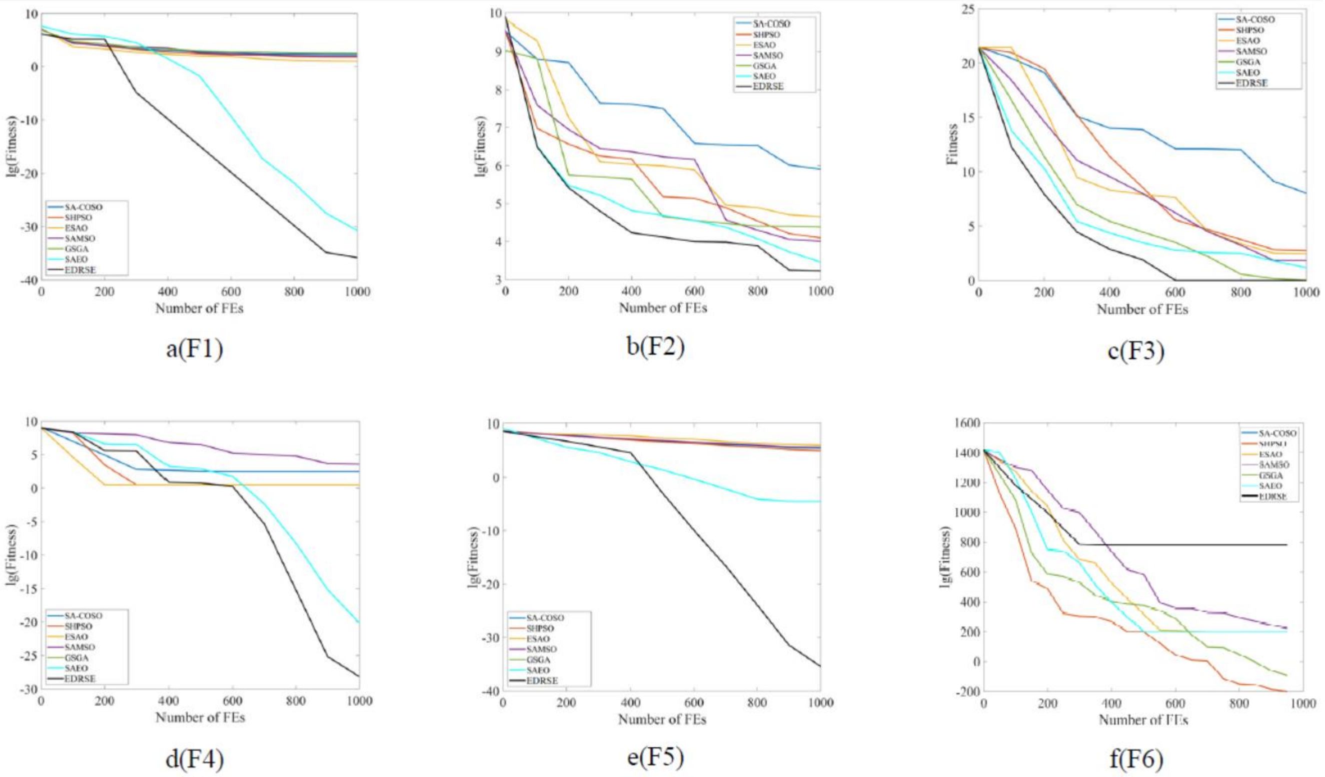
Table IV demonstrates that DRSAE surpasses most comparable algorithms in efficiently identifying the vicinity of the optimal solution within a limited number of evaluations, while many others fall short of reaching the optimum. Notably, DRSAE exhibits strong performance on F5 with a minimum value characteristic indicative of a regular distribution. Among these, SAEO [7] shows relatively superior performance due to its utilization of dimensionality reduction. Furthermore, the ELM-AE dimensionality reduction network employed by DRSAE accurately restores high-dimensional space without information loss, resulting in overall better performance compared to SAEO. ESAO showcases commendable exploration ability owing to differential evolution, leading to superior performance on F2 with 50 dimensions compared to DRSAE.

Plot convergence curves for the 100D and 200D problems, as illustrated in Fig. 6 and Fig. 7, respectively. Analysis of figures indicates that DRSAE demonstrates faster convergence to the optimal solution compared to other surrogate-assisted algorithms such as SAEO. With increasing dimensionality, DRSAE exhibits superior optimization speed relative to other algorithms due to its enhanced capability in identifying promising regions within low-dimensional space. GSGA [13] performs exceptionally well on F6, a complex multimodal function, by iteratively employing alternative functions at the cost of computational burden to enhance opportunities for finding the optimal solution. The suboptimal performance of DRSAE on F6 is primarily attributed to the inability of dimensionality reduction in simplifying the complexity of F6's landscape. Overall, DRSAE ranks first in average for both optimization speed and convergence performance, thus fully demonstrating its superiority. Upon activation of the surrogate model, fitness value sharply decreases within the same range of function evaluations.

E. *Effectiveness Analysis of Dimensionality Reduction Strategies*

To assess the efficacy of incorporating a low-dimensional search space, this study selects the F1 and F5 functions with dimensions ranging from 50 to 200 for evaluating both DRSAE and its non-low-dimensional version, DRSAE*. The optimization results for DRSAE and DRSAE* are presented in Table V.

Drawing upon the findings presented in Table V, it is evident that ELM-AE has significantly augmented the search efficiency of DRSAE* while maintaining an equivalent number of FEs. The results signify a noteworthy reduction in the average optimal values, leading to expedited optimization speed for DRSAE. ELM-AE effectively steers the exploration of low-dimensional search space, integrates high-dimensional information, and extracts latent features, thereby facilitating rapid identification of promising regions and substantial reduction in unnecessary FE consumption, particularly for problems with potential convergence issues. For problems up to 200 dimensions, leveraging the low-dimensional search space in ELM-AE has widened the gap between optimal values, demonstrating its advantage in precise and efficient dimensionality reduction.



a(F1)

b(F2)

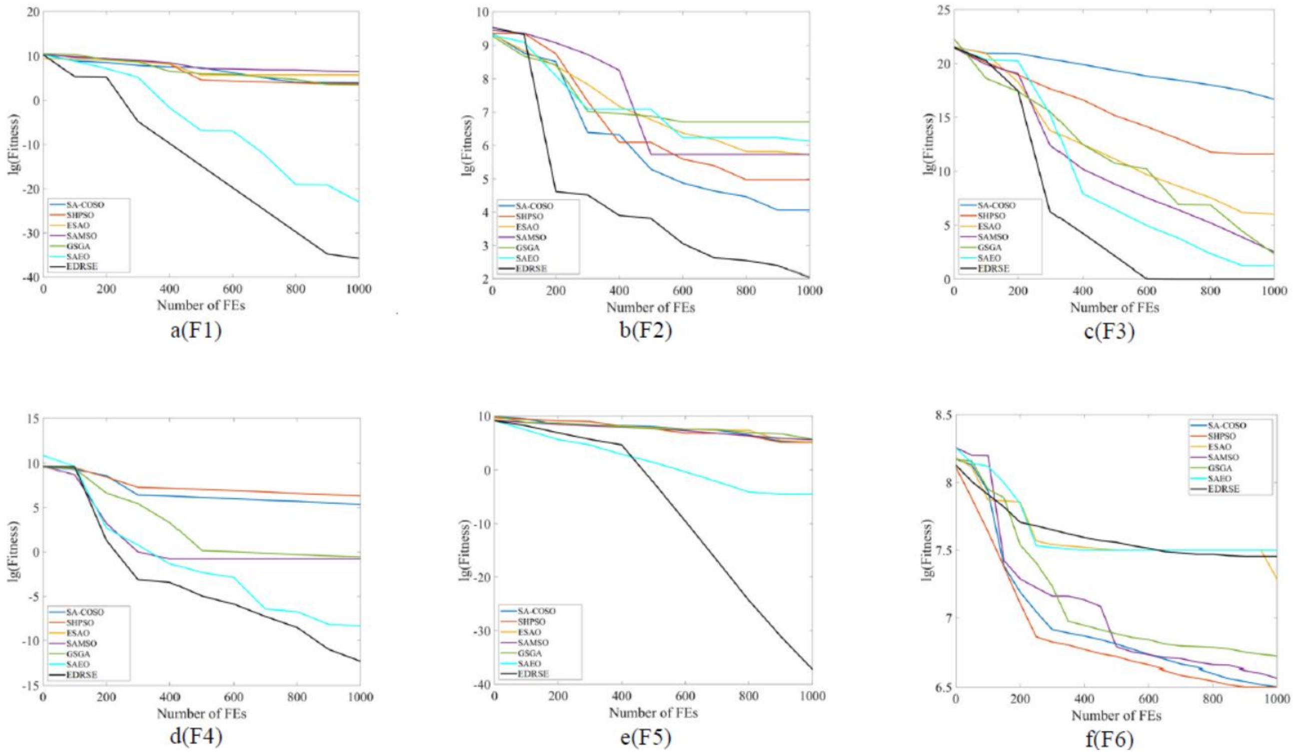
c(F3)

d(F4)

e(F5)

f(F6)

Fig. 6. Convergence curves of different algorithms on 100D benchmark problems.



a(F1) b(F2) c(F3)

d(F4) e(F5) f(F6)

Fig. 7. Convergence curves of different algorithms on 200D benchmark problems.

TABLE V. OPTIMIZATION RESULTS OF DRSAE* AND DRSAE

Fun & Dimension	Metrics	DRSAE*	DRSAE
F1(50)	**Mean**	4.18e-03	**1.51e-26**
	Std deviation	1.49e-03	**2.46e-26**
F5(50)	**Mean**	8.06e-02	**0.00e+00**
	Std deviation	7.11e-02	**0.00e+00**
F1(100)	**Mean**	1.55e-02	**8.99e-13**
	Std deviation	7.42e-03	**3.12e-12**
F5(100)	**Mean**	8.51e-02	**2.22e-11**
	Std deviation	4.53e+00	**8.87e-11**
F1(200)	**Mean**	1.75e+01	**1.79e-04**
	Std deviation	1.52e+01	**2.45e-04**
F5(200)	**Mean**	8.81e+02	**2.26e-04**
	Std deviation	5.14e+02	**5.15e-04**

V. CONCLUSION

This paper introduces an effective dimensionality reduction assisted evolutionary framework (DRSAE) for addressing high-dimensional expensive evaluation function problems (HEPs). The primary challenge in HEPs lies in the high cost of evaluating the function, necessitating the rapid identification of optimal solutions within a limited number of function evaluations. The algorithm makes two key contributions: 1) Incorporating an efficient and accurate low-dimensional search space into the traditional agent-based algorithm, enabling precise reduction and reconstruction of high-dimensional space to minimize position information errors resulting from switching between spaces and expedite the discovery of a more promising solution space; 2) Implementing hierarchical learning of particles in high-dimensional space, allowing lower-level particles to learn from superior ones to enhance population diversity. Furthermore, specific activation conditions for the agent model are established for different dimensional problems.

In order to assess the performance of DRSAE, it was compared with other established algorithms across seven commonly utilized functions. The experimental results indicate that in most cases, DRSAE performs admirably, with SAEO demonstrating relatively superior performance. This can be attributed to the fact that SAEO also operates within a low-dimensional search space; however, its low-dimensional model is based on AE. In contrast to DRSAE, where ELM-AE produces a smaller reconstruction error from the low-dimensional space to the high-dimensional space. As a result, DRSAE is able to more accurately reconstruct features in the high-dimensional space and achieve improved performance. In addition, better results may be achieved by improving the ELM network structure to stack ELM hidden layers.

VI. FUTURE WORK

We endeavor to integrate state-of-the-art Evolutionary Algorithms (EAs) [19-23] into the DRSAE in order to tackle High-dimensional Expensive Problems (HEPs), while exploring the underlying theory in future research. It is worth noting that this framework has the potential for extension to address multi-objective optimization problems, dynamic optimization problems, and constrained optimization problems, thereby validating the effectiveness of DRSAE in relevant real-world scenarios and expanding its applicability in high-dimensional expensive optimization domains.

REFERENCES

[1] Hu Rong, Chen Wenbo, Qian Bin, Guo Ning, and Xiang Fenghong. "Learning more than ant colony algorithm to solve the green yard vehicle routing problem". *Journal of System Simulation*, China,33(09):2095-2108, 2021.

[2] Li Han, Du Peng, Du Ying, et al. "Multi-path routing selection method for wireless body area networks based on genetic algorithm". *Journal of Jilin University (Engineering and Technology Edition)*, China, 52(11): 2706-2711, 2022.

[3] Yang Hua. "Research on feature gene selection method based on particle swarm algorithm." *Hunan University*, China, 2010.

[4] Xu Kangyu, Liu Yuan, Li Miqing, Yang Shengxiang, Zou Juan, and Zheng Jinhua. "A review of evolutionary high-dimensional multi-objective optimization." *Control Engineering*, China, 30(08): 1436-1449, 2023.

[5] Zhao Yuxiang, Li Qiang, Liu Ziyu. "Application of gaussian process surrogate Model in Large-Scale Global Optimization." *Control and Decision*, China, 36(3): 577-583, 2021.

[6] C. Sheng, J. Hundley. "Data-dimensionality reduction. " *Whitman College*,https://www.whiman.edu/Docments/Acadeics/Mathematics/2019/Sheng-Hundley.pdf. USA, 2019.

[7] M. Cui, L. Li, M. Zhou and A. Abusorrah, "Surrogate-assisted autoencoder-embedded evolutionary optimization algorithm to solve high-dimensional expensive problems," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 4, pp. 676-689, Aug. 2022.

[8] F. Li, X. Cai, L. Gao and W. Shen, "A Surrogate-Assisted Multiswarm Optimization Algorithm for High-Dimensional Computationally Expensive Problems," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1390-1402,2020.

[9] Huang Guangbin, Song Shiji, and You Kai. "Trends in extreme learning machines." *Neural Networks*, 61: 32-48, 2015.

[10] Zhang Ting and Yu Li. "Extreme learning machine: algorithms and applications." *Neural Computing and Applications*, 32(11):7041-7059, 2020.

[11] R. G. Regis, "Evolutionary programming for high-dimensional constrained expensive black-box optimization using radial basis functions," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 3, pp. 326-347, June 2014.

[12] R. Mendes and A. S. Mohais, "DynDE: a differential evolution for dynamic optimization problems," *IEEE Congress on Evolutionary Computation*, Edinburgh, UK, pp. 2808-2815 Vol. 3, 2005.

[13] X. Cai, L. Gao and X. Li, "Efficient generalized surrogate-assisted evolutionary algorithm for high-dimensional expensive problems," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 365-379, April 2020.

[14] C. Sun, Y. Jin, R. Cheng, J. Ding and J. Zeng, "Surrogate-assisted cooperative swarm optimization of high-dimensional expensive problems," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 4, pp. 644-660, Aug. 2017.

[15] X. Wang, G. G. Wang, B. Song, P. Wang and Y. Wang, "A novel evolutionary sampling assisted optimization method for high-dimensional expensive problems," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 815-827, Oct. 2019.

[16] Haibo Yu, Ying Tan, Jianchao Zeng, Chaoli Sun, and Yaochu Jin. "Surrogate-assisted hierarchical particle swarm optimization," *Information Sciences*, 454(2):59–72. 2018.

[17] F. Li, X. Cai, L. Gao and W. Shen, "A surrogate-assisted multiswarm optimization algorithm for high-dimensional computationally expensive problems," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1390-1402, March 2021.

[18] Sun, C., Ding, J., and Zeng, J. "A fitness approximation assisted competitive swarm optimizer for large scale expensive optimization problems. "*Memetic Comp.*10(2), 123–134, 2018.

[19] J. Zhang, P. Wen and A. Xiong, " Application of improved quantum particle swarm optimization algorithm to multi-task assignment for heterogeneous UAVs,"*2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Changzhou, China, pp. 1-5, 2022.

[20] K. Gao, Z. Cao, L. Zhang, Z. Chen, Y. Han and Q. Pan, "A review on swarm intelligence and evolutionary algorithms for solving flexible job shop scheduling problems," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 4, pp. 904-916, July 2019.

[21] Y. Yu, S. Gao, Y. Wang and Y. Todo, "Global optimum-based search differential evolution," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 2, pp. 379-394, March 2019.

[22] W. Qingling and J. Yubo, "Research on improved particle swarm optimization algorithm based on simulated annealing algorithm,"*2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, Ottawa, ON, Canada, pp. 360-362,2023.

[23] T. Wang, W. Miao and Z. Zeng, "Optimization method of Data interaction in power IoT based on particle swarm algorithm," *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, pp. 351-355, 2022.

Optimization of the Energy-Saving Data Storage Algorithm for Differentiated Cloud Computing Tasks

Optimization of the Energy-Saving Data Storage Algorithm

Peichen Zhao
School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

***Abstract*—This study presents a novel energy-saving data storage algorithm designed to enhance data storage efficiency and reduce energy consumption in cloud computing environments. By intelligently discerning and categorizing various cloud computing tasks, the algorithm dynamically adapts data storage strategies, resulting in a targeted optimization methodology that is both devised and experimentally validated. The study findings demonstrate that the optimized model surpasses comparative models in accuracy, precision, recall, and F1-score, achieving peak values of 0.863, 0.812, 0.784, and 0.798, respectively, thereby affirming the efficacy of the optimized approach. In simulation experiments involving tasks with varying data volumes, the optimized model consistently exhibits lower latency compared to Attention-based Long Short-Term Memory Encoder-Decoder Network and Deep Reinforcement Learning Task Scheduling models. Furthermore, across tasks with differing data volumes, the optimized model maintains high throughput levels, with only marginal reductions in throughput as data volume increases, indicating sustained and stable performance. Consequently, this study is pertinent to cloud computing data storage and energy-saving optimization, offering valuable insights for future research and practical applications.**

***Keywords*—Energy-saving data storage algorithm; differentiated task recognition; cloud computing; intelligent storage strategy; data classification and distribution**

I. Introduction

With the rapid development of cloud computing, the demand for data storage has surged, rendering traditional centralized storage solutions inadequate for managing vast data volumes and diverse task requirements [1]. The data storage landscape in cloud computing is characterized by its large scale and the variety of data types, encompassing different task types and service models, such as big data analysis, real-time processing, and archival storage. Each task possesses unique resource requirements and storage characteristics, leading to significant variations in performance requirements across tasks [2-4]. To enhance overall performance, data storage algorithms in cloud computing must balance data accuracy, reliability, real-time performance, and energy consumption. However, existing storage strategies frequently overlook the distinctions between diverse tasks, resulting in wasted storage resources, performance degradation, and elevated energy consumption.

Previous studies have highlighted the substantial differences in resource requirements among various task types in cloud computing environments. For instance, Yang et al. demonstrated that these disparities significantly influence the selection and optimization of data storage strategies [5]. Saravanan et al. conducted an analysis of big data and real-time processing tasks, emphasizing fundamental differences in data access patterns and response time requirements, thereby underscoring the necessity for targeted algorithm design [6]. Additionally, Manukumar and Muthuswamy elucidated the variations in persistence and reliability requirements between archival storage and high-frequency access data, providing a theoretical foundation for the diversification of data storage strategies [7]. Furthermore, Hamid et al. proposed an energy-saving storage algorithm predicated on data access frequency, which markedly reduced energy consumption in data centers through intelligent data migration and caching strategies [8]. Zhang et al. developed a storage management system based on data hotness and task priority, effectively enhancing the utilization of storage resources and decreasing the energy consumption associated with redundant data [9]. Finally, Rahimikhanghah et al. conducted simulation experiments to compare the performance of various data storage algorithms under diverse cloud computing tasks, revealing that targeted optimization algorithms exhibit significant advantages in terms of accuracy and response time [10]. El-Menbawy et al. validated the improvements in throughput and storage efficiency of their proposed energy-saving storage algorithm through testing in a real cloud environment, providing robust data to support its practical application [11]. Dong et al. assessed the performance of various storage algorithms using metrics such as accuracy, recall, F1-score, and precision. Their findings indicated that algorithms that comprehensively accounted for task differences demonstrated superior performance across multiple indicators [12]. Al-Masri et al. applied machine learning techniques to predict and adjust the performance of storage systems, optimizing data distribution and replication strategies while minimizing energy consumption, all while maintaining data reliability [13].

Although previous studies have recognized the differences in cloud computing tasks, many have lacked in-depth analysis of their specific characteristics and requirements, leading to suboptimal performance of storage algorithms when faced with diverse task types. Moreover, most energy-saving storage algorithms are typically designed for specific tasks or scenarios, making it difficult for them to adapt flexibly to varied task environments, thereby diminishing overall performance and energy-saving effectiveness. By conducting a detailed analysis of the data characteristics and performance requirements of different cloud computing tasks, this study provides a clear optimization framework and decision-making basis for the design of energy-saving data storage algorithms. Additionally, the proposed algorithm can identify and adapt to diverse task environments, dynamically adjusting data storage strategies according to task types, thereby achieving an optimal balance between energy efficiency and performance under different task scenarios. This study first analyzes the differences in cloud computing tasks, revealing that a detailed analysis and understanding of these tasks can provide clearer directions for optimizing energy-saving data storage algorithms, enabling them to adapt to task variability and achieve a balance between energy efficiency and performance in diverse task environments. Secondly, the design of the energy-saving data storage algorithm is studied, emphasizing that through the comprehensive application of these strategies, the algorithm can effectively meet the demands of different cloud computing tasks, ensuring an optimal balance between performance, reliability, and energy consumption in data storage systems. Furthermore, by incorporating strategies for data distribution, load balancing, data replication, fault tolerance, data access, migration, and energy consumption optimization with performance evaluation, the model is refined. Finally, the effectiveness of the model is validated through experiments.

II. Optimization of the Energy-Saving Data Storage Algorithm for Differentiated Cloud Computing Tasks

A. Analysis of Differentiated Cloud Computing Tasks

Before designing and optimizing energy-saving data storage algorithms tailored to differentiated cloud computing tasks, a comprehensive analysis of the characteristics and requirements of each task is essential. Additionally, it is critical to establish a clear mapping between data storage models and task types, while identifying the key factors that influence the selection of storage algorithms [14-16]. The variety of cloud computing tasks is considerable, and the common task types are presented in Table I:

TABLE I. Types of Cloud Computing Tasks

Task types	Description
Real-time processing tasks	Real-time tasks, such as financial transactions and online games, require extremely low latency to guarantee a quick response to user requests. They often rely on caching strategies that ensure efficient reading and immediate updating of data.
Big data analysis tasks	They involve large-scale data processing and analysis, such as data mining, machine learning, and business intelligence. These tasks require high storage performance and require fast data transfer between distributed storage systems to meet computing and analysis requirements.
Archive storage tasks	For instance, long-term storage of logs, backups, and legal files requires high data persistence and security, but relatively low access speed. Data backup, fault tolerance, and tiered storage are often required to balance cost and performance.
Content delivery tasks	Examples include video streaming and large file downloads, emphasizing high throughput and efficiency of data distribution. Content Delivery Network (CDN) and multi-layer caching are needed to accelerate content distribution.

A well-designed data storage model is crucial for meeting the performance requirements of different tasks. Cache modes are commonly employed for real-time processing and content delivery tasks, with the primary goal of minimizing data access latency while dynamically adjusting data distribution across various cache layers. Distributed file systems are typically utilized in large-scale data processing and transmission, ensuring reliable data distribution and fast access for big data analysis tasks [17-18]. For archival storage tasks, hierarchical storage strategies are applied, utilizing cold and hot data layers to manage data accessed at varying frequencies, thereby reducing overall storage costs while maintaining data persistence and security. Replication strategies are adapted based on task characteristics: real-time processing tasks often require leader-follower replication to ensure low-latency responses, while big data analysis tasks generally employ distributed replication with high redundancy to guarantee data reliability and persistence [19-21]. The selection and optimization of storage algorithms are influenced by a range of critical factors, depending on the task types and storage modes, as summarized in Table II:

TABLE II. Key Factors Influencing the Selection of Storage Algorithms

Factor	Analysis
Task priority	Various tasks have different priorities in the system. High-priority tasks such as financial transactions should be allocated more resources, while low-priority archiving tasks can use more energy-efficient storage strategies.
Data access mode	The data access mode of the task directly affects the design of the storage algorithm. The frequent read and write requirements of real-time processing tasks are different from the sequential batch processing of big data analysis tasks, and differentiated caching and migration strategies are required.
Data consistency	Different tasks have diverse requirements for data consistency. Real-time tasks require strong consistency, while analytical tasks can accept some degree of ultimate consistency.
Energy consumption and cost	Low energy consumption and storage costs are the focus of most missions. According to the performance and budget requirements of different tasks, a proper storage strategy can effectively reduce the cost of data migration and redundant copies.

A detailed analysis and understanding of various cloud computing tasks can provide clearer guidance for optimizing energy-saving data storage algorithms. This allows them to adapt to task variability and achieve a balance between energy efficiency and performance across diverse task environments [22].

B. Design of the Energy-Saving Data Storage Algorithm

The design of the energy-saving data storage algorithm adheres to the following principles and objectives. First, the algorithm must be adaptive, capable of recognizing and responding to the diverse requirements of different cloud computing tasks, and dynamically adjusting data storage and distribution strategies based on the tasks' characteristics and priorities. Second, reducing energy consumption in data centers through efficient data migration and replication strategies is a primary goal, focusing on minimizing hardware resource idleness and reducing redundant data replication. Third, the algorithm must ensure data availability and persistence across various task scenarios, maintaining data consistency even under high-load conditions. Lastly, it is essential to balance performance metrics such as response time, throughput, and accuracy while addressing the diverse demands of different tasks. This study proposes an energy-saving data storage algorithm designed to accommodate the diversity of cloud computing tasks while minimizing energy consumption. The framework of the proposed algorithm comprises several key modules, as illustrated in Fig. 1.

The identification and classification of differentiated task data form the foundational basis of the energy-saving storage algorithm. By analyzing task-specific characteristics, such as real-time requirements, priority levels, and data access modes, the data can be categorized into distinct types, as outlined in Table III:

TABLE III. DATA TYPES

Type	Analysis
High-priority real-time data	This type of data is used for real-time processing tasks, usually has high priority and low latency requirements, and is mainly stored in the cache layer to meet fast access demands.
Batch analysis data	It is suitable for big data analysis tasks that need to maintain the reliability and distribution of data in a distributed file system to ensure efficient computation and analysis.
Long-term archived data	The data access frequency is low and data is stored at the cold layer. The hierarchical storage strategy minimizes space and power consumption.

Building on the identification and classification of differentiated task data, appropriate energy-saving data storage strategies can be developed. First, the intelligent data migration strategy dynamically adjusts the distribution of data across various storage system tiers based on task characteristics and data access patterns. This ensures that high-frequency data is cached, while low-frequency data is archived, optimizing the hierarchical management of data. Second, the replication strategy is determined by task type and priority. High-priority tasks utilize a leader-follower replication model to meet low-latency response requirements, whereas batch analysis tasks implement a multi-replica distribution strategy to guarantee data reliability and consistency. In addition, the resource balancing strategy employs load-balancing algorithms to distribute access loads evenly across data nodes, mitigating resource idleness and avoiding bottlenecks caused by hot data, thereby enhancing resource utilization. Moreover, the dynamic optimization strategy continuously monitors the storage system's performance and energy consumption in real-time, using machine learning algorithms to optimize data storage strategies and adapt to evolving task demands. Through the comprehensive implementation of these energy-saving data storage strategies, the proposed algorithm effectively addresses the requirements of diverse cloud computing tasks, ensuring an optimal balance between performance, reliability, and energy efficiency in the data storage system.

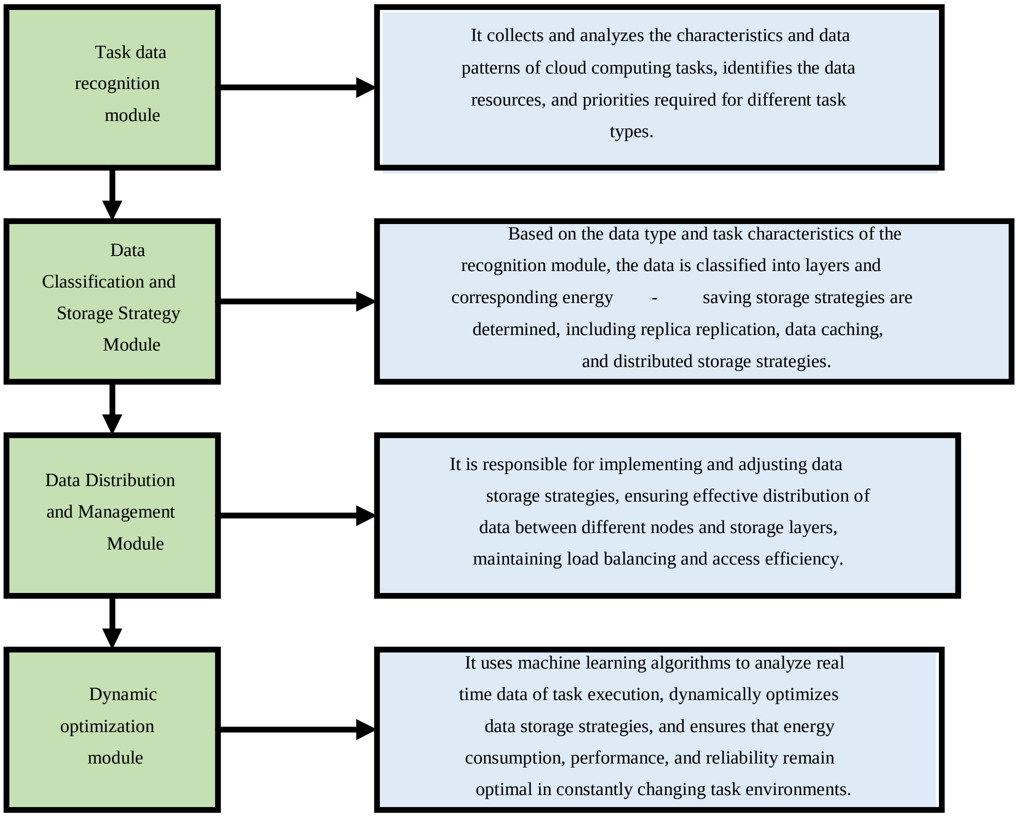

Task data recognition module
It collects and analyzes the characteristics and data patterns of cloud computing tasks, identifies the data resources, and priorities required for different task types.
Data Classification and Storage Strategy Module
Based on the data type and task characteristics of the recognition module, the data is classified into layers and corresponding energy - saving storage strategies are determined, including replica replication, data caching, and distributed storage strategies.
Data Distribution and Management Module
It is responsible for implementing and adjusting data storage strategies, ensuring effective distribution of data between different nodes and storage layers, maintaining load balancing and access efficiency.
Dynamic optimization module
It uses machine learning algorithms to analyze real time data of task execution, dynamically optimizes data storage strategies, and ensures that energy consumption, performance, and reliability remain optimal in constantly changing task environments.

Fig. 1. The framework of energy-saving data storage algorithms.

C. Algorithm Optimization and Implementation

The optimization and implementation of the energy-saving data storage algorithm are designed to enhance the efficiency and reliability of the data storage system while minimizing energy consumption, all without compromising performance. Central to cloud computing data storage are effective data distribution strategies, which significantly influence the system's performance and stability. The algorithm proposed here is founded on three key principles: task priority, data hotness, and load balancing. Task data is allocated to various storage nodes based on the real-time nature and priority of the tasks, facilitating hierarchical management of data access. High-priority tasks with stringent real-time requirements are directed to nodes that offer faster response times, while lower-priority tasks are assigned to nodes with slower response capabilities. By analyzing the frequency and hotness of data access, high-demand (hot) data is stored in the high-speed cache layer to ensure efficient access, whereas low-demand (cold) data is migrated to the cold storage layer to alleviate pressure on the cache. The load balancing algorithm dynamically adjusts the data distribution across nodes to achieve an equitable distribution of access loads among all storage nodes, thus preventing resource idleness and mitigating the occurrence of bottlenecks due to hot data. Furthermore, replication and fault tolerance mechanisms within the data storage framework ensure data persistence and reliability, as illustrated in Table IV.

TABLE IV. DATA REPLICAS AND FAULT TOLERANCE MECHANISMS

Dimension	Analysis
Copy replication strategy	According to the requirements of different tasks, various copy replication strategies are designed. Leader-follower replication is adopted for high-priority tasks to ensure real-time performance and fast recovery. Multiple replicas are used for batch processing and archiving tasks to ensure reliable data recovery in case of faults.
Fault tolerance	Regular data integrity checks and replica status monitoring should be implemented to promptly detect and handle storage node failures, restore damaged data to healthy nodes, and ensure data availability and consistency.

Data access and migration strategies significantly influence the performance and flexibility of the data repository. Access optimization employs caching strategies and data layering to refine access pathways in accordance with task type and data popularity, ensuring that frequently accessed data resides in the cache layer while infrequently accessed data is relegated to secondary storage. Dynamic migration adjusts data storage locations in response to evolving access patterns, transferring data between cold and hot layers to meet shifting task requirements and sustain optimal system performance. In pursuit of a comprehensive reduction in energy consumption and an evaluation of storage system performance, this study delineates the following strategies, as outlined in Table V:

TABLE V. OPTIMIZATION STRATEGY

Strategy	Analysis
Energy consumption optimization	By adjusting data distribution, and reducing node idle rates and redundant copies, unnecessary energy consumption is reduced. Meanwhile, energy-saving storage hardware and resource hibernation mechanisms are employed to mitigate system energy consumption while maintaining high performance.
Performance evaluation	Comprehensive performance evaluation indicators are established to identify system bottlenecks and optimize storage strategies, thereby achieving a balance between performance and energy consumption.

By integrating a comprehensive array of data distribution strategies, load balancing, data duplication, fault tolerance mechanisms, data access, migration strategies, energy consumption optimization, and performance assessment, the algorithm proposed herein ensures that the data storage system achieves an optimal amalgamation of high performance and low energy consumption across various task environments.

III. ANALYSIS OF PERFORMANCE AND SIMULATION RESULTS OF ENERGY-SAVING STORAGE ALGORITHMS

A. Analysis of Performance Comparison Results of Energy-Saving Storage Algorithms

The dataset utilized for this experiment is the Alibaba Cluster dataset, a comprehensive resource derived from Alibaba's production cluster, specifically designed to facilitate research on cluster management. The dataset spans multiple versions from 2017 to 2023, offering valuable insights into various facets of cloud computing tasks. It is publicly accessible via the official repository at https://github.com/alibaba/clusterdata. Details of the experimental environment are presented in Table VI.

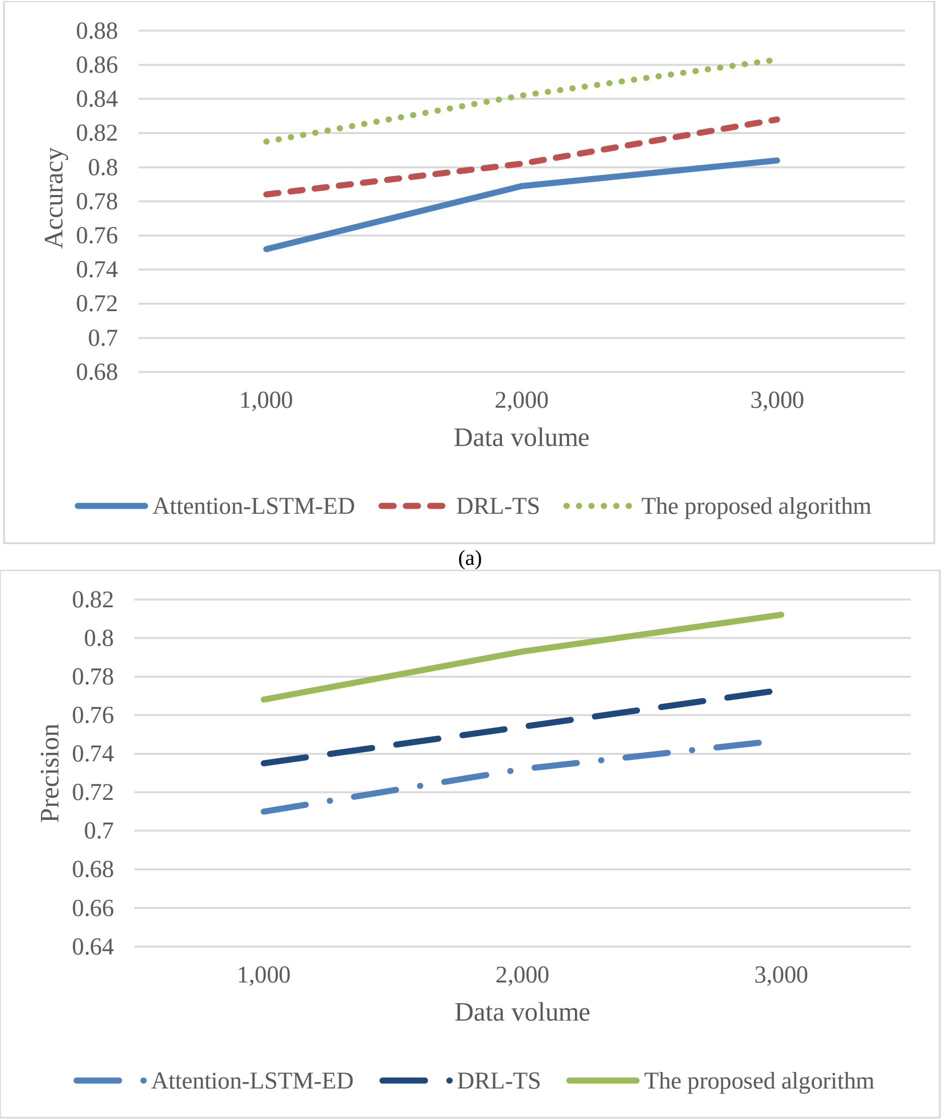
TABLE VI. EXPERIMENTAL ENVIRONMENT

Equipment type	Parameter configuration
Processor	Inter(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
Graphics card	NVIDIA Titan Xp 12GB
Memory	128 GB
Operating system	Ubuntu 16.04 LTS
Programming language	Python 3.6

The model parameters are uniformly configured to ensure experimental accuracy. Specifically, the learning rate is set to 0.01, with a batch size of 64, two hidden layers, each containing 128 hidden units. The Adam optimizer is employed, and training is conducted over 50 epochs with a dropout probability of 0.02. For comparative analysis, the Attention-based Long Short-Term Memory Encoder-Decoder (Attention-LSTM-ED) and Deep Reinforcement Learning Task Scheduling (DRL-TS) models are selected due to their representative and practical applications within the cloud computing domain. The Attention-LSTM-ED model leverages the LSTM network and attention mechanism to effectively handle time series data, making it suitable for applications such as task load prediction. In contrast, the DRL-TS model utilizes deep reinforcement learning for task scheduling, offering an efficient solution to challenges related to resource allocation and task scheduling. These models provide a robust comparison, enabling a thorough evaluation of the proposed model's improvements in terms of performance, efficiency, and energy consumption. The performance metrics compared in this study include accuracy, recall, precision, and F1-score, with the results illustrated in Fig. 2.

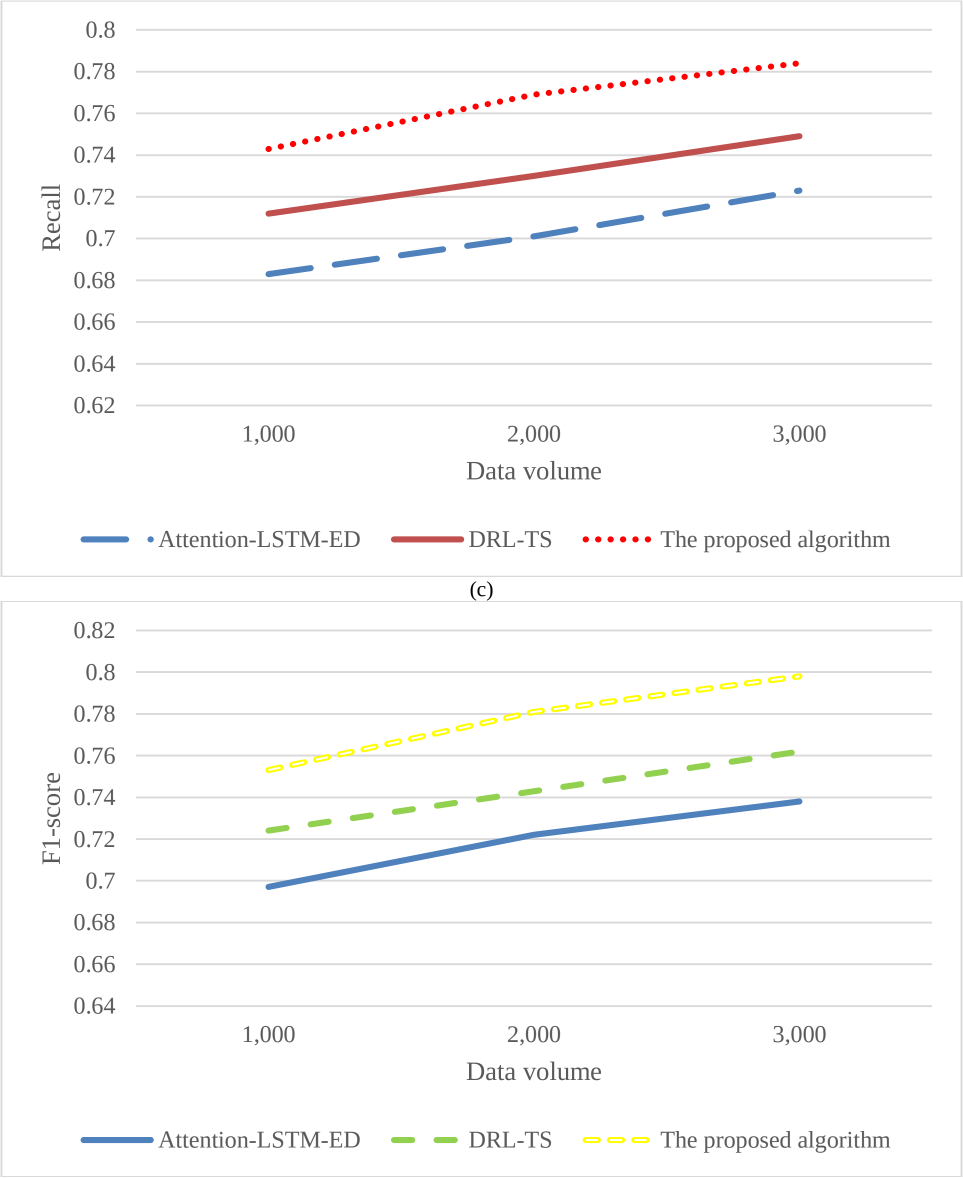
Fig. 2 demonstrates that the accuracy of the optimized model reaches 0.815, 0.842, and 0.863 for data volumes of 1000, 2000, and 3000, respectively, consistently surpassing the performance of the Attention-LSTM-ED and DRL-TS models. This suggests that the optimized model maintains high accuracy across varying data sizes, highlighting its superior generalization ability and stability. While the Attention-LSTM-ED and DRL-TS models exhibit relatively strong performance with smaller data volumes, their accuracy decreases as data volume increases, underscoring the optimized model's advantage in handling large-scale datasets.

In terms of precision, the optimized model consistently achieves stable and high precision levels of 0.768, 0.793, and 0.812 across different data volumes. In comparison, the precision of the Attention-LSTM-ED and DRL-TS models is 0.710 and 0.735, respectively, at a data volume of 1000, and 0.732 and 0.754, respectively, at a data volume of 2000. Overall, the optimized model outperforms the other two models across all data volumes. This performance advantage likely stems from the optimized model's enhanced data recognition and classification capabilities, which allow it to adapt more effectively to varying data volumes and task requirements. Even at smaller data volumes, the optimized model demonstrates efficient precision, which improves further as data volume increases. In contrast, the precision improvement of the Attention-LSTM-ED and DRL-TS models is relatively slow, suggesting that they may struggle to maintain stable performance as data size grows. These comparative results highlight the significant advantage of the optimized model in terms of precision, particularly in large-scale data environments, where it can more accurately process cloud computing tasks.



(a)

(b)



(c)

(d)

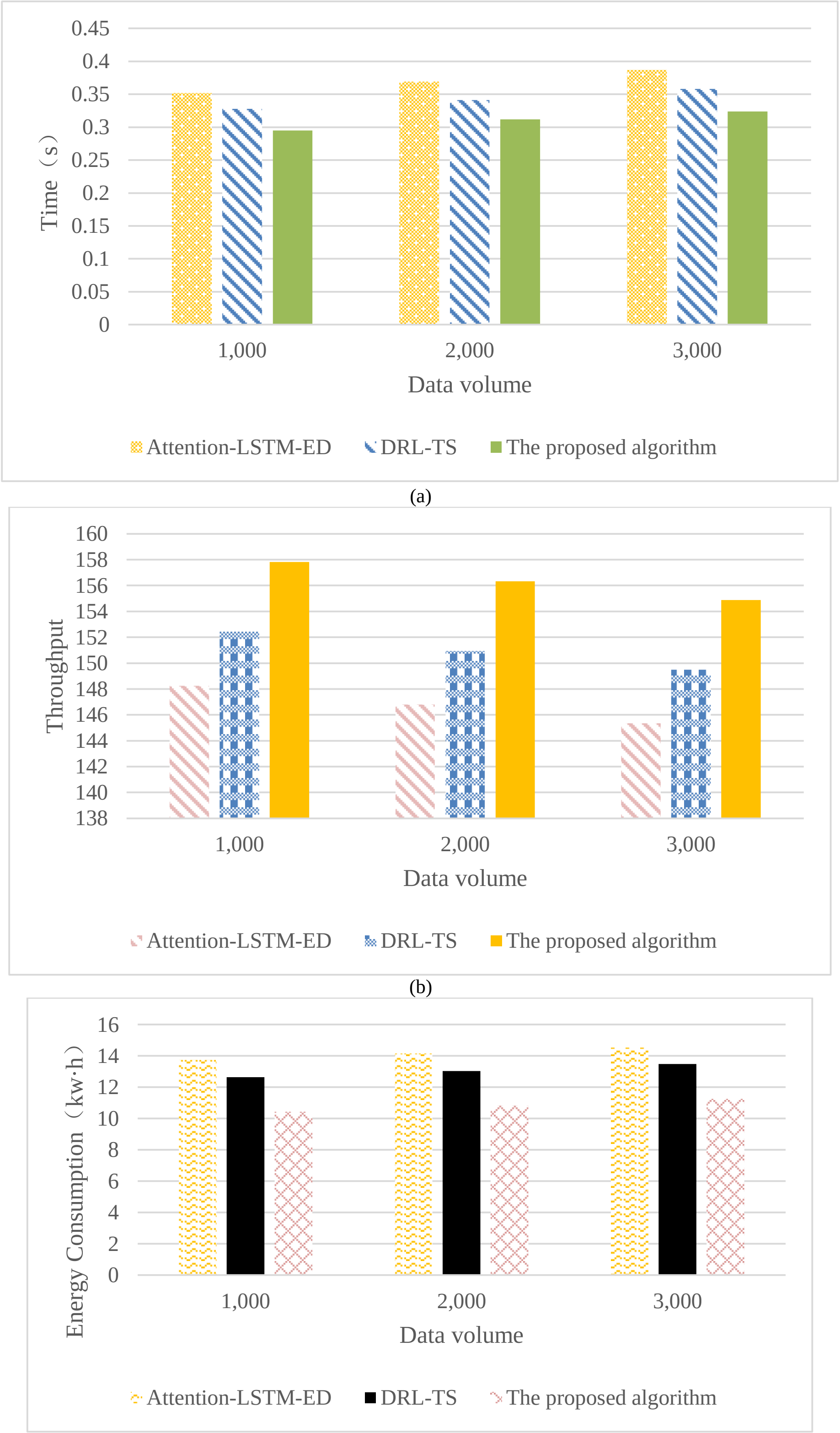
Fig. 2. Performance comparison results (a): Accuracy; (b): Precision; (c): Recall; (d): F1-score.

In terms of recall, the optimized model achieves recall values of 0.743, 0.769, and 0.784 for data volumes of 1000, 2000, and 3000, respectively, consistently outperforming the other models at each data level. For a data volume of 1000, the recall for the Attention-LSTM-ED model is 0.683, while the DRL-TS model achieves 0.712. At a data volume of 2000, their recall values are 0.701 and 0.730, respectively. This consistency indicates that the optimized model exhibits lower sensitivity to varying data scales in terms of recall, allowing it to effectively capture relevant information and enhance data processing performance across a range of task conditions. As data volume increases, the recall of the optimized model improves steadily, demonstrating its robustness and reliability in handling large-scale datasets.

With respect to the F1-score, the optimized model achieves values of 0.753, 0.781, and 0.798 for data volumes of 1000, 2000, and 3000, respectively, outperforming both the Attention-LSTM-ED and DRL-TS models. At a data volume of 1000, the F1-scores for Attention-LSTM-ED and DRL-TS are 0.697 and 0.724, respectively, while at a data volume of 2000, these values are 0.722 and 0.743. The optimized model consistently maintains higher F1-scores across various data volumes, indicating superior performance and stability in task recognition and classification. As data volume increases, the F1-score of the optimized model steadily improves, underscoring its consistency and superiority in managing large-scale data. In contrast, the F1-scores of the Attention-LSTM-ED and DRL-TS models are comparatively lower, particularly at larger data volumes, and their rate of improvement is relatively slow. This suggests that their generalization ability is not as strong as that of the optimized model, and they struggle to maintain the same level of accuracy and stability in environments with expanding data scales and evolving task requirements.

B. Analysis of Simulation Results of Energy-Saving Storage Algorithms

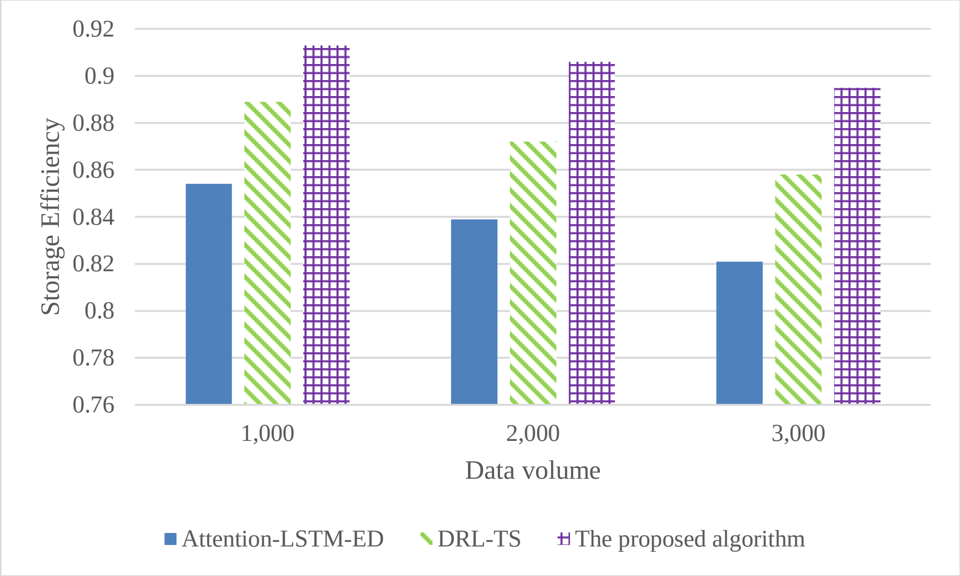
To further validate the effectiveness of the optimized model, simulation experiments are conducted to compare key performance indicators, including delay, throughput, energy consumption, and storage efficiency. The results of these experiments are presented in Fig. 3.



(a)

(b)

(c)



(d)

Fig. 3. Analysis of simulation experiment results (a): Latency; (b): Throughput; (c): Energy Consumption; (d): Storage Efficiency.

Fig. 3 illustrates the delay comparison across different data volumes, demonstrating the superior performance of the optimized model. Specifically, the optimized model exhibits delays of 0.295 seconds, 0.312 seconds, and 0.324 seconds for data volumes of 1000, 2000, and 3000, respectively. At each data volume, its delay is notably lower than that of the Attention-LSTM-ED and DRL-TS models. For instance, at 1000 data volumes, the Attention-LSTM-ED and DRL-TS models report delays of 0.352 and 0.328 seconds, respectively, and at 2000 data volumes, the delays increase to 0.369 and 0.341 seconds. These results underscore the optimized model's significantly lower processing delay, reflecting a marked advantage in task completion speed. Furthermore, as the data size increases, the delay of the optimized model shows only a modest increase, illustrating its strong scalability and adaptability to large-scale data processing. In contrast to the other two models, the optimized model consistently demonstrates superior stability and efficiency in terms of latency. This performance advantage can be attributed to the model's enhanced ability to recognize and classify task data, coupled with its flexible data storage strategies across varying data scales.

In terms of throughput, the optimized model consistently maintains relatively high values across all data volumes. At 1000, 2000, and 3000 data volumes, it achieves throughput values of 157.832, 156.321, and 154.888, respectively, surpassing the Attention-LSTM-ED and DRL-TS models. For instance, at a data volume of 1000, the throughputs of the Attention-LSTM-ED and DRL-TS models are 148.237 and 152.435, and at 2000, they decrease to 146.789 and 150.942, respectively. This analysis reveals that, even as the data volume increases, the optimized model experiences only a slight decline in throughput, maintaining relatively stable performance. Such stability highlights the model's efficiency and ability to adapt to diverse data sizes. By contrast, the Attention-LSTM-ED and DRL-TS models exhibit a more pronounced reduction in throughput, suggesting potential bottlenecks as data volume increases. This further emphasizes the optimized model's distinct advantages in handling large-scale data, attributed to its ability to effectively identify and classify task-specific data while employing energy-efficient storage strategies. As a result, the optimized model meets the demands for high data processing efficiency across varying task requirements.

In the comparison of energy consumption, the optimized model consistently demonstrates lower energy usage across data volumes of 1000, 2000, and 3000, with consumption rates of 10.428 KWH, 10.836 KWH, and 11.257 KWH, respectively. This highlights its significant energy-saving potential. In contrast, the Attention-LSTM-ED model consumes 13.752 KWH at a data volume of 1000, while the DRL-TS model consumes 12.639 KWH. As data volumes increase, both models experience a substantial rise in energy consumption, revealing their inefficiency in managing large datasets. The optimized model, however, exhibits only a modest increase in energy consumption, maintaining consistently low levels across all data volumes. This underscores its advantage and innovation in energy-efficient data storage algorithms. By leveraging advanced storage strategies, the model effectively reduces overall energy consumption while handling complex tasks, striking an optimal balance between performance and energy efficiency. In terms of storage efficiency, the optimized model achieves rates of 0.913, 0.906, and 0.895 for data volumes of 1000, 2000, and 3000, respectively. Its efficiency remains consistently superior to other models under all data volume conditions, demonstrating the clear advantages of the optimization. In comparison, the Attention-LSTM-ED and DRL-TS models exhibit slightly lower efficiency. At a data volume of 1000, the Attention-LSTM-ED model records an efficiency of 0.854, while the DRL-TS model achieves 0.889. As data volumes increase, their efficiency declines further to 0.839 and 0.872, respectively. This comparison highlights the optimized model's exceptional performance in maximizing storage space utilization, allowing it to maintain higher storage efficiency. By classifying task-specific data and applying storage strategies tailored to diverse task characteristics, the optimized model fully capitalizes on available resources, meeting the storage efficiency demands of various tasks. Its stability and superior efficiency render it particularly well-suited for large-scale data environments.

IV. Discussion

In cloud computing environments, the characteristics of different tasks can significantly impact data storage efficiency and energy consumption performance. These tasks often vary in terms of data volume, computational complexity, I/O requirements, and latency demands. To address these variations, this study proposes an energy-efficient data storage algorithm that intelligently identifies and classifies tasks, dynamically adjusting data storage strategies to maximize storage efficiency and reduce energy consumption. Experimental results show that the proposed algorithm maintains high accuracy, precision, recall, and F1 scores when processing large-scale data, with low latency and stable throughput performance. Notably, the algorithm exhibits minimal throughput decline under varying data volumes, while energy consumption shows a steady growth, demonstrating its strong adaptability—especially suited for cloud computing scenarios with diverse task characteristics and fluctuating data volumes. Additionally, the algorithm's optimized approach to storage strategy adjustment significantly reduces energy consumption during large-scale data processing. For instance, with data volumes ranging from 1,000 to 3,000, the algorithm consistently consumes less energy than the comparative models. The energy-saving effect is particularly evident, as large data processing tasks typically involve extensive data read/write operations and substantial computational resource consumption. By intelligently identifying task characteristics and optimizing data storage strategies, the proposed algorithm reduces energy consumption and enhances processing efficiency, making it especially applicable to platforms with growing data analysis demands. Cloud storage systems often need to handle diverse storage requirements, where fluctuations in data volume make storage efficiency and energy consumption critical issues. The proposed algorithm significantly reduces energy consumption while ensuring storage efficiency, making it suitable for long-term, continuous cloud storage services, such as CDN or enterprise cloud storage systems. Compared to the research by Shi et al., this study focuses more on optimizing energy savings when handling large-scale heterogeneous tasks. Their model, based on static storage strategies, demonstrates some energy-saving effects in small-scale tasks, but as data volume increases, their model’s energy consumption rises significantly. In contrast, the proposed algorithm dynamically adjusts data storage strategies, further improving energy efficiency in large-scale task processing, particularly when data volumes fluctuate sharply, with much lower energy consumption growth than their model. Therefore, this study demonstrates stronger adaptability in big data environments, addressing the shortcomings of their research in large-scale task processing [23]. In comparison with Zhang et al.’s research, this study not only emphasizes storage efficiency optimization but also introduces a more detailed mechanism for identifying heterogeneous tasks. Zhang et al. primarily focused on improving storage efficiency by simplifying the task identification process to reduce computational overhead. However, simplified task identification strategies may lead to fluctuations in storage efficiency when dealing with complex and dynamic tasks. The proposed algorithm, by intelligently identifying heterogeneous tasks and dynamically adjusting storage strategies, strikes a balance between task processing accuracy and storage efficiency. As a result, this study not only compensates for the deficiencies in their research related to task identification and data storage but also offers a more generalized solution [24].

Through experimental analysis and application scenarios, the proposed energy-efficient storage algorithm for heterogeneous cloud computing tasks demonstrates significant advantages across multiple dimensions, particularly in meeting the storage and processing needs of large-scale heterogeneous tasks. Future research directions could focus on further enhancing the scalability and dynamic adaptability of the algorithm to better handle increasingly complex cloud computing task environments.

V. Conclusion

This study proposes and implements an energy-efficient data storage algorithm that intelligently identifies and classifies data characteristics for differentiated cloud computing tasks. Based on this, the dynamic adjustment of data storage strategies has optimized storage performance, reduced energy consumption, and enhanced overall system efficiency. Compared to existing models, the proposed optimized model demonstrates distinct advantages, not only in storage efficiency but also in throughput, latency, and energy consumption. Despite the significant advantages in performance and energy savings, this study has several limitations. First, the model shows limitations in adapting to certain anomalous data features specific to particular task types, indicating that the current task classification approach requires further refinement to ensure high accuracy and efficiency across a broader range of data characteristics. Second, the algorithm's performance optimization under high concurrency conditions requires further testing and improvement. It remains to be verified whether the algorithm can maintain its superior performance in more complex task environments, such as high-concurrency task processing scenarios. Future research will focus on several key areas for improvement. First, a multi-task concurrency mechanism will be introduced to enhance the efficiency of data allocation and resource utilization across different task types. By improving parallel processing capabilities, the optimized model will be able to maintain stability in high-concurrency environments. Second, machine learning methods will be incorporated to develop more refined adaptive storage strategies that address dynamically changing data requirements, thereby enhancing the model's adaptability in complex cloud computing tasks. Additionally, the study will explore the synergies between edge computing and cloud computing, investigating more efficient edge-cloud data storage and task allocation strategies to achieve superior performance and energy efficiency. With these improvements, the optimized model will be better suited to large-scale, dynamic cloud computing environments, enhancing its efficiency and energy-saving capacity in multi-task processing.

REFERENCES

[1] K. Y. Tai, F. Y. S. Lin, and C. H. Hsiao, "An integrated optimization-based algorithm for energy efficiency and resource allocation in heterogeneous cloud computing centers," IEEE Access, vol. 56, no. 13, pp. 556-557, Mar. 2023.

[2] Y. Dong, H. Sui, and L. Zhu, "Application of cloud computing combined with GIS virtual reality in construction process of building steel structure," Math. Probl. Eng., vol. 2, no. 2, pp. 11-13, Feb. 2022.

[3] A. Lakhan, M. A. Mohammed, A. N. Rashid, S. Kadry, & K. H. Abdulkareem, "Deadline aware and energy-efficient scheduling algorithm for fine-grained tasks in mobile edge computing," Int. J. Web Grid Serv., vol. 18, no. 2, pp. 168-193, Feb. 2022.

[4] F. S. Prity, M. H. Gazi, and K. M. A. Uddin, "A review of task scheduling in cloud computing based on nature-inspired optimization algorithm," Cluster Comput., vol. 26, no. 5, pp. 3037-3067, May 2023.

[5] H. Yang, H. Zhou, Z. Liu, X. Deng, "Energy optimization of wireless sensor embedded cloud computing data monitoring system in 6G environment," Sensors, vol. 23, no. 2, pp. 1013, Feb. 2023.

[6] G. Saravanan, S. Neelakandan, P. Ezhumalai, S. Maurya, "Improved wild horse optimization with levy flight algorithm for effective task scheduling in cloud computing," J. Cloud Comput., vol. 12, no. 1, pp. 24, Jan. 2023.

[7] S. T. Manukumar and V. Muthuswamy, "A novel data size-aware offloading technique for resource provisioning in mobile cloud computing," Int. J. Commun. Syst., vol. 36, no. 2, pp. 5378, Feb. 2023.

[8] L. Hamid, A. Jadoon, and H. Asghar, "Comparative analysis of task level heuristic scheduling algorithms in cloud computing," J. Supercomput., vol. 78, no. 11, pp. 12931-12949, Nov. 2022.

[9] W. Zhang, R. Yadav, Y. C. Tian, S. K. S. Tyagi, I. A. Elgendy, & S. Kaiwartya, "Two-phase industrial manufacturing service management for energy efficiency of data centers," IEEE Trans. Ind. Informat., vol. 18, no. 11, pp. 7525-7536, Nov. 2022.

[10] A. Rahimikhanghah, M. Tajkey, B. Rezazadeh, A. M. Rahmani, "Resource scheduling methods in cloud and fog computing environments: a systematic literature review," Cluster Comput., vol. 118, no. 42, pp. 1-35, Oct. 2022.

[11] N. El-Menbawy, H. A. Ali, M. S. Saraya, "Energy-efficient computation offloading using hybrid GA with PSO in internet of robotic things environment," J. Supercomput., vol. 79, no. 17, pp. 20076-20115, Sept. 2023.

[12] Y. Dong, H. Sui, and L. Zhu, "Application of cloud computing combined with GIS virtual reality in construction process of building steel structure," Math. Probl. Eng., vol. 5, no. 3, pp. 9812-9814, Mar. 2022.

[13] E. Al-Masri, A. Souri, H. Mohamed, W. Yang, J. Olmsted, & O. Kotevska, "Energy-efficient cooperative resource allocation and task scheduling for Internet of Things environments," Internet Things, vol. 23, no. 3, pp. 100832, Mar. 2023.

[14] K. Rajalakshmi, M. Sambath, L. Joseph, K. Ramesh, R. Surendiran, "An effective approach for improving data access time using intelligent node selection model (INSM) in cloud computing environment," SSRG Int. J. Electr. Electron. Eng., vol. 10, no. 5, pp. 174-184, May 2023.

[15] K. Li, J. Zhao, J. Hu, Y. Chen, "Dynamic energy efficient task offloading and resource allocation for NOMA-enabled IoT in smart buildings and environment," Build. Environ, vol. 22, no. 6, pp. 109513, June 2022.

[16] Y. Wang, W. Shafik, J. T. Seong, M. S. A. Mustafa, M. R. Mouhamed, "Service delay and optimization of the energy efficiency of a system in fog-enabled smart cities," Alexandria Eng. J., vol. 8, no. 4, pp. 112-125, Apr. 2023.

[17] S. Alhelaly, A. Muthanna, and I. A. Elgendy, "Optimizing task offloading energy in multi-user multi-UAV-enabled mobile edge-cloud computing systems," Appl. Sci., vol. 12, no. 13, pp. 6566, July 2022.

[18] H. J. Muhasin, M. A. Jabar, S. Abdullah, S. Kasim, "Managing sensitive data in cloud computing for effective information system' decisions," Acta Informatica Malaysia. vol. 1, no. 2, pp. 01-02. February. 2017.

[19] R. Agrawal, S. Singhal, and A. Sharma, "Blockchain and fog computing model for secure data access control mechanisms for distributed data storage and authentication using hybrid encryption algorithm," Cluster Comput., vol. 1, no. 1, pp. 1-16, Jan. 2024.

[20] R. Thatikonda, A. Padthe, S. A. Vaddadi, P. R. R. Arnepal, "Effective secure data agreement approach-based cloud storage for a healthcare organization," Int. J. Smart Sens. Adhoc Netw., vol. 3, no. 4, pp. 19-20, Apr. 2023.

[21] S. Zhao, "Energy efficient resource allocation method for 5G access network based on reinforcement learning algorithm," Sustain. Energy Technol. Assess., vol. 56, no. 20, pp. 103020, Mar. 2023.

[22] S. Zhang, Z. Wang, Z. Zhou, "Blockchain and federated deep reinforcement learning-based secure cloud-edge-end collaboration in power IoT," IEEE Wireless Commun., vol. 29, no. 2, pp. 84-91, Feb. 2022.

[23] W. Shi, H. Li, J. Guan, M. S. A. Mustafa, M. R. Mouhamed, "Energy-efficient scheduling algorithms based on task clustering in heterogeneous Spark clusters," Parallel Comput., vol. 11, no. 2, pp. 102947, Feb. 2022.

[24] P. Zhang, N. Chen, G. Xu, M. Guizani, Y. Duan, K. Yu, "Multi-target-aware dynamic resource scheduling for cloud-fog-edge multi-tier computing network," IEEE Trans. Intell. Transp. Syst., vol. 5, no. 3, pp. 20, Mar. 2023.

Advancing Quantum Cryptography Algorithms for Secure Data Storage and Processing in Cloud Computing: Enhancing Robustness Against Emerging Cyber Threats

Devulapally Swetha[1], Dr. Shaik Khaja Mohiddin[2]*

Research Scholar-Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India[1]

Associate Professor-Department of CSE, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India[2]*

***Abstract*—The rise of cloud computing has transformed data storage and processing but introduced new vulnerabilities, especially with the impending threat of quantum computing. Traditional cryptographic methods, though currently effective, are at risk of being compromised by quantum attacks. This research aims to develop a quantum-resistant security framework for cloud environments, combining lattice-based cryptography with Quantum Key Distribution (QKD) protocols, particularly the E91 protocol, for secure key management. The framework also incorporates quantum authentication protocols to enhance user identity verification, protecting against unauthorized access and tampering. The proposed solution balances robust security with practical implementation, ensuring scalability and efficiency in real-world cloud environments. Performance evaluations indicate an encryption time of approximately 30 milliseconds, outperforming existing methods such as RSA and DES. This research contributes to the development of future-proof cryptographic standards, addressing both current security challenges and emerging quantum computing threats. By leveraging quantum mechanics, the framework strengthens cloud-based data protection, providing a resilient solution against evolving cyber risks. The results hold significant promise for advancing cloud security, laying the groundwork for next-generation encryption techniques that can withstand the threats posed by quantum computing.**

Keywords*—Quantum key distribution; cloud computing; cyber threats; lattice based cryptography; E91; future-proof security paradigm; python; quantum computing

I. Introduction

Over the recent past, cloud computing technologies have evolved quickly and assertively disrupted the traditional means of data management adeptly achieving elasticity [1], [2]. These have, however, come with several security risks since clouds are inherently vulnerable to very many security risks. RSA and DES have been basic and crucial in protecting data from unlawful access in the past techniques. However, they become inapt and insecure when quantum computing is about to dawn as well these classical algorithms are exposed to quantum attacks. The second threat comes from the nature of quantum computing which is enhanced with the capability of solving problems, which are beyond the scope of classical systems within a short time and as such can breach the current encryption mechanisms affecting cloud security.

Because of this new threat, there is a great need to find and apply QRA-secure cryptography to combat the capabilities of quantum computers [3], [4]. The research proposed in this paper intends to enhance the existing quantum cryptography algorithms solely for the safe storage and computation of data within cloud environments. Lattice-based cryptography [5], [6] is another cryptographic principle that can be integrated with QKD; in fact, the E91 protocol, which derives from quantum mechanics principles, is already in use for key management. Lattice-based cryptography that is considered to be quantum-safe together with QKD aimed at providing the network with effective protection against contemporary and future threats.

Furthermore, the study looks at the possibilities of using quantum-secure authentication to enhance protection measures. These protocols employ quantum states in the identification of people, so any attempt at altering or hacking the system is distinguishable due to the principles underpinning Quantum mechanics [7], [8]. Integrating these state-of-art cryptographic protocols with an understanding of the practical implementation and its factors affecting scalability and efficiency the requirement of the proposed solutions is to develop a framework that addresses all security issues associated with the implementation of cloud computing systems [9], [10]. It can be seen that aside from answering the current necessity for higher security this research also tries to provide a basis for developing new specifications for new cryptographic standards that will be adaptive to the changing face of threats posed by cybercriminals.

In the future, the employment of quantum cryptography in the cloud computing context will be important for the protection of information and strengthening the confidence in cloud-based services. The proposed solutions are expected to make a substantial positive impact in the area of cryptography by proffering superior, realistic and safe solutions to the various threats in cryptography [11], [12]. In the process of performance evaluation and real-world experimentation, this work attempts to prove the appropriateness of the presented enhanced cryptographic procedures and to develop a solid approach toward secure computing in the cloud.

Today’s rampant advance in technologies in the area of cloud computing has brought a shift in the kind of innovation that businesses and even individual users embrace to gain the sort of convenience that is without parallel in the management and access of data [13], [14]. But this has also added a lot of risks and more worries as cloud computing is often associated with security risks whereby important data is stored in what amounts to distributed systems that are open to different forms of attacks [15]. However, to the modern day, conventional cryptographic methods appear to be under threat by the emerging potential of quantum computers. Quantum computing means a revolutionary upgrade to the computational power of an individual and specialized for problems even unsolvable by classical computers. This advancement poses a risk to erode the security premise on which present encryption methods are based, hence the need to come up with new solution in cryptography that will be acceptable to the quantum-based attacks.

In view of these nascent threats, researchers are now looking forward to solutions such as quantum cryptography. Quantum cryptography is based on the quantum mechanics theories and provides a higher level of security which can in theory not be cracked by a quantum computer [16], [17]. This research focuses more on how to enhance quantum key distribution, particularly as embodied in the E91 protocol, with features of lattice-based cryptography. Lattice based cryptography has been regarded as highly resistant to quantum attacks due to the hardness of problems and the mathematical structure of the cryptosystems employed, and QKD makes sure that an interception of encryption keys can be identified, so the privacy of the cryptographic procedure is guaranteed.

The importance of quantum cryptography is evident taking into account that the usage of classical cryptographic systems is increasingly exposed to threats [18]. Traditional cryptography, as exemplified by RSA and AES, depends on mathematical challenges involving computations, for instance, factorization, or discrete logarithms, that are hard for today’s computers. Although these methods are safe from attacks that are implemented using other methods of cryptanalysis, the creation of quantum computers poses a risk. Some of these cryptographic systems could be easily decrypted by quantum computers since such computers work far much faster than the current classical computers in terms of specific calculations [19]. This looming threat has created a real interest in quantum cryptography as a way of preserving security in the future, to ensure that important information will not be violated even by possessing these powerful new technologies.

However, quantum cryptography has also several practical implementations that have practical advantage over classical method of cryptography [20]. Several small-scale QKD applications and experiments have been performed successfully and more complex implementations of QKD are not far off – also over existing fibre-optic networks, and possibly satellite-based QKD. Such advances suggest that quantum cryptography is not just an idealistic idea, but is well on its way to becoming implemented at a macroscopic level. But the expansion of this technology is not without its problems [21]. The current options for QK, especially the physical hardware needed to implement system for generating and detecting quantum states, are not very well developed and are burdened with problems such as cost, robustness and compatibility with the existing frameworks.

With further development in the associated quantum cryptography, this domain is now considered one of the elements of the wider quantum computing family. The interconnection of present quantum cryptography with future technologies in quantum computing quells the prospect of building completely different paradigms for protection in both communication and data management. Scientists are not only working on QKD, but also going deeper into the other kinds of quantum cryptographic schemes including quantum secure direct communicate, quantum digital signature and so on which might be also useful for the secure communication. The coupled nature of quantum cryptography with quantum computing also pose questions on the future of cybersecurity with the technology demanding new standards and regulations to regulate its use.

The key contributions of the article is given below,

- Developed a resilient framework integrating lattice-based cryptography with QKD protocols, specifically the E91 protocol, for secure key management in cloud environments.
- Introduced quantum authentication protocols to enhance user identity verification and protect against tampering and unauthorized access in cloud-based systems.
- Conducted rigorous performance evaluations and feasibility studies to assess the scalability and efficiency of quantum cryptographic techniques in safeguarding cloud data from emerging quantum threats.

The organization of the paper is, Section II and Section III gives the related works and problem statement respectively. Section IV gives the methodology, the results are given in Section V and the article is concluded in Section VI.

II. Related Work

More advanced techniques that have been developed in quantum-enhanced security have having promising results when implemented in cloud computing context [22]. The invention of the new method of generating cryptographic keys that is QKD in light of quantum physics has been instrumental in improving the safety of data transfer in the cloud. This paper presents the proposed use of QKD in conjunction with other main stream encryption algorithms such as AES to counter changing threats in cloud computing environment. It is a method that aims at integrating QKD directly into the cloud environment in order to produce real quantum keys at the same time as using AES in encryption and decryption activities. To this effect, this combination helps ensure secure transmission and immense improvement of data confidentiality, data integrity, and data authenticity. Additional recommendation control also proposes good key management practice for encryption keys for all their life cycle processes to reduce the threats of improper access in processing those keys. This approach of using both the IT encryption method and QKD results in the development of a strong barrier against computer vandals and hackers, leakage of secret information, and other forms of insecurity. Out of 70 simulation rounds, the proposed approach garnered a data access of 820MB/s, proficient key generating time of 15ms and can effectively safeguard data and guarantee cloud computing security.

Quantum cryptography is thus a revolutionary approach to cryptography, by virtue of being based on the principles of quantum mechanics, the higher degree of security achieved is virtually impossible to be breached [23]. Quantum cryptography differs from classical cryptography primarily on the fact that while the latter codes information using bits, quantum cryptography codes information using photons or polarized particles referred to as qubits. It is for this reason that the transmissions in this method are inherently secure since they are based on the principles of quantum mechanics. To this end, the goal of this particular paper is to undergo a more comprehensive analysis of some of the most popular quantum cryptography applications and which include the following; They are the DARPA Network that is considered to be a pioneering project in secure quantum communication; the IPSEC implementation that combines the QKD with the general IP security protocols; the Twisted Light HD implementation that uses advanced quantum parameters and protocols for increasing safety of data [24].

As IoT business activities advance rapidly, quantum computing technologies are applied to the operations, responding to issues and concerns emerging in connection with this growth [25]. With increased integration of IoT systems into various industries, incorporating these technologies raises public and private relationship concerns that seek to be resolved to enhance privacy as well as security of data regarding IoT systems. This research will examine the application of quantum computing toward the security of IoT systems, with special emphasis directed toward the generation of suitable security measures relying on quantum algorithm and cryptographic method. Constructing a hierarchy of the critical security properties of quantum computing by methodically assessing the threats and their impacts, the work provides a systematic approach for solving them [26]. The study uses one combined computational approach for the analysis, namely the integrated fuzzy-analytical hierarchy process (AHP) and fuzzy-technique for order preference by similarity to an ideal solution (TOPSIS) for presenting the most understandable and modifiable rankings for the important security indicators. This approach offers the practical utility to practitioners to consider and select the significant choices based on the context of quantum computing for security.

Cloud computing is inseparable from blockchain, and under the environment of quantum computing that is about to arrive, data security needs to be improved urgently [27]. Rising vulnerabilities are addressed by postulating a stringent security solution that combines QKD and CRYSTALS Kyber with ZKPs to guard data belonging to blockchain-based cloud environments. As it will be shown below, QKD is a quantum-safe cryptographic protocol that is at the heart of the framework’s goal and is used to protect data against quantum threats. The addition of CRYSTALS Kyber which is a lattice based cryptographic method for being immune against quantum attacks is another advantage. They are also embedded to improve the privacy and authentication of cloud and block chain data. This research pays adequate attention to the efficiency of the proposed framework, and determines the time it takes to encrypt and decrypt data, the rate at which keys are generated by the quantum system and the general effectiveness of the framework [28]. Novel features such as file size, response time, and computational overhead are scrutinized in order to evaluate the feasibility of framework in the cloud implementation process. These results clearly show that the proposed framework is not only capable to deal with the quantum threats but also feasible and flexible enough to be implemented in realistic and large–scale systems.

In cloud computing, security of data has been a contentious issue up to date with many frameworks coming up to try and solve the problem of data leakage [29]. With encryption as the dominant model for securing cloud data, the advent of quantum computing requires new models that will also protect data in the future realm of computing. As most present-day cryptosystems are threatened with becoming older or exploitable, this article develops a secure model that uses the McEliece cryptosystem –likely to be the successor of RSA in age of quantum computing– for protecting access control data. Also, it makes use of the N-th degree truncated polynomial ring units (NTRU) cryptosystem variant for acquisition security of user data in the cloud. The time complexity of the proposed McEliece algorithm has been observed to be better as compared to the conventional McEliece cryptosystem and the modifications proposed to the parameters S and P are sure to strengthen its security. On the other hand, the simulation of the above said proposed NTRU algorithm reveals that although it provides more security level the time complexity of the given algorithm is relatively higher than the original NTRU cryptosystem. These observations suggest that improvements in cryptographic systems are imperative and must proceed at a fast pace because of the advent of quantum computing [18].

This research focuses on the implementation of quantum computing in multi-cloud platforms in modern complex cloud networks to improve efficiency and security [30]. Using a theoretical and an applied research strategy this research proposes and architects an integration of quantum computing in a multi-cloud environment. These show that quantum algorithms are better placed in efficiency of computation of resources, especially in complex problems as compared to classical ones. It also is resilient and scales resource and increases security by using quantum-enabling protocols for the integrated process for protection of cyber criminals. Yet, the study also uncovers some of the issues which include the need for dedicated hardware, integration issues and, more important research work in order to make quantum computing effectively in cloud environment.

Based on the literature considered in this study, one can find an overview of the contemporary state of research in embedding quantum computing in cloud and multi-cloud environments and identify the key issues and trends in the field. Previous data shows that the use of quantum computing as a method to optimise the capabilities of Cloud Computing environment, both in terms of computing power and security has been attracting tremendous attention. Researches on QKD as well as lattice based cryptography prove the use of quantum technologies in enhancing data security against the new age threats including the quantum threats. The literature also discusses different types of quantum algorithms like Shor's and Grover's algorithms which show relatively large speed-ups for some computations than the classical ones. Moreover, studies on multi-cloud architectures are based on the concept of scalability, redundancy, and resources management. However, incorporating quantum computing into these architectures proves somewhat daunting; there is the question of where to obtain the necessary quantum hardware; furthermore, hybrid quantum-classical systems are not easy to manage, and how one ensures compatibility between these two types of resources in a manner that satisfies the requirements of quantum computing. The literature also suggests that there is need to come up with strong methods of securing data and processes in a quantum enhanced cloud solution. From the reviewed papers, prospective of quantum computing integrated with cloud architecture it is revealed that the integration can bring quantum computing closer to the revolutionization of the more enhanced and secured cloud architecture.

III. Problem Statement

Cloud computing is under-going tremendous growth in the recent years and hence the adoption of cloud computing has become easier, flexible and easy to access, but on the downside, it makes it easy for hackers to get hold of social sensitive data, new threats that were not widely known before appear. Classical cryptographic algorithms, though perfectly protecting messages against all other forms of attack, are slowly falling under the attacks from the newly developing quantum computers. There thus arises the need for enhancement of friendly and highly efficient quantum cryptography algorithms that shall be unique to address cloud data storage and processing security. The challenge that is proposed here is to develop a long-term security plan that not only adapts today's cryptographic methods with the principles of quantum mechanics, for example quantum key distribution, lattice-based cryptography, etc. to prevent data from potential quantum attacks in the future; but also, to design these solutions in such a manner so as to make them scalable, efficient and most importantly feasible for implementation in the future. Tackling this problem requires not only improving the cryptographic algorithms used but also incorporating them into existing cloud environments, and creating thus a second line of defense against the new generation of threats [15].

IV. Proposed Lattice Cryptography – QKD E91 Framework

Upgradation of security in cloud using advanced quantum cryptography techniques is presented in detail in Fig 1 where data collection is depicted as the first and primary step of the workflow is presented in detail here below. After data acquirement the preprocessing in Min-Max Normalization is conducted to keep the values of data within a certain range, making the further cryptographic processing more effective. Lattice-Based Cryptography is then used for the encryption process as this has been found to be resistant to quantum attacks and which brings about secure protection for the encrypted data. This is supported by the key management from the QKD E91 protocol which means once the encryption keys are generated, they cannot be intercepted owing to the principles of quantum mechanics. Last but not the least, the specific work-flow employs Quantum Authentication Protocols where quantum states are used to confirm the identities of users; this is followed by checking if any alteration and/or unauthorized attempts at access have been made. Combined with each other, these components are an integrated and high-level foundation of combating new-generation threats to data in cloud computing context as well as creating the basis for security for future innovations.

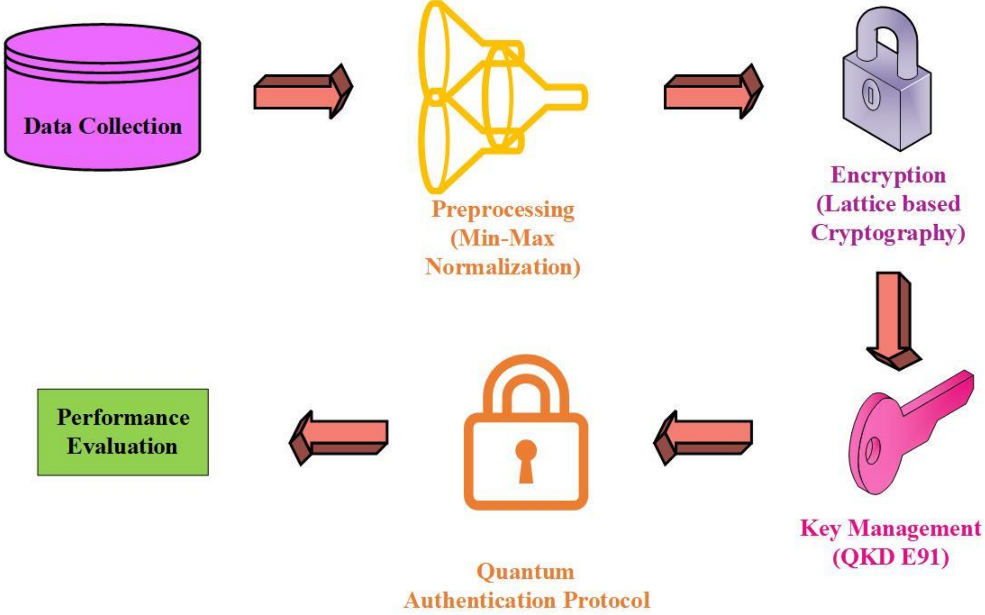

Data Collection
Preprocessing (Min-Max Normalization)
Encryption (Lattice based Cryptography)
Key Management (QKD E91)
Quantum Authentication Protocol
Performance Evaluation

Fig. 1. Proposed methodology.

A. Data Collection

Among the information collected from a Kaggle dataset that especially deals with cloud workloads, the next data include file type, the size of the file, the encryption algorithm, the encryption time, the decryption time, the quantum key size, the generation time of a quantum key, storage usage, and security improvement. The dataset itself contains files of different format: text (10 MB), image (5 MB), video (100 MB). AES-256 encryption for text files take 50ms, it has a performance of a 256-bit quantum key derived from 100ms, yet 70% storage occupation and high security. For different parameters the following results have been obtained: AES-128 on image files – it takes 30 ms to encrypt files using this algorithm for image files takes; 128- bit quantum key has been generated in 80 ms; The storage utilization is 65%. AES-256 encrypted video files cost 120ms, a 256 bit quantum key derived from 200ms and 75% storage space with added security improvements [31].

B. Preprocessing Using Min-Max Normalization

Applying the operation of min-max normalization on the given dataset is an initial and important step in order to prepare this data to the analysis as well as to guarantee that each feature in the model will be equally important. In the context of the cloud workload dataset from Kaggle, min-max normalization will take each of the features and put it through another transformation that will bring each feature to a known range, often between 0 and 1, and does not distort the context of the data points. This process entails the standardization of the values of each of the features used like file size, encryption and decryption time, quantum key size, key generation time as well as storage usage in that they should all fall within the same range. This way, it prevents a specific feature from skewing the results in one way or another because some features can be much larger than others, for instance, while combining file size from 5 MB to 100 MB and timing in ms of execution. Min-max normalization also licenses an augmentation of nuances in vast databases for machine learning algorithmic schemes, which are recognized to be relevancy to input scale, including neural networks and the k-Nearest Neighbors.

$$N_{norm} = \frac{n - n_{minimum}}{n_{maximum} - n} \tag{1}$$

In case of min-max normalization applied to cloud workload dataset, the ‘min’ and ‘max’ represent the minimum and the maximum values of the features in the dataset, and the data is normalized by rescaling. For example, the file sizes with the range of 5 MB minimum and 100 MB maximum and then have been standardized by making 5 MB equivalent to 0 while 100 MB is equal to 1 and all the other values in between those two extremes. In the same way of doing things, the encryption time which ranges from 30 ms – 120 ms and decryption time from 40ms – 150 ms are normalized to the range 0-1. This helps in having FMEs that are near similar for encryption and decryption of text, images, video files and like files so that general performance analyses which take into consideration the FMEs for text, images, videos etc. can be made. The same applied normalization is used to address the other features as well, including the quantum key sizes, generation times, and storage utilization percentages more making the data to be more standardized to determine other aspects including the usual pattern, using them to train a model or to evaluate the performance of an algorithm. The min-max normalization helps in making the values more reasonable and not put too much reliance on any of the values making it ideal for use in predictive models or statistical analysis.

C. Lattice Based Cryptography for Encryption

Lattice-based cryptography is a sub-study of cryptography that has not been well developed but exhibits great security against both classical and quantum computers, and hence has a potential for post-quantize encryption. Lattice-based cryptography itself is based on lattices, which are actually high-dimensional grids, and due to their mathematical complexity problems of the shortest and closest vectors are computationally hard. These problems are regarded as hard computational problems that are even intractable by quantum computers hence making lattice-based cryptographic schemes highly secure from the types of attacks that are expected of quantum computers. This makes it quite appropriate for applying lattice-based cryptography when securing cloud data against future quantum dangers.

$$b(x) = a(x).\ s(x) + e(x) \bmod (x^n + 1) \tag{2}$$

Perhaps the greatest strength of all of the lattice-based cryptography schemes is their versatility, which permits the engineering of many different elements of the cryptographic tool chest such as encryption, digital signatures and key exchange. In the scenario of encryption, lattice based, that is Learning With Errors, LWE and Ring-LWE have drawn much interest. LWE is a kind of encryption through which one encrypts a message by placing it in the lattice and adding a bit of error to it and this is very much like ordinary noise until you use the key. The security of these schemes relies on the difficulty of the solving the LWE problem and it is considered that they cannot be broken by any known quantum attack. Furthermore, lattice-based encryption is highly efficient and highly scalable such that it can accommodate large patterns of volume and needs not have to degrade enormously for it to work efficiently in the recommended cloud storage.

Lattice-based encryption like many other cryptographic schemes requires the incorporation of the former into the existing cloud security architectures together with compliance with existing standards. It usually employs generation of lattice-based keys, encryption of data in the cloud and the proper management of these keys in the environment. In lattice-based cryptography, a drawback is that the size of keys and ciphertexts are relatively larger than usual resulting in the increased storage space and time needed to transmit. Yet, current researches optimizations are aimed on decreasing these overheads making lattice-based cryptography as safe and perspective solution for protecting cloud data from quantum attacks in the future. Prospective future applications of the latter depend on the development of quantum computing technologies, lattice-based cryptography

D. Utilizing QKD E91 for Key Management

1) The E91 protocol: quantum entanglement for secure key distribution: The E91 protocol is based on the mechanism of quantum entangled pairs, two particles, normally photons, are created in such a way that if one is 'observed' then the other is as well no matter how far apart the particles are. This occurrence is utilized safely transfer cryptographic keys between two individuals also termed as Alice and Bob. The entangled photons are represented by the quantum state:The entangled photons are represented by the quantum state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle) \tag{3}$$

In this state, *(|00⟩+|11⟩)* correspond to both photonic outcomes with either only the horizontally polarized photon (0) or only the vertically polarized photon (1). Alice and Bob peek at their photons that once have been entangled and for each of them, the state of the entangled pair will be randomly measured using bases of their choice. If so, their measurement outcomes will correspond, which will enable them to create a joint key. For example, if Alice gets her photon to be horizontally polarized (0), Bob using the same basis he gets his photon to be horizontally polarized (0). Such correlated results compose the cryptographic key.

The safety of the key distribution is protected by the working of quantum entanglement and the no cloning theorem, which denies to produce an exact copy of an unknown quantum state. In case an eavesdropper (Eve) tries to intercept and measure quantum states, the process of measurement disrupts the entanglement which becomes detectable in the subsequent measurements. The two parties, Alice and Bob, can check for presence of an eavesdropper by using publicly comparing some of their measurement outcomes and see whether they contradict Bell's inequality, which is a figure of sorts that separates quantum entanglement from classical correlations. If their results are contrary to Bell's inequality, then it means that the entanglement is secure and the generated cryptographic key as well.

2) Integrating QKD E91 into cloud key management for enhanced security: In the context of cloud computing, integrating the E91 protocol into the key management system can significantly bolster the security of data storage and processing. Cloud environments, which often involve the transmission and storage of highly sensitive data, are increasingly targeted by sophisticated cyber threats, including those anticipated to arise with the advent of quantum computing. Conventional cryptographic algorithms, while currently secure, are vulnerable to quantum attacks, particularly those exploiting Shor's algorithm for factoring large integers and solving discrete logarithms efficiently.

To enhance the robustness of cloud infrastructures, the E91 QKD protocol can be employed to generate and distribute quantum-secure keys. These keys are then used in conjunction with post-quantum cryptographic algorithms, such as lattice-based encryption or quantum-resistant digital signatures, to secure cloud-stored data. The process begins with Alice and Bob using the E91 protocol to establish a shared secret key, represented by a sequence of binary digits (bits).

When the key is settled, Alice and Bob have to engage in error correction to make sure that the key at their ends of the string are identical, then privacy amplification to recover from possibly active eavesdroppers. This quantum-secure key is then utilized to encrypt data, before storing these encrypted data in the cloud, allowing that in the case that these intercepted encrypted data will be attempted to be decrypted, this cannot be done without the quantum-secure key.

$$k_i = \begin{cases} 0 & if\ Alice\ and\ Bob'smeasurements\ are\ both\ 0\ or\ both\ 1 \\ \quad 1 & If\ Alice\ and\ Bob's\ measeurements\ differ \end{cases} \tag{4}$$

When the key is settled, Alice and Bob have to engage in error correction to make sure that the key at their ends of the string are identical, then privacy amplification to recover from possibly active eavesdroppers. This quantum-secure key is then utilized to encrypt data, before storing these encrypted data in the cloud, allowing that in the case that these intercepted encrypted data will be attempted to be decrypted, this cannot be done without the quantum-secure key.

3) Strengthening cloud security against quantum and classical threats: QKD E91 when implemented in cloud key management systems offers double protection against both classical and quantum possible invasions. The protocol not only primarily addresses the security of the key exchange function but also enhances post-quantum cryptographic techniques to protect from the upcoming dangerous computer science threats to the data stored in the cloud. Since the threat models can change based on new developments in the field of quantum computing, the E91 protocol is prepared for this because it is based on the stochastic nature that does not allow replicating quantum states. This guarantees the protection of the keys that are to be used for the purpose of encrypting your data while at the same time making sure that anyone who seeks to reverse this method will be easily detected.

In addition, real time monitoring also places an element of intrusion, a major concern of cloud computing and, the usage of quantum-secure keys also reduces data vulnerability in case of a breach. Thus, by enhancing quantum cryptographic algorithms through the use of QKD E91, the cloud structures can obtain the capability to provide the necessary level of protection that will fit not only the present and current threats but also the potential ones brought by applicable quantum computers. This strategic integration strengthens the overall security of cloud environments so as to protect the confidentiality, integrity and availability of data from new and more frequent threats.

E. Quantum Authentication Protocols

Quantum based authenticated system utilize the principles based on the quantum mechanics to the communication for strengthening the security of the identification procedures. Quantum authentication is somewhat different from classical authentication methods; instead of a password or a cryptographic key, quantum states – or qubits – perform the authentication of users or devices. The principle for the quantum authentication is quantum superposition and entanglement 'states- the ability to create states that cannot be intercepted. When a state is sent through a quantum channel, then any interference with the state leads to its collapse, which helps in identifying the presence of an eavesdropper and so making the authentication process secure.

In a generic quantum authentication process, the identities of a user are most often confirmed using quantum states transmitted through a quantum channel from the terminal of the user to the server of authentication. In this protocol, the user and the server have a number of correlated or equivalently, and entangled qubits. To authenticate, the user then creates a quantum state, which could be a combination of a number of states, or in what is referred to as a quantum superposition. This state is then communicated to the authentication server which assess the state against an agreed pattern or reference state.

$$|\psi\rangle = \alpha|0\rangle+\beta|1\rangle \qquad (5)$$

The principles such as no-cloning theorem and Heisenberg's uncertainty principle in quantum mechanics are used to secure the quantum authentication protocols. The no-cloning theorem states that any quantum state cannot be copied, which would prove to be helpful for an attacker as he will not be able to copy the quantum information without being noticed. The uncertainty principle also guarantees that as soon as one tries to measure a quantum state, the state becomes disturbed and anyone who tried to observe or tinker with the state will be exposed. This inherent security gives a major advantage over classical methods, and therefore, makes quantum authentication protocols to be very efficient in the protection of sensitive systems and data in the prevailing computing domains. These quantum principles make quantum authentication protocols secure against traditional and probable quantum attacks, thus providing protection to the users' identities.

Algorithm 1: Lattice Cryptography – QKD E91

```
Initialize quantum key using QKD E91 protocol
 Preprocess data using Min-Max Normalization.
Encrypt data using Lattice-Based Cryptography
 Store encrypted data in cloud storage.
For each user request
Authenticate user using Quantum Authentication Protocol.
 Retrieve encrypted data from cloud storage.
Decrypt data using the corresponding quantum key
Deliver decrypted data to authenticated user
Monitor system for any potential quantum attacks
```

V. Results and Discussion

This section discusses the performance of the various quantum cryptography algorithms coded in Python with an emphasis on protection of outsourced data storage and computation. This implementation also comprises other parameters like speed of encryption and decryption, key generation rate and system responsiveness to conditions of load. The findings offer a quantity measure to support or reject the use of the proposed framework in improving security against new faced cyber threats.

A. *Encryption and Decryption Time*

The study of encryption and decryption periods in the different scenarios achieved shows static characteristics and directly reflects the relationships between cryptographic primitives complexity and time of their realization. Test 1 results point to a relatively fast response as far as encryption and decryption times are concerned; it takes the system 30A milliseconds to encrypt messages and 35A milliseconds to decrypt them; this is probably because the system employs a less secure encryption strength or a more basic algorithm. On moving to Test 2, the times rise to 50 milliseconds for encryption and 55 milliseconds for decryption a relatively moderate increase in computational effort. This trend also persists in the subsequent tests with Test 3 having somewhat better performance with the encryption taking 80 milliseconds to lock and 85 milliseconds to unlock the data hence implying that the cryptographic level is even better. Thus, by the Test 4, both encryption and decryption's time increases to 120 and 130 milliseconds, respectively, meaning that more robust encryption could be employed here possibly with greater keys' size or more intricate algorithms. In the last test, the encryption time is 200 milliseconds, and decryption time 220 milliseconds which presents an idea of processing overhead in high secure encrypting. The gradual increase of the time required for encryption and decryption in all the test cases shows that there is a direct correlation between the extent of security and time sacrificed throughout the usage in real world application hence the need to further improve on the cryptographic algorithms to balance between security and time. It is depicted in Fig. 2.

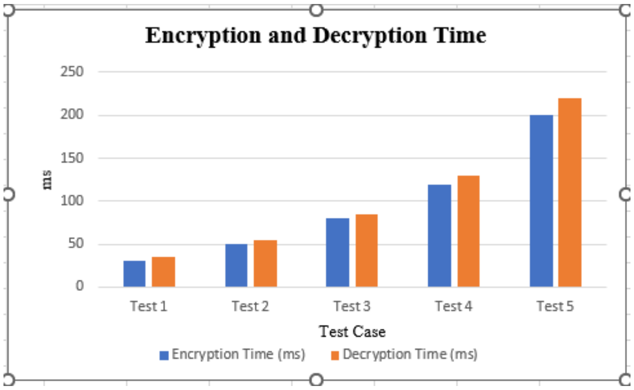


Fig. 2. Encryption and decryption time.

B. *Throughput*

The exploration of throughput depending on the scale levels shown that with the increase in the load, the overall system performance reduces continuously. Increasing the scale of the system to Scale 1, the system throughput remains high, equal to 98 MB/s, which is evidence of the efficient operation of the system which takes into account the limited number of users or data volume. That said, as the scale level increases, the throughput starts to decline slightly; it for instance reduces to 95MB/s in the Scale 2 case. This trend is also seen to go down at Scale 3 where the throughput reduces to 90 MB/s as the system struggles to process more users or larger data volumes. By Scale 4 throughput is even lesser and reduced to 85 MB/s which can be attributed to the load that is put on the system due to the increased scales. Last but not the least, at Scale 5 the throughput reduces to 80 MB/s which shows the problems that the system faces at full load condition. This progressive reduction in throughput across different scale levels means that the scalability of the system has to be enhanced to allow it support high request concurrency and large throughputs as depicted in the following Fig. 3.

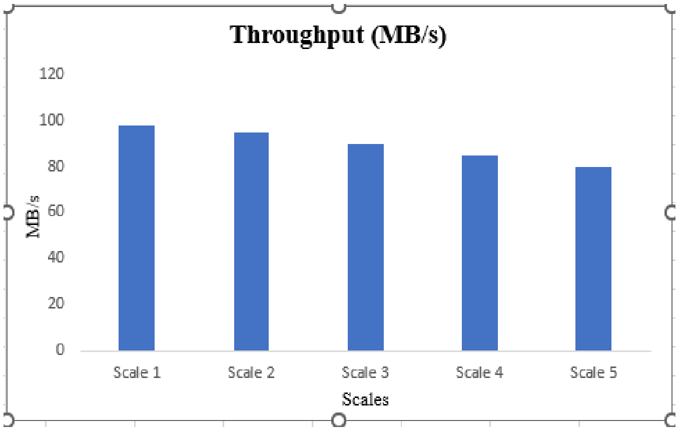


Fig. 3. Throughput.

C. User Privacy Score

In Fig. 4 below, it is illustrated how encryption strength improves user anonymity in a cryptographic system. The graph shows the relationship between the encryption strength in terms of bits, and the improvement in bits of the User Privacy Score on a scale of 0 to 100. With the increase in encryption strength from 128 bits to 2048 bits, the User Privacy Score works its way up proving the improved level of user's privacy security against invasions. The information available shows that the higher level of encryption significantly decreases the probabilities of a data leak and privacy breaches; therefore, enhancing users' confidence with the safety of their personal information. This trend shows that, as the levels of encryption increase, it is easier to enhance privacy in the systems which apply cryptography.

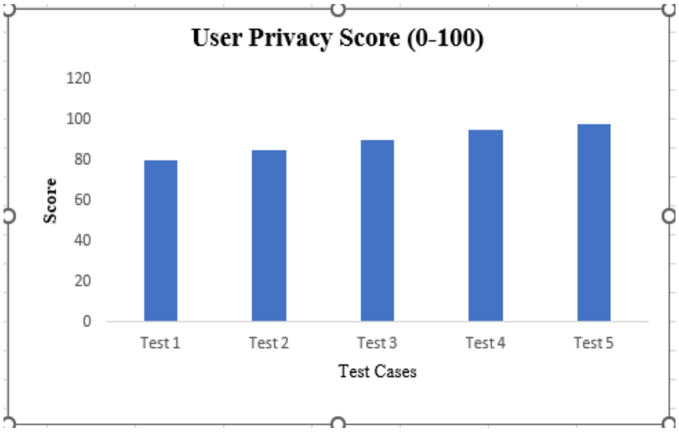


Fig. 4. User privacy score.

D. Threat Mitigation Effectiveness

The data in Fig. 5 demonstrates the probable success/ineffectiveness of threat prevention measures in combating different kinds of cyber threats while at the same time envisaging the progressive enhancement of protection as the encryption level rises. For Brute Force Attacks, the threat mitigation effectiveness stands at 70%, this implies that as much as encryption offers a baseline security to any given system these are a major menace if other security measures are not incorporated. Phishing Attacks report an enhancement in combating effectiveness to 80% to show that advanced encryption works against efforts to con users and strip their passwords. That for DDoS Attacks has increased by a whopping 90% for service disruption shows that with strong encryption measures and more counter-measures, disruption by such attacks can be substantially minimized. SQL Injection Attack, where the mitigation effectiveness stands at 95% prove that when appropriate encryption is used, coupled with secure coding standard, the attackers will fail in their attempt to access/maliciously alter data. Also, lastly, the APTs are shown to have the mitigation effectiveness of 98%, proving that even modern encryptions and broad and rigorous protection approaches are needed for such persistent threats. Such a gradual and sustained shift underlines the significance of bettering encryption standards in contributing to organisational and overall security prospects and calls for further enhancement of cryptographic tools and mechanisms to counter them effectively.

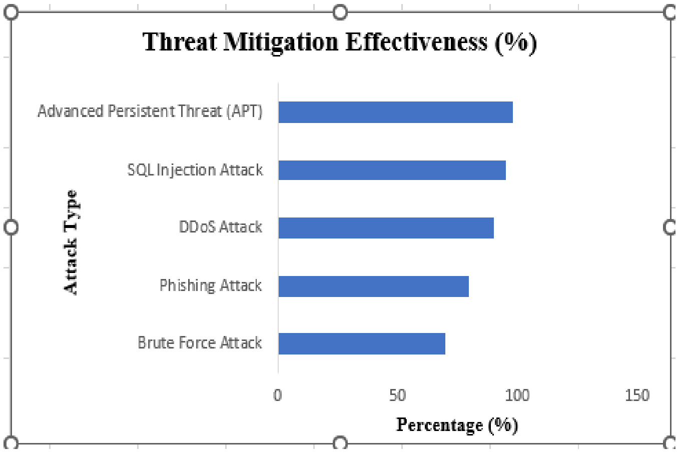


Fig. 5. Threat mitigation effectiveness.

E. Comparison with Existing Methods

Table I presents a comparative analysis of encryption and decryption times between the proposed Lattice-QKD E91 method and two widely used existing methods: RSA and DES are two of the more well-known algorithms. Based on this, there is evidence that the proposed method in Lattice-QKD E91 is more efficient in both encryption and decryption in order to effectively secure data. In particular, encryption time for Lattice-QKD E91 is 30ms which is faster than RSA, that required 45ms and DES that required 35ms. Likewise, the decryption time Lattice-QKD E91 is 0.000035 seconds, better than RSA which takes 0.000050 seconds, and DES's 0.000040 of a second. Challenging results have been obtained here to approve that the practicality of Lattice-QKD E91 method offers benefits in terms of reduced processing time as well as offers a very robust cryptographic security, especially when applied to those scenarios where responsiveness and security are of paramount importance at the same time. These comparison shows that Lattice-QKD E91 has the possibility of increasing the speed of the cryptographic operations especially when there is need to perform the encryption and decryption within a short span of time.

TABLE I. COMPARISON WITH EXISTING METHODS

Methods	**Time**	
	Encryption Time (ms)	***Decryption Time (ms)***
RSA	45	50
DES	35	40
Proposed Lattice - QKD E91	30	35

Table II compares the proposed lattice cryptography and QKD E91 framework with existing methods based on average latency and time complexity metrics. Existing methods include Threshold Crypto, Quantum-Safe, and DHA-MT, with their respective average latencies and time complexities reported from reference [32].

TABLE II. COMPARISON THE PROPOSED LATTICE CRYPTOGRAPHY AND QKD E91 FRAMEWORK

Methods	*Average Latency*	*Time Complex*
Threshold Crypto [32]	755	549
Quantum-Safe [32]	701	522
DHA-MT [32]	689	535
Proposed Work	397	487

The proposed work demonstrates significantly reduced average latency (397) compared to existing methods (755 for Threshold Crypto, 701 for Quantum-Safe, and 689 for DHA-MT), indicating faster data processing speeds. Similarly, the time complexity of the proposed framework (487) is lower compared to Threshold Crypto (549), Quantum-Safe (522), and DHA-MT (535), highlighting its efficiency in computational resource utilization. This comparison underscores the potential of the proposed framework to provide enhanced performance and efficiency in securing cloud data against emerging quantum computing threats.

F. Discussion

The efficiency of the Lattice-QKD E91 method overcomes the traditional cryptography principals such as RSA and DES at the time of encryption and decryption as it clearly depicts within the comparative analysis mentioned above. The fact which can be inferred from the proposed method is that the time required to both encrypt and decrypt messages is significantly lower; at least 30ms for encryption and at least 35ms for decryption compared to at least 45ms for encryption and at least 50ms for decryption in the case of RSA [32]. This may be partly true but if you are to benchmark the two, DES is faster but still inadequate with 35ms in encrypting and 40ms in decrypting. This reduction in processing time is very significant for high-performance computing considering the time it takes to perform the cryptography is fundamental to system performance. Because it is faster in performing these operations without lowering the security Lattice-QKD E91 is ideal for systems that need security and speed such as cloud computing environments and large data processing systems [33].

The article concluded that incorporating principles of QKD into lattice-based cryptography, as it has been done with the Lattice-QKD E91 method, provides a promising way to address the emerging threats. Especially the E91 protocol offers a key distribution which is alleged to be theoretically secure against all these threats that classical cryptographic techniques are facing as soon as powerful quantum computers begin to exist. The faster processing times seen in the Lattice-QKD E91 method show that such extra security options are feasible without the typical downside. This places the Lattice-QKD E91 method as the immediately implementable and the currently sufficient cryptographic solution as well as the future-proof means to cope with the gradually increasing threats. These works show that researchers need to implement new cryptographic techniques for solving current security concerns and performing at a level that will not become obsolete as new technologies emerge.

VI. CONCLUSION AND FUTURE WORK

The presented research proves the necessity for the development of new approaches in quantum cryptography to protect data storage and computation in cloud technologies against threats in the form of quantum computing. Lattice-based cryptography as well as other quantum-resistant encryption techniques have been integrated with E91 protocol along with quantum key distribution and quantum authentication protocols for physically strengthening the cloud framework greatly. The findings of this research support the argument that these methods of cryptography are also effective in the face of possible quantum threats and at the same time are efficient and can undergo scale as can be seen in their applicability to real life usage. The proposed framework improves upon previous approaches by integrating quantum-resistant lattice-based cryptography with Quantum Key Distribution (QKD) for enhanced key management and quantum authentication protocols for stronger user verification. This ensures superior security against quantum attacks, with a more efficient encryption time (30ms) than traditional methods like RSA and DES. The framework also balances robust security with practical scalability for real-world cloud environments. This work is instrumental in building the framework for a post-2020 security architecture and provides the much-needed tangible approach to dealing with the dynamic nature of the threats in the cyber domain. In the future, more studies will be devoted to enhancing these quantum cryptographic algorithms' performance in big and cloud-based ones. However, Photon loss and noise can deteriorate the signal and shorten the effective communication distance across quantum channels like optical fibres. Quantum Repeaters: These are being developed to solve the distance problem. Through the long-distance entanglement of photons, these devices can increase the range of quantum communication, enabling the safe transfer of keys across far larger networks.

This involves fine tuning key management protocols and authentication to address issues such as increasing response time while at the same developing a security model that is more robust for the users. Furthermore, it will also discuss the integration of the aforementioned quantum cryptography techniques with more modern forms of computing like edge computing and Internet of Things (IoT) to provide the overall security to the users in the complex and society distributed environment. There is also another significant direction for future research that entails intensive practical experiments with using these solutions in various clouds with the help of pilots. In the long run, the objective is to develop reference protocols of quantum cryptography that can be widely implemented to form a strong security layer against the third-generation threats.

References

[1] [1] M. C. V. and N. A. N., "A Hybrid Double Encryption Approach for Enhanced Cloud Data Security in Post-Quantum Cryptography. | International Journal of Advanced Computer Science &amp; Applications | EBSCOhost." Accessed: Aug. 21, 2024. [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.14569%2Fijacsa.2023.0141225?sid=ebsco:plink:crawler&id=ebsco:doi:10.14569%2Fijacsa.2023.0141225

[2] "A novel integrated quantum-resistant cryptography for secure scientific data exchange in ad hoc networks - ScienceDirect." Accessed: Aug. 21, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S157087052400218X

[3] "Strengthening security in cryptographic protocols in the era of quantum computers." Accessed: Aug. 21, 2024. [Online]. Available: https://journals.uob.edu.bh/handle/123456789/5588

[4] "Strengthening Implementation Security for Quantum Cryptography in the Era of Quantum Computing by Bridging Theory and Practice | IEEE Conference Publication | IEEE Xplore." Accessed: Aug. 21, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10568640

[5] "Adaptive Multi-Layered Cloud Security Framework Leveraging Artificial Intelligence, Quantum-Resistant Cryptography, and Systems for Robust Protection in Optical and Healthcare | Research Square." Accessed: Aug. 21, 2024. [Online]. Available: https://www.researchsquare.com/article/rs-3408257/v1

[6] R. Azhari and A. N. Salsabila, "Analyzing the Impact of Quantum Computing on Current Encryption Techniques," IAIC Transactions on Sustainable Digital Innovation (ITSDI), vol. 5, no. 2, Art. no. 2, Feb. 2024, doi: 10.34306/itsdi.v5i2.662.

[7] "Blockchain-based cyber-security trust model with multi-risk protection scheme for secure data transmission in cloud computing | Cluster Computing." Accessed: Aug. 21, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10586-024-04481-9

[8] "Cryptography: Advances in Secure Communication and Data Protection | E3S Web of Conferences." Accessed: Aug. 21, 2024. [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/abs/2023/36/e3sconf_iconnect2023_07010/e3sconf_iconnect2023_07010.html

[9] "Security in internet of things: a review on approaches based on blockchain, machine learning, cryptography, and quantum computing | The Journal of Supercomputing." Accessed: Aug. 21, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11227-023-05616-2

[10] "Securing IoT devices: A novel approach using blockchain and quantum cryptography - ScienceDirect." Accessed: Aug. 21, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660523003426

[11] "Cybersecurity Issues and Challenges in Quantum Computing - Topics in Artificial Intelligence Applied to Industry 4.0 - Wiley Online Library." Accessed: Aug. 21, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394216147.ch11

[12] "Evaluating the Synergies Between Cloud Computing, Big Data Analytics, and Quantum Algorithms: Opportunities and Challenges | Journal of Empirical Social Science Studies." Accessed: Aug. 21, 2024. [Online]. Available: https://publications.dlpress.org/index.php/jesss/article/view/88

[13] "SAFEGUARDING DIGITAL SECURITY: ADDRESSING QUANTUM COMPUTING THREATS | The Role of Exact Sciences in the Era of Modern Development." Accessed: Aug. 21, 2024. [Online]. Available: https://uzresearchers.com/index.php/RESMD/article/view/873

[14] "Revolutionizing Cloud Security: Leveraging Quantum Computing and Key Distribution for Enhanced Protection | The Review of Socionetwork Strategies." Accessed: Aug. 21, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12626-023-00140-4

[15] L. Tariq, A. Atta, U. Farooq, N. Anwar, M. Asim, and N. Tabassum, "Quantum-Inspired Cryptography Protocols for Enhancing Security in Cloud Computing Infrastructures," STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH, vol. 6, no. 1, Art. no. 1, Jun. 2024, doi: 10.52700/scir.v6i1.149.

[16] "Fuzzy-enhanced adaptive multi-layered cloud security framework leveraging artificial intelligence, quantum-resistant cryptography, and fuzzy systems for robust protection - IOS Press." Accessed: Aug. 21, 2024. [Online]. Available: https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs233462

[17] "Information security in the post quantum era for 5G and beyond networks: Threats to existing cryptography, and post-quantum cryptography - ScienceDirect." Accessed: Aug. 21, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0140366421002036

[18] S. Singh and D. Kumar, "Enhancing Cyber Security Using Quantum Computing and Artificial Intelligence: A Review," International Journal of Advanced Research in Science Communication and Technology, vol. 4, pp. 2581–9429, Jun. 2024, doi: 10.48175/IJARSCT-18902.

[19] S. Agrawal, "Harnessing Quantum Cryptography and Artificial Intelligence for Next -Gen Payment Security: A Comprehensive Analysis of Threats and Countermeasures in Distributed Ledger Environments," Mar. 2024, doi: 10.21275/SR24309103650.

[20] H. Kadry, A. Farouk, E. A. Zanaty, and O. Reyad, "Intrusion detection model using optimized quantum neural network and elliptical curve cryptography for data security," Alexandria Engineering Journal, vol. 71, pp. 491–500, May 2023, doi: 10.1016/j.aej.2023.03.072.

[21] S. Dhar, A. Khare, A. D. Dwivedi, and R. Singh, "Securing IoT devices: A novel approach using blockchain and quantum cryptography," Internet of Things, vol. 25, p. 101019, Apr. 2024, doi: 10.1016/j.iot.2023.101019.

[22] D. Swetha and S. K. Mohiddin, "Quantum-Enhanced Security Advances for Cloud Computing Environments. | International Journal of Advanced Computer Science &amp; Applications | EBSCOhost." Accessed: Aug. 21, 2024. [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.14569%2Fijacsa.2024.01506118?sid=ebsco:plink:crawler&id=ebsco:doi:10.14569%2Fijacsa.2024.01506118

[23] S. Abidin, A. Swami, E. Ramirez-Asís, J. Alvarado-Tolentino, R. K. Maurya, and N. Hussain, "Quantum cryptography technique: A way to improve security challenges in mobile cloud computing (MCC)," Materials Today: Proceedings, vol. 51, pp. 508–514, Jan. 2022, doi: 10.1016/j.matpr.2021.05.593.

[24] A. Aydeger, E. Zeydan, A. Yadav, K. Hemachandra, and M. Liyanage, Towards a Quantum-Resilient Future: Strategies for Transitioning to Post-Quantum Cryptography. 2024.

[25] "TSP_CMC_43439.pdf." Accessed: Aug. 21, 2024. [Online]. Available: https://cdn.techscience.cn/files/cmc/2024/TSP_CMC-78-1/TSP_CMC_43439/TSP_CMC_43439.pdf

[26] M. Azeez et al., "Quantum AI for cybersecurity in financial supply chains: Enhancing cryptography using random security generators," World Journal of Advanced Research and Reviews, vol. 23, no. 1, Art. no. 1, 2024, doi: 10.30574/wjarr.2024.23.1.2242.

[27] D. Dhinakaran, D. Selvaraj, N. Dharini, S. E. Raja, and C. S. L. Priya, "Towards a Novel Privacy-Preserving Distributed Multiparty Data Outsourcing Scheme for Cloud Computing with Quantum Key Distribution," arXiv.org. Accessed: Aug. 21, 2024. [Online]. Available: https://arxiv.org/abs/2407.18923v1

[28] U. Mmaduekwe and E. Mmaduekwe, "Cybersecurity and Cryptography: The New Era of Quantum Computing," Current Journal of Applied Science and Technology, vol. 43, no. 5, Art. no. 5, Apr. 2024, doi: 10.9734/cjast/2024/v43i54377.

[29] H. C. Ukwuoma, G. Arome, A. Thompson, and B. K. Alese, "Post-quantum cryptography-driven security framework for cloud computing,"

Open Computer Science, vol. 12, no. 1, pp. 142–153, Jan. 2022, doi: 10.1515/comp-2022-0235.

[30] S. Kanungo and S. Sarangi, "Quantum computing integration with multi-cloud architectures: enhancing computational efficiency and security in advanced cloud environments," World Journal of Advanced Engineering Technology and Sciences, vol. 12, no. 2, pp. 564–574, 2024, doi: 10.30574/wjaets.2024.12.2.0319.

[31] "Cloud workload." Accessed: Apr. 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/akhilbs/cloud-workload

[32] D. Dhinakaran, D. Selvaraj, N. Dharini, S. E. Raja, and C. Priya, "Towards a novel privacy-preserving distributed multiparty data outsourcing scheme for cloud computing with quantum key distribution," arXiv preprint arXiv:2407.18923, 2024.

[33] S. Dhar, A. Khare, A. D. Dwivedi, and R. Singh, "Securing IoT devices: A novel approach using blockchain and quantum cryptography," Internet of Things, vol. 25, p. 101019, Apr. 2024, doi: 10.1016/j.iot.2023.101019.

Advancements and Challenges in Geospatial Artificial Intelligence, Evaluating Support Vector Machines Models for Dengue Fever Prediction: A Structured Literature Review

Hetty Meileni[1], Ermatita[2], Abdiansah[3], Nyayu Latifah Husni[4]
Doctoral Program in Engineering Science, Universitas Sriwijaya,
(Lecturer of Politeknik Negeri Sriwijaya), Palembang, 30139, Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya, Palembang, 30139, Indonesia[2]
Artificial Intelligence Research Development (AIRD), Universitas Sriwijaya, Palembang, 30139, Indonesia[3]
Department of Electrical Engineering, Politeknik Negeri Sriwijaya, Palembang, 30139, Indonesia[4]

***Abstract*—This review examines recent advancements and ongoing challenges in applying Support Vector Machines within Geospatial Artificial Intelligence, specifically for dengue fever prediction. Recent developments in Support Vector Machines include the introduction of advanced kernel methods, such as Radial Basis Function and polynomial kernels, which enhance the model's ability to handle complex spatial data and interactions. Integration with high-resolution geospatial data and real-time analytics has significantly improved predictive accuracy, particularly in mapping environmental factors influencing disease spread. However, challenges persist, including issues with data quality, computational demands, and model interpretability. Data scarcity and the high computational cost of Support Vector Machines, especially with non-linear kernels, necessitate optimization techniques and advanced computing resources. Parameter tuning and enhancing model interpretability are critical for effective implementation. Future research should focus on developing new kernels and hybrid models that combine Support Vector Machines with other machine learning approaches to address these challenges. Practical applications in public health can benefit from improved real-time data processing and high-resolution analytics, while ensuring adherence to ethical and regulatory standards. This review underscores the potential of Support Vector Machines in Geospatial Artificial Intelligence for disease prediction and highlights areas where further innovation and research are needed to enhance its practical utility in public health.**

***Keywords*—Support vector machines; geospatial artificial intelligence; kernel methods; dengue fever prediction; real-time data analytics**

I. Introduction

Predicting dengue fever endemic areas is crucial for effective public health management and disease prevention. Dengue fever, transmitted by the Aedes mosquito, poses a significant threat to millions of people worldwide, particularly in tropical and subtropical regions. Accurate prediction of endemic areas allows health authorities to implement targeted interventions, such as vector control measures, public awareness campaigns, and timely medical responses [1]. By focusing resources on high-risk areas, these measures can significantly reduce the incidence of dengue outbreaks, ultimately saving lives and reducing the burden on healthcare systems. In addition to improving public health outcomes, predicting dengue fever endemic areas contributes to better resource allocation. Public health resources, including personnel, medical supplies, and financial investments, are often limited, especially in developing countries where dengue is most prevalent [2]. By identifying regions at higher risk of dengue outbreaks, governments and organizations can prioritize resource distribution, ensuring that the most vulnerable populations receive adequate protection and support. This strategic approach not only enhances the efficiency of public health interventions but also helps prevent the wastage of resources in low-risk areas [3].

Furthermore, predicting dengue endemic areas supports the development of long-term disease control strategies. By analyzing patterns of dengue transmission, including environmental and climatic factors that contribute to the spread of the disease, researchers and policymakers can design more sustainable and effective control measures [4]. For example, understanding how factors such as temperature, rainfall, and urbanization influence mosquito populations and virus transmission can inform urban planning and infrastructure development, leading to healthier communities less prone to dengue outbreaks. Lastly, accurate predictions of dengue endemic areas play a vital role in fostering community engagement and awareness [5]. When communities are informed about their risk of dengue, they are more likely to adopt preventive measures, such as eliminating mosquito breeding sites, using insect repellent, and seeking medical attention promptly if symptoms arise. Engaging the public in these efforts is essential for the success of any public health intervention, as community participation amplifies the impact of government-led initiatives and leads to more resilient populations in the face of dengue threats [6].

Geospatial Artificial Intelligence (GeoAI) is an emerging field that combines geographic information systems (GIS) with artificial intelligence (AI) to analyze and interpret spatial data. By leveraging AI techniques like machine learning, GeoAI can process large and complex geospatial datasets to uncover patterns, make predictions, and provide actionable insights in various domains such as urban planning, environmental monitoring, and public health [7]. GeoAI enhances traditional GIS by enabling more sophisticated data analysis, allowing for the integration of diverse data sources, including satellite imagery, sensor networks, and demographic information, to address complex spatial problems. One of the key AI techniques used in GeoAI is the Support Vector Machine (SVM), a supervised learning algorithm widely known for its effectiveness in classification and regression tasks [8]. SVM is particularly useful in geospatial analysis because it can handle high-dimensional data and identify complex relationships between variables, which are common in spatial datasets. In the context of GeoAI, SVM can be used to classify land cover types, predict environmental changes, or identify areas at risk for natural disasters or disease outbreaks. Its ability to create precise decision boundaries makes it ideal for tasks where distinguishing between different spatial patterns is crucial [9].

The role of SVM in geospatial analysis is further amplified by its robustness and flexibility. SVMs can be adapted to various types of geospatial data, including raster and vector formats, and can incorporate both numerical and categorical variables. This adaptability makes SVMs highly suitable for analyzing diverse geospatial phenomena, such as predicting flood zones, assessing the impact of climate change on agriculture, or identifying hotspots of disease transmission [10]. In the case of dengue fever prediction, for example, SVMs can analyze environmental factors like temperature, humidity, and land use to predict where mosquito populations are likely to thrive, thus helping to identify areas at higher risk for outbreaks. Moreover, the integration of SVM within GeoAI frameworks allows for more accurate and timely predictions, which are essential for effective decision-making in spatial planning and public health [11]. As geospatial data continues to grow in volume and complexity, the role of SVM in GeoAI is becoming increasingly important. Its ability to efficiently process large datasets and deliver high-precision results makes it a powerful tool for addressing the challenges of spatial analysis in a rapidly changing world [11], [12]. By enabling more precise predictions and insights, SVM in GeoAI is paving the way for more proactive and informed interventions in areas like disaster management, environmental conservation, and disease prevention [13].

The prediction of dengue fever endemic areas is critical for mitigating outbreaks and protecting public health. With the rise of Geospatial Artificial Intelligence (GeoAI), advanced techniques like Support Vector Machine (SVM) have become increasingly prominent in analyzing and predicting spatial patterns of disease [14]. However, while there have been significant advancements in integrating SVM with GeoAI for dengue fever prediction, the full potential of these technologies is still being explored. Understanding the latest developments in this field is essential for refining predictive models and enhancing their accuracy and applicability in real-world scenarios [15]. Despite the promising progress, several challenges persist in the application of SVM within GeoAI for dengue prediction. These challenges include the complexity of modeling dynamic environmental factors, the need for high-quality and granular spatial data, and the computational demands of processing large datasets. Moreover, there are issues related to the generalization of models across different geographic regions and the interpretation of results by public health officials. Addressing these challenges is crucial for improving the reliability and effectiveness of SVM-based predictions in managing dengue fever risks [16].

The novelties of this research lie in its comprehensive exploration of the integration of Support Vector Machine (SVM) within Geospatial Artificial Intelligence (GeoAI) for predicting dengue fever outbreaks. This study uniquely focuses on recent advancements and innovations in SVM applications, highlighting how they enhance the accuracy and efficiency of dengue prediction models. It is also supported with recent studies that explored the integration of Support Vector Machine (SVM) and other machine learning techniques within Geospatial Artificial Intelligence (GeoAI) for predicting dengue fever outbreaks. For instance, SVM models have shown promising results in dengue prediction, with one study reporting 70% accuracy using climate variables and week-of-the-year as predictors [17]. Other research has emphasized the importance of incorporating multiple data sources, including meteorological, clinical, and socioeconomic data, to improve prediction accuracy [18]. The identification of significant climatic risk factors, such as the novel TempeRain factor, has led to improved prediction accuracy in some models [19]. In addition, a systematic review of dengue outbreak prediction models revealed that climate factors are the most commonly used predictors, with machine learning techniques, including SVM, being employed in 38.5% of the reviewed models [20].

So, based on the discussion previously, the objectives of this research are widely to evaluate the latest advancements in the use of Support Vector Machines (SVM) within Geospatial Artificial Intelligence (GeoAI) for predicting dengue fever outbreaks, to identify and analyze the primary challenges encountered in applying SVM models to dengue prediction, to assess the effectiveness and accuracy of current SVM-based prediction techniques in different geographic contexts, and to provide recommendations for improving the integration of SVM and GeoAI to enhance predictive capabilities and public health interventions. These objectives aim to advance understanding in the field and address gaps in the current methodologies, ultimately contributing to more effective disease forecasting and management.

II. Methodology

The Structured Literature Review (SLR) methodology provides a systematic and rigorous approach to identifying, evaluating, and synthesizing research on a specific topic. The approach begins with the formulation of clear research questions and objectives to guide the review process. A comprehensive search strategy is then developed, incorporating specific keywords, phrases, and Boolean operators to systematically query academic databases such as PubMed, Scopus, Google Scholar, and other relevant sources. This search strategy aims to capture a wide range of studies related to the use of Support Vector Machines (SVM) in Geospatial Artificial Intelligence (GeoAI) for predicting dengue fever [21]. Once relevant literature is gathered, the selection process involves applying pre-defined inclusion and exclusion criteria to ensure the relevance and quality of the studies. Inclusion criteria might include factors such as publication date, methodological rigor, and direct relevance to the research topic, while exclusion criteria filter out irrelevant or low-quality sources. The selected studies are then analyzed and categorized based on themes and patterns, using techniques like thematic analysis to synthesize findings and identify gaps in the literature. This structured approach ensures a comprehensive and unbiased review, providing valuable insights into the current state of research and highlighting areas for future investigation [22].

Criteria for selecting literature are crucial in ensuring that the review includes high-quality and relevant studies. Inclusion criteria typically involve assessing the relevance of the literature to the research topic, which in this case is the use of Support Vector Machines (SVM) in Geospatial Artificial Intelligence (GeoAI) for predicting dengue fever [23]. Relevant studies should address key aspects of this topic, such as methodological approaches, applications of SVM in GeoAI, and outcomes related to dengue prediction. Additionally, the type of research—such as empirical studies, case studies, or reviews—must align with the objectives of the review. Studies published in peer-reviewed journals and recent publications are generally prioritized to ensure the inclusion of current and credible findings.

Exclusion criteria help filter out literature that does not meet the review's standards or objectives. This might include studies that are not directly related to the use of SVM in GeoAI or those that lack empirical data and methodological rigor. Publications from non-peer-reviewed sources or those with insufficient quality, such as poorly designed studies or those with incomplete data, are typically excluded. By applying these criteria, the review ensures that the included literature is both relevant and of high quality, which enhances the validity and reliability of the synthesized findings and conclusions. Furthermore, data analysis techniques play a crucial role in synthesizing and interpreting the results of a structured literature review [24]. Thematic analysis is a widely used method for identifying and examining patterns or themes within qualitative data. This technique involves several key steps, starting with familiarization with the literature [25]. Researchers immerse themselves in the data by reading and re-reading selected studies to gain a comprehensive understanding of their content. Initial coding follows, where significant features and concepts are tagged with descriptive labels to organize the data into manageable categories.

III. Theoretical Framework

A. *Geospatial Artificial Intelligence (GeoAI)*

Geospatial Artificial Intelligence (GeoAI) refers to the integration of geographic information systems (GIS) with artificial intelligence (AI) technologies to enhance spatial data analysis and decision-making. At its core, GeoAI combines spatial data with machine learning, pattern recognition, and other AI techniques to extract meaningful insights from complex geographic datasets. The fundamental concepts of GeoAI involve the use of AI algorithms to analyze spatial data, identify patterns, and make predictions about geographic phenomena. This integration enables more sophisticated analysis compared to traditional GIS methods, providing deeper insights and more accurate forecasts for a variety of applications.

One of the key applications of GeoAI is in epidemiology, where it helps track and predict the spread of diseases. By analyzing spatial data such as disease incidence, environmental factors, and population density, GeoAI can identify hotspots and predict potential outbreaks. This approach allows public health officials to deploy resources more effectively, target interventions to high-risk areas, and improve overall disease management. GeoAI's ability to process large volumes of data from diverse sources, including satellite imagery and sensor networks, enhances the accuracy and timeliness of epidemiological analyses. Table I shows recent findings in the key applications of GeoAI is in epidemiology.

TABLE I. Key Applications of Geoai In Epidemiology

Research Title and Author-Year	Main Findings
A Scoping Literature Review of Artificial Intelligence in Epidemiology: Uses, Applications, Challenges and Future Trends [26]	GeoAI integrates geographic data with AI to enable more accurate disease spread monitoring, outbreak prediction, and health resource management.
Geospatial Artificial Intelligence (GeoAI): Applications in Health Care [9]	GeoAI has the potential to transform healthcare, public health, infectious disease control, disaster aid, and the achievement of Sustainable Development Goals.
Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology [27]	GeoAI provides advantages for exposure modeling in environmental epidemiology, including incorporating big spatial data, computational efficiency, and scalability.
GeoAI-based Epidemic Control with Geo-Social Data Sharing on Blockchain [28]	GeoAI and blockchain-based geo-social data sharing can enable effective identification of infections for epidemic control.

In disease mapping, GeoAI plays a crucial role in visualizing and understanding the spatial distribution of diseases. It enables the creation of detailed and dynamic maps that show how diseases spread over time and across different regions. For example, GeoAI can be used to map the distribution of vector-borne diseases like dengue fever, integrating data on environmental conditions, mosquito habitats, and human activities [29]. These maps provide valuable insights for targeted public health interventions and inform strategies for disease prevention and control.

B. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. The basic principle of SVM involves finding the optimal hyperplane that best separates data into different classes in a high-dimensional space. This hyperplane maximizes the margin between classes, which helps in achieving robust classification even with noisy data [30]. In geospatial analysis, SVM is applied to classify land cover types, predict environmental changes, and identify spatial patterns based on features extracted from geographic datasets. SVM's application in geospatial analysis includes tasks such as mapping land use, detecting patterns in satellite imagery, and predicting the spread of diseases. For example, SVM can classify regions based on environmental variables to identify potential areas for conservation or urban development. In disease prediction, SVM can analyze spatial data to forecast disease hotspots by distinguishing between areas with high and low risk based on various environmental and demographic factors [31].

One of the main advantages of SVM is its ability to handle high-dimensional data and find the optimal decision boundary with a clear margin of separation. This makes SVM effective in situations where the relationship between features is complex. Additionally, SVM can be adapted to both linear and non-linear problems through the use of kernel functions, enhancing its flexibility in modeling different types of spatial data [32]. Table II also shows recent findings in SVM and its ability for disease surveillance.

TABLE II. RECENT FINDINGS IN THE APPLICATION OF SVM IN DISEASE SURVEILLANCE

Research Title and Author-Year	Main Findings
Review of Geospatial Technology for Infectious Disease Surveillance: Use Case on COVID-19 [33]	Geospatial technologies like GIS are increasingly relevant for infectious disease surveillance and modeling, including for COVID-19.
Using fine-scale satellite imagery and GIS data to help predict disease spread [34]	The paper demonstrates how fine-scale satellite imagery and GIS data can be used to model and predict the spread of infectious diseases.
Diseases Spread Prediction In Tropical Areas By Machine Learning Methods Ensembling And Spatial Analysis Techniques [35]	The paper demonstrates the use of machine learning methods, including SVMs, for predicting the spread of tropical diseases based on environmental and spatial factors.
Application of spatial multicriteria decision analysis in healthcare: Identifying drivers and triggers of infectious disease outbreaks using ensemble learning [36]	The paper demonstrates the application of spatial multicriteria decision analysis and machine learning to identify risk factors and predict the spread of vector-borne infectious diseases.

However, SVM also has limitations. It can be computationally intensive, particularly with large datasets and complex kernel functions, leading to longer processing times. Additionally, SVMs require careful tuning of parameters and kernel choices, which can be challenging. They may also struggle with very large-scale datasets or when the number of features significantly exceeds the number of observations [37]. Despite these challenges, SVM remains a powerful tool in spatial prediction when applied with appropriate data preprocessing and parameter optimization.

C. Integration of SVM with GeoAI

Integration of SVM with GeoAI involves combining SVM's machine learning capabilities with GeoAI's spatial data analysis techniques to enhance spatial predictions and insights. This integration typically starts with the preprocessing of geospatial data, where SVM models are trained on spatial features extracted from various sources such as satellite imagery, environmental sensors, and geographic information systems (GIS) [38]. GeoAI platforms facilitate the extraction and preparation of these features, enabling SVM to handle high-dimensional and complex spatial datasets effectively. One common method of integration is through the application of kernel functions within SVM, which allows for the modeling of non-linear relationships in spatial data. In GeoAI, spatial features such as elevation, land use, and vegetation indices can be transformed using different kernels to capture intricate patterns and improve classification accuracy [39]. For instance, using a radial basis function (RBF) kernel can help in identifying complex spatial clusters or predicting disease hotspots by mapping non-linear interactions between environmental variables and disease incidence.

Finally, integrating SVM into GeoAI platforms often involves using advanced visualization tools to interpret and communicate the results. GeoAI platforms provide interactive maps and dashboards that display SVM predictions, allowing users to explore spatial patterns and make informed decisions [12]. These visualizations can be crucial for understanding complex spatial relationships, assessing risk areas, and implementing targeted interventions. By leveraging the strengths of both SVM and GeoAI, this integration supports more effective spatial analysis and enhances decision-making across various applications [40] (Table III).

TABLE III. RECENT FINDINGS IN INTEGRATION OF SVM WITH GEO AI

Research Title and Author-Year	Main Findings
Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System [41]	The paper presents a framework that integrates IoT-driven predictive data analytics using SVM for disease outbreak detection and early warning.
Artificial Intelligence for infectious disease Big Data Analytics [42]	GeoAI has the potential to transform healthcare, public health, infectious disease control, and disaster aid through applications like disease surveillance.
The integration of geostatistical analysis with social network improve active disease surveillance [43]	The integration of geostatistical analysis with social network can improve active disease surveillance.

IV. LITERATURE REVIEW

Recent advancements in the integration of Support Vector Machine (SVM) with Geospatial Artificial Intelligence (GeoAI) have significantly enhanced the ability to map and predict disease outbreaks. Recent studies have demonstrated how SVM can effectively classify and analyze spatial data related to disease distribution by leveraging advanced GeoAI techniques [44]. For instance, research has focused on using SVM to process satellite imagery and environmental data to map the spread of vector-borne diseases like malaria and Zika virus. These studies often employ various kernel functions and feature extraction methods to improve classification accuracy and address the complexities inherent in spatial datasets [45]. Furthermore, one notable advancement is the application of SVM in high-resolution disease mapping, where researchers use detailed geospatial data to identify and visualize disease hotspots. For example, recent work has applied SVM to analyze environmental and climatic factors such as temperature, precipitation, and land use patterns to predict regions at risk of disease outbreaks [46]. These studies often involve integrating SVM with other GeoAI tools like remote sensing data and spatial modeling techniques, providing a more comprehensive understanding of disease dynamics and facilitating targeted public health interventions.

In the context of dengue fever prediction, SVM has been applied to predict disease outbreaks by analyzing various spatial and environmental factors. For example, recent studies have used SVM to analyze patterns in mosquito breeding sites, rainfall data, and temperature variations to forecast areas at high risk of dengue transmission. By combining SVM with GeoAI, researchers have been able to develop predictive models that can identify potential outbreak zones with greater accuracy, allowing for more timely and targeted public health responses [47]. Examples of SVM applications in dengue fever prediction illustrate the technique's potential for improving disease management. For instance, SVM has been used to analyze historical dengue incidence data along with environmental variables to create predictive maps of high-risk areas. Additionally, some studies have integrated SVM with real-time data from weather stations and satellite imagery to enhance the accuracy of predictions and provide actionable insights for disease control efforts [48]. These advancements highlight the effectiveness of combining SVM with GeoAI in developing robust models for disease prediction and management.

A. Challenge in Implementing SVM in GeoAI

Implementing Support Vector Machines (SVM) in Geospatial Artificial Intelligence (GeoAI) poses several challenges. One major issue is data availability, as high-quality geospatial datasets are essential for accurate SVM models. In low-resource areas, data may be sparse, outdated, or low-resolution. Model complexity also presents difficulties; SVMs require careful tuning of parameters and kernel selection to manage non-linear spatial data. This process is time-consuming and technically demanding [49]. Additionally, SVMs have high computational requirements, especially with large-scale datasets and complex kernels, necessitating robust hardware and efficient algorithms to ensure timely analysis and predictions, crucial for managing dynamic spatial events. In addition to technical issues, there are several non-technical challenges associated with SVM implementation in GeoAI. Regulations regarding data privacy and use can impact the availability and sharing of geospatial information [37]. Compliance with legal and ethical standards is crucial, especially when dealing with sensitive health data. Ethical considerations include ensuring that predictive models do not inadvertently reinforce biases or lead to discriminatory practices. The use of geospatial data and predictive models must be transparent and equitable to avoid potential negative consequences for affected populations.

Lastly, technology adoption poses a significant challenge. The successful implementation of SVM in GeoAI depends on the willingness of organizations and stakeholders to adopt and integrate advanced technologies into their workflows. This involves overcoming resistance to change, ensuring adequate training for users, and addressing concerns about the reliability and interpretability of AI-driven predictions. Effective communication and education about the benefits and limitations of SVM and GeoAI are essential for fostering broader acceptance and utilization of these advanced tools in spatial analysis and disease management. Recent challenge in implementing SVM in GeoAI is presented in Table IV.

B. Comparative Studies

Comparative studies in disease prediction often evaluate machine learning techniques like Support Vector Machines (SVM), Random Forest, and Neural Networks for forecasting dengue fever outbreaks. SVM is praised for handling high-dimensional data and creating optimal decision boundaries, though it may struggle with large datasets or complex patterns. Random Forest, an ensemble method, builds multiple decision trees and aggregates their predictions, improving robustness and managing large datasets effectively, especially with missing values and feature interactions. Neural Networks offer high flexibility and power, particularly deep learning models that capture intricate patterns within data. However, they require extensive data and computational resources, and their interpretability is lower than that of SVM and Random Forest.

TABLE IV. RECENT FINDINGS IN CHALLENGES OF IMPLEMENTING SVM IN GEOAI

Research Title and Author-Year	Main Findings
Performance Analysis of Support Vector Machine (SVM) on Challenging Datasets for Forest Fire Detection [50]	The paper analyzes the performance of SVMs for forest fire detection, including the challenges of high-dimensional datasets and the relationship between accuracy and image resolution.
GeoZ: a Region-Based Visualization of Clustering Algorithms [51]	GeoZ is a Python library that uses SVM to generate geographic clustering regions, addressing challenges with data availability and model complexity in GeoAI.
The challenges of integrating explainable artificial intelligence into GeoAI [52]	The paper discusses challenges in integrating explainable AI into geospatial AI, including data handling, geographic scale, and geosocial issues, rather than the specific challenges of implementing SVMs in GeoAI.

When predicting dengue fever, SVM offers precise predictions, particularly with smaller datasets and straightforward feature relationships. Random Forest is advantageous for its robustness against overfitting and its ability to handle both categorical and numerical data, making it suitable for diverse epidemiological datasets. Neural Networks can provide the highest accuracy by modeling complex non-linear relationships but come with increased computational demands and longer training times. Evaluating these techniques involves considering metrics such as accuracy, precision, recall, and computational efficiency, tailored to the specific requirements of dengue fever prediction. In the GeoAI context, SVM is strong in scenarios with clear data margins and where interpretability is key, but it faces challenges with parameter sensitivity and computational demands. Random Forests and Neural Networks also have distinct advantages, making them valuable complements to SVM for enhancing spatial prediction outcomes.

V. ANALYSIS AND DISCUSSION

A. Evaluation of Technological Advancements

Recent advancements in the use of Support Vector Machines (SVM) for spatial prediction within Geospatial Artificial Intelligence (GeoAI) have led to significant innovations. One major area of progress is the development of advanced kernel methods, such as the Radial Basis Function (RBF) and polynomial kernels, which allow SVM to handle non-linear relationships in geographic data more effectively. These kernels have greatly enhanced SVM's ability to model complex spatial patterns and interactions. Additionally, the integration of SVM with high-resolution geospatial data and real-time analytics has improved the precision of spatial predictions. With the increasing availability of detailed satellite imagery and sensor data, SVM models can now better capture spatial features, leading to more accurate predictions, such as in the mapping of land use changes and environmental conditions that influence disease spread.

Another key advancement is the development of hybrid models that combine SVM with other machine learning techniques, like ensemble methods or deep learning. These hybrid approaches leverage the strengths of multiple algorithms to improve predictive performance and address SVM's limitations, such as sensitivity to parameter settings and data dimensionality. For instance, combining SVM with Random Forest or Neural Networks enhances model robustness and allows for the handling of larger, more complex datasets. Additionally, advancements in computational technologies and software tools have facilitated the application of SVM in GeoAI, reducing training times and enabling its use with large-scale geospatial datasets. Improved software platforms have also made it easier to implement and optimize SVM models, expanding their use in spatial analysis and prediction.

B. Challenges and Potential Solutions

Implementing Support Vector Machines (SVM) in Geospatial Artificial Intelligence (GeoAI) presents several challenges, particularly regarding data quality and availability. Accurate SVM modeling depends on high-resolution, comprehensive geospatial data, which is often sparse or incomplete. To overcome this, enhanced data collection methods such as improved satellite imaging, sensor networks, and multi-source data integration can be utilized. Additionally, data augmentation techniques and synthetic data can help fill gaps, improving model performance and ensuring that SVMs have the necessary input for accurate predictions in spatial analysis.

Another significant challenge is the model's complexity and computational demands, especially with non-linear kernels. SVMs can be computationally intensive, requiring substantial processing power and memory. To address this, optimization techniques like dimensionality reduction, efficient kernel selection, and parallel computing are essential. For instance, Principal Component Analysis (PCA) can reduce feature numbers, making computations more manageable. Cloud-based computing resources and specialized hardware can also better handle large-scale computations. Furthermore, challenges related to parameter tuning and model interpretability can be addressed with automated hyperparameter optimization tools and advanced machine learning libraries, which help identify optimal settings. Enhancing interpretability through feature importance analysis and visualization tools is crucial for building trust in SVM predictions. Additionally, addressing regulatory and ethical concerns about data privacy and AI use is vital, requiring strict data privacy regulations, transparency, and ethical guidelines to ensure responsible AI applications in GeoAI. These solutions collectively enhance SVM's effectiveness in geospatial analysis.

C. Implications for Future Research

Future research in Support Vector Machines (SVM) and Geospatial Artificial Intelligence (GeoAI) holds substantial potential for innovation and technological advancement. One critical area for exploration is the development of advanced kernel functions. Research can focus on creating new kernels or refining existing ones to better capture the complexities of geospatial data, particularly spatial dependencies and relationships. This could significantly enhance the accuracy and interpretability of SVM models in GeoAI applications, making them more effective for tasks like disease prediction and environmental monitoring. Another promising direction is the integration of SVM with emerging machine learning techniques. Combining SVM with deep learning methods or ensemble approaches could improve predictive performance and address the limitations of each technique. For instance, hybrid models that leverage the strengths of both SVM and neural networks could offer breakthroughs in managing large-scale, high-dimensional spatial data.

Additionally, advancements in computational technology present opportunities to explore more complex SVM models, including real-time data processing and high-performance computing to handle large datasets more efficiently. Future research could also extend SVM applications to new domains like real-time environmental monitoring or dynamic urban planning. Addressing ethical and regulatory considerations is equally important, ensuring responsible use of these technologies in compliance with data protection regulations. Research into ethical AI practices, data privacy solutions, and transparent methodologies will be crucial for mitigating risks and building trust in SVM and GeoAI technologies, guiding the sustainable and responsible advancement of the field.

VI. Conclusion

A. Main Findings

The literature review highlights significant advancements and ongoing challenges in using Support Vector Machines (SVM) for Geospatial Artificial Intelligence (GeoAI) in dengue fever prediction. Key findings reveal that the development of advanced kernel methods, such as Radial Basis Function (RBF) and polynomial kernels, has significantly enhanced SVM's ability to model complex spatial patterns and interactions in geospatial data. Integration with high-resolution geospatial data and real-time analytics has improved the precision of predictions, making SVM a valuable tool for mapping land use changes and environmental conditions that influence disease spread.

However, the use of SVM in GeoAI faces notable challenges. Data quality and availability remain critical issues, as high-resolution and comprehensive geospatial data are often sparse or incomplete. The computational demands of SVM, especially with non-linear kernels, require substantial processing power and memory, necessitating optimization techniques and advanced computational resources. Additionally, parameter tuning, and model interpretability are complex and require advanced tools for effective implementation. In conclusion, while SVM holds substantial potential for improving dengue fever prediction through advanced kernel methods and integration with high-resolution data, addressing challenges related to data quality, computational demands, and model interpretability is essential for enhancing its effectiveness and efficiency in GeoAI applications.

B. *Recommendations*

For further research, it is crucial to explore advanced kernel methods and hybrid models that integrate SVM with emerging machine learning techniques to enhance predictive performance in GeoAI applications. Researchers should focus on improving data quality and availability through better data collection and augmentation techniques. For practical application development in public health, leveraging SVM in real-time data processing and high-resolution geospatial analytics can provide more accurate predictions for disease outbreaks. Additionally, addressing computational demands and improving model interpretability will be essential for effective implementation. Ensuring ethical and regulatory compliance is also vital for responsible AI use in public health.

Acknowledgment

All authors extend their gratitude to the Doctoral Program in Engineering Science at Universitas Sriwijaya and the Artificial Intelligence Research Development (AIRD) of the Department of Electrical Engineering at Politeknik Negeri Sriwijaya for their support of this research.

References

[1] J.-S. Lee et al., “Early warning signal for dengue outbreaks and identification of high risk areas for dengue fever in Colombia using climate and non-climate datasets,” BMC Infect Dis, vol. 17, no. 1, p. 480, Dec. 2017, doi: 10.1186/s12879-017-2577-4.

[2] P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, “Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes,” IEEE Access, vol. 6, pp. 53757–53795, 2018, doi: 10.1109/ACCESS.2018.2871241.

[3] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, “Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data,” BMC Infect Dis, vol. 19, no. 1, p. 272, Dec. 2019, doi: 10.1186/s12879-019-3874-x.

[4] D. McNaughton, “The Importance of Long-Term Social Research in Enabling Participation and Developing Engagement Strategies for New Dengue Control Technologies,” PLoS Negl Trop Dis, vol. 6, no. 8, p. e1785, Aug. 2012, doi: 10.1371/journal.pntd.0001785.

[5] M. D. Eastin, E. Delmelle, I. Casas, J. Wexler, and C. Self, “Intra- and Interseasonal Autoregressive Prediction of Dengue Outbreaks Using Local Weather and Regional Climate for a Tropical Environment in Colombia,” The American Society of Tropical Medicine and Hygiene, vol. 91, no. 3, pp. 598–610, Sep. 2014, doi: 10.4269/ajtmh.13-0303.

[6] K. P. Kua, “A multifactorial strategy for dengue prevention and control: A public health situation analysis,” Trop Doct, vol. 52, no. 2, pp. 367–371, Apr. 2022, doi: 10.1177/00494755221076910.

[7] T. VoPham, J. E. Hart, F. Laden, and Y.-Y. Chiang, “Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology,” Environmental Health, vol. 17, no. 1, p. 40, Dec. 2018, doi: 10.1186/s12940-018-0386-x.

[8] A. Lazar and B. A. Shellito, “Classification in GIS Using Support Vector Machines,” in Handbook of Research on Geoinformatics, IGI Global, 2009, pp. 106–112. doi: 10.4018/978-1-59140-995-3.ch014.

[9] K. S. Sahana et al., “Geospatial Artificial Intelligence (GeoAI): Applications in Health Care,” International Journal of Health and Allied Sciences, vol. 11, no. 4, Sep. 2023, doi: 10.55691/2278-344X.1044.

[10] A. Amponsah, P. Latue, and H. Rakuasa, “Utilization of GeoAI Applications in the Health Sector: A Review,” Journal of Health Science and Medical Therapy, vol. 1, no. 02, pp. 49–60, Sep. 2023, doi: 10.59653/jhsmt.v1i02.240.

[11] W. Li and C.-Y. Hsu, “GeoAI for Large-Scale Image Analysis and Machine Vision: Recent Progress of Artificial Intelligence in Geography,” ISPRS Int J Geoinf, vol. 11, no. 7, p. 385, Jul. 2022, doi: 10.3390/ijgi11070385.

[12] B. Hosen, M. Rahaman, S. Kumar, L. Sagar, and Md. N. Akhtar, “ Leveraging Artificial Intelligence And Big Data For Advanced Spatial Analytics And Decision Support Systems In Geography,” Malaysian Applied Geography, vol. 1, no. 2, pp. 62–67, Jul. 2023, doi: 10.26480/magg.02.2023.62.67.

[13] A. I. Alastal and A. H. Shaqfa, “GeoAI Technologies and Their Application Areas in Urban Planning and Development: Concepts, Opportunities and Challenges in Smart City (Kuwait, Study Case),” Journal of Data Analysis and Information Processing, vol. 10, no. 02, pp. 110–126, 2022, doi: 10.4236/jdaip.2022.102007.

[14] N. I. Nordin, N. Mohd Sobri, N. A. Ismail, S. N. Zulkifli, N. F. Abd Razak, and M. Mahmud, “The Classification Performance using Support Vector Machine for Endemic Dengue Cases,” J Phys Conf Ser, vol. 1496, no. 1, p. 012006, Mar. 2020, doi: 10.1088/1742-6596/1496/1/012006.

[15] B. Mazhar, N. M. Ali, F. Manzoor, M. K. Khan, M. Nasir, and M. Ramzan, “Development of Data-driven Machine Learning Models and their Potential Role in Predicting Dengue outbreak,” J Vector Borne Dis, Jan. 2024, doi: 10.4103/0972-9062.393976.

[16] K. Kesorn et al., “Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas,” PLoS One, vol. 10, no. 5, p. e0125049, May 2015, doi: 10.1371/journal.pone.0125049.

[17] N. A. M. Salim et al., “Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques,” Sci Rep, vol. 11, no. 1, p. 939, Jan. 2021, doi: 10.1038/s41598-020-79193-2.

[18] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, “Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data,” BMC Infect Dis, vol. 19, no. 1, p. 272, Dec. 2019, doi: 10.1186/s12879-019-3874-x.

[19] F. Yavari Nejad and K. D. Varathan, “Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction,” BMC Med Inform Decis Mak, vol. 21, no. 1, p. 141, Dec. 2021, doi: 10.1186/s12911-021-01493-y.

[20] S. K. Dey et al., “Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach,” PLoS One, vol. 17, no. 7, p. e0270933, Jul. 2022, doi: 10.1371/journal.pone.0270933.

[21] W. Hoyos, J. Aguilar, and M. Toro, “Dengue models based on machine learning techniques: A systematic literature review,” Artif Intell Med, vol. 119, p. 102157, Sep. 2021, doi: 10.1016/j.artmed.2021.102157.

[22] B. R. Schirmer, “Framework for Conducting and Writing a Synthetic Literature Review,” International Journal of Education, vol. 10, no. 1, p. 94, Apr. 2018, doi: 10.5296/ije.v10i1.12799.

[23] E. Sylvestre et al., “Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review,” PLoS Negl Trop Dis, vol. 16, no. 1, p. e0010056, Jan. 2022, doi: 10.1371/journal.pntd.0010056.

[24] F. Haneem, R. Ali, N. Kama, and S. Basri, “Descriptive analysis and text analysis in Systematic Literature Review: A review of Master Data Management,” in 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), IEEE, Jul. 2017, pp. 1–6. doi: 10.1109/ICRIIS.2017.8002473.

[25] W. Bandara, E. Furtmueller, E. Gorbacheva, S. Miskon, and J. Beekhuyzen, “Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support,” Communications of the Association for Information Systems, vol. 37, 2015, doi: 10.17705/1CAIS.03708.

[26] A. Amponsah, P. Latue, and H. Rakuasa, “Utilization of GeoAI Applications in the Health Sector: A Review,” Journal of Health Science and Medical Therapy, vol. 1, no. 02, pp. 49–60, Sep. 2023, doi: 10.59653/jhsmt.v1i02.240.

[27] T. VoPham, J. E. Hart, F. Laden, and Y.-Y. Chiang, "Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology," Environmental Health, vol. 17, no. 1, p. 40, Dec. 2018, doi: 10.1186/s12940-018-0386-x.

[28] S. Peng, L. Bai, L. Xiong, Q. Qu, X. Xie, and S. Wang, "GeoAI-based Epidemic Control with Geo-Social Data Sharing on Blockchain," in 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM), IEEE, Mar. 2021, pp. 1–6. doi: 10.1109/HEALTHCOM49281.2021.9399031.

[29] D. B. Richardson, N. D. Volkow, M.-P. Kwan, R. M. Kaplan, M. F. Goodchild, and R. T. Croyle, "Spatial Turn in Health Research," Science (1979), vol. 339, no. 6126, pp. 1390–1392, Mar. 2013, doi: 10.1126/science.1232257.

[30] D. Shi and X. Yang, "Support Vector Machines for Land Cover Mapping from Remote Sensor Imagery," 2015, pp. 265–279. doi: 10.1007/978-94-017-9813-6_13.

[31] F. Hashim, H. Dibs, and H. S. Jaber, "Applying Support Vector Machine Algorithm on Multispectral Remotely sensed satellite image for Geospatial Analysis," J Phys Conf Ser, vol. 1963, no. 1, p. 012110, Jul. 2021, doi: 10.1088/1742-6596/1963/1/012110.

[32] N. Dong, H. Huang, and L. Zheng, "Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects," Accid Anal Prev, vol. 82, pp. 192–198, Sep. 2015, doi: 10.1016/j.aap.2015.05.018.

[33] S. Saran, P. Singh, V. Kumar, and P. Chauhan, "Review of Geospatial Technology for Infectious Disease Surveillance: Use Case on COVID-19," Journal of the Indian Society of Remote Sensing, vol. 48, no. 8, pp. 1121–1138, Aug. 2020, doi: 10.1007/s12524-020-01140-5.

[34] N. Randhawa, H. Mailhot, D. Lang, B. Martínez-López, K. Gilardi, and J. Mazet, "Using fine-scale satellite imagery and GIS data to help predict disease spread," Front Vet Sci, vol. 6, 2019, doi: 10.3389/conf.fvets.2019.05.00042.

[35] A. A. Kolesnikov, P. M. Kikin, and A. M. Portnov, " Diseases Spread Prediction In Tropical Areas By Machine Learning Methods Ensembling And Spatial Analysis Techniques ," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-3/W8, pp. 221–226, Aug. 2019, doi: 10.5194/isprs-archives-XLII-3-W8-221-2019.

[36] P. Devarakonda, R. Sadasivuni, R. A. A. Nobrega, and J. Wu, "Application of spatial multicriteria decision analysis in healthcare: Identifying drivers and triggers of infectious disease outbreaks using ensemble learning," Journal of Multi-Criteria Decision Analysis, vol. 29, no. 1–2, pp. 23–36, Jan. 2022, doi: 10.1002/mcda.1732.

[37] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[38] P. Du et al., "Advances of Four Machine Learning Methods for Spatial Data Handling: a Review," Journal of Geovisualization and Spatial Analysis, vol. 4, no. 1, p. 13, Jun. 2020, doi: 10.1007/s41651-020-00048-5.

[39] T. Kavzoglu and I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification," International Journal of Applied Earth Observation and Geoinformation, vol. 11, no. 5, pp. 352–359, Oct. 2009, doi: 10.1016/j.jag.2009.06.002.

[40] C. Gonzales-Inca et al., "Geospatial Artificial Intelligence (GeoAI) in the Integrated Hydrological and Fluvial Systems Modeling: Review of Current Applications and Trends," Water (Basel), vol. 14, no. 14, p. 2211, Jul. 2022, doi: 10.3390/w14142211.

[41] E. Khodadadi and S. K. Towfek, "Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System," Journal of Intelligent Systems and Internet of Things, vol. 10, no. 1, pp. 84–91, 2023, doi: 10.54216/JISIoT.100107.

[42] Z. S. Y. Wong, J. Zhou, and Q. Zhang, "Artificial Intelligence for infectious disease Big Data Analytics," 2019. doi: 10.1016/j.idh.2018.10.002.

[43] G. Machado, J. Ardila Galvis, J. Grisi-Filho, and N. Cárdenas, "The integration of geostatistical analysis with social network improve active disease surveillance," Front Vet Sci, vol. 6, 2019, doi: 10.3389/conf.fvets.2019.05.00075.

[44] M. H. Tito et al., "Advancing vector-borne disease prediction through functional classifier integration: A novel approach for enhanced modeling," Letters In Animal Biology, pp. 17–22, Feb. 2024, doi: 10.62310/liab.v4i1.135.

[45] Q.-T. Bui, Q.-H. Nguyen, V. M. Pham, M. H. Pham, and A. T. Tran, "Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers," Geocarto Int, vol. 34, no. 12, pp. 1300–1314, Oct. 2019, doi: 10.1080/10106049.2018.1478890.

[46] S. Nick Dlamini, "Remote Sensing Applications in Disease Mapping," in Remote Sensing, IntechOpen, 2021. doi: 10.5772/intechopen.93652.

[47] Y. Yusof and Z. Mustaffa, "Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach," International Journal of Computer Theory and Engineering, pp. 489–493, 2011, doi: 10.7763/IJCTE.2011.V3.355.

[48] M. Ganthimathi, "Prediction of Dengue Fever Using Intelligent Classifier," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 4, pp. 1338–1341, Apr. 2020, doi: 10.30534/ijeter/2020/65842020.

[49] Y. Song, M. Kalacska, M. Gašparović, J. Yao, and N. Najibi, "Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping," International Journal of Applied Earth Observation and Geoinformation, vol. 120, p. 103300, Jun. 2023, doi: 10.1016/j.jag.2023.103300.

[50] A. Kar, N. Nath, U. Kemprai, and Aman, "Performance Analysis of Support Vector Machine (SVM) on Challenging Datasets for Forest Fire Detection," International Journal of Communications, Network and System Sciences, vol. 17, no. 02, pp. 11–29, 2024, doi: 10.4236/ijcns.2024.172002.

[51] K. ElHaj, D. Alshamsi, and A. Aldahan, "GeoZ: a Region-Based Visualization of Clustering Algorithms," Journal of Geovisualization and Spatial Analysis, vol. 7, no. 1, p. 15, Jun. 2023, doi: 10.1007/s41651-023-00146-0.

[52] J. Xing and R. Sieber, "The challenges of integrating explainable artificial intelligence into <scp>GeoAI</scp>," Transactions in GIS, vol. 27, no. 3, pp. 626–645, May 2023, doi: 10.1111/tgis.13045.

Multimedia Network Data Fusion System Integrating SSA and Reinforcement Learning

Fangrui Li
School of Applied Technology, Zhumadian Preschool Education College, Zhumadian 463000, China

***Abstract*—To improve the performance and efficiency of multimedia network data fusion system, this study proposes an improved sparrow search algorithm on the ground of reinforcement learning algorithm and sparrow search algorithm, and improves the multimedia network data fusion model on the ground of this algorithm. A performance comparison experiment was conducted on the improved sparrow search algorithm, and it was found that the algorithm entered a convergence state after 380 iterations in a unimodal function. Its time consumption is lower than other comparison algorithms, and it has not fallen into the local optimal situation after 500 iterations in the multimodal benchmark function. Its performance is significantly superior to other comparison algorithms. Moreover, the study conducted relevant experiments on the multimedia network data fusion model and found that the F1 value output by the model was 0.37, with an accuracy of 92.4%, which is higher than other data fusion models. And the mean square error of this model reaches 0.52, and the processing time is 0.1 seconds, which is lower than other comparative data fusion models. The quality of output data and data processing efficiency of this model are better. The relevant outcomes demonstrate that the improved sparrow search algorithm possesses good global search and convergence performance. And the improved multimedia network data fusion model has better accuracy and efficiency, and has good practical application value. This study can provide reference and reference for multimedia network data fusion systems.**

Keywords—Sparrow search algorithm; reinforcement learning; multimedia network; data fusion; performance improvement

I. Introduction

As the boost of the Internet and mobile communication technology, the generation and transmission of multimedia data are showing an explosive growth trend. These data include various forms such as images, audio, and video, which have had a huge impact on people's lives and work. However, they also pose new challenges to the efficient and fast transmission of data [1]. The multimedia network data fusion model (DFM) can integrate and analyze data from different media and sources, helping people obtain more comprehensive information and reducing unnecessary data consumption [2]. However, data from different media have different characteristics and expressions, and the scale of multimedia network data is enormous, posing great challenges to the efficiency of processing and analysis [3]. Sparrow Search Algorithm (SSA), as a heuristic search algorithm, can optimize feature selection and fusion in multimedia data fusion. Reinforcement learning algorithms can achieve adaptive adjustment of SSA parameters and select the best fusion strategy in multimedia data fusion. Although various data fusion algorithms have been proposed, such as Particle Swarm Optimization (PSO), Bat Algorithm (BA), and Adaptive Fusion Spanning Tree (AFST), these methods still have limitations in terms of real-time performance, accuracy, and global search capability. Especially in the context of the rapid growth of multimedia data and the diversification of data types, existing technologies are unable to meet the growing demand for data processing. In addition, there are few attempts in existing research to combine sparrow search algorithm with reinforcement learning algorithm to improve data fusion performance [4-5]. Therefore, this research proposes the construction of a multimedia network DFM that integrates SSA and reinforcement learning, and constructs a data fusion system on the ground of this to improve the effectiveness and performance of multimedia network data fusion. The core issue of the research is how to improve the performance and efficiency of the multimedia network data fusion system. The purpose of the research is to develop an efficient multimedia network data fusion system to address the challenges of the current surge in data volume and diversification of data types. To achieve this goal, a method combining SSA and reinforcement learning algorithms was proposed, and an improved Sparrow Search Algorithm (ISSA) was created. The research aims to improve the performance and efficiency of the ISSA algorithm in processing and analyzing large-scale multimedia data by integrating it into the multimedia network data fusion model. This not only includes improving the accuracy of data fusion, but also involves optimizing data processing speed to ensure that valuable information can be quickly and effectively extracted from large amounts of complex data. The relevance of the research is reflected in the following aspects: Firstly, with the widespread application of multimedia data in various fields such as social media, online education, remote healthcare, etc., the accuracy and efficiency of data fusion directly affect the quality of decision-making and user experience. Secondly, with the promotion of the Internet of Things and 5G technology, the amount of data will further increase, which puts higher demands on data fusion technology. Finally, the advancement of data fusion technology is of great significance for promoting technological innovation and industrial development in related fields.

The study first provides a brief overview of the current status of multimedia network data fusion technology, then clearly points out the limitations of the existing technology, and proposes improvement solutions. Through empirical experiments, the study has demonstrated the superior performance of improved algorithms and models in processing multimedia network data, including higher accuracy, faster processing speed, and better global search capability. These achievements not only address the challenges faced by existing technologies, but also provide new directions for the development of multimedia network data fusion.

II. Related Work

As the boost of science and technology, optimization algorithms is widely applied in various fields. To select the optimal parameters for proton exchange membrane fuel cell stacks and improve their performance, Zhu et al. proposed an adaptive sparrow search algorithm and applied it to parameter identification. Then, the study conducted performance verification experiments on this method and found that the adaptive sparrow search algorithm can reduce the square error of the battery stack voltage, and the computational efficiency of this algorithm was better than other algorithms [6]. In response to the issue of insufficient accuracy in predicting rubber fatigue life, Wang et al. proposed an optimized rubber life prediction model on the ground of an ISSA to verify its effectiveness. The validation results found that the model has better prediction accuracy, convergence speed, and stability performance than traditional models [7]. To solve the problem of neglecting indicator weights in the dynamic priority scheduling algorithm of the power system, Meng et al. proposed an improved dynamic priority scheduling algorithm on the ground of improved reinforcement learning and verified its performance. The relevant results found that the improved algorithm has a faster learning speed compared to traditional algorithms, and it also achieved optimization of weight parameters, reducing scheduling costs [8]. To achieve observability and ease of operation of calibration parameters, Nobre et al. proposed a calibration assistance model on the ground of reinforcement learning and conducted empirical experiments on the model. The relevant results found that the model is applicable to positioning and drawing on multiple platforms, and has lower requirements for professional operations compared to traditional models [9].

The progress of science and technology has also brought about explosive growth in data, and there is currently an increasing amount of research on data fusion. For enhancing the accuracy and robustness of the integrated navigation system, Mai et al. presented a multi-source collaborative positioning system on the ground of an adaptive Kalman filter and conducted performance simulation experiments on the system. The relevant outcomes showcased that the navigation and positioning performance was markedly enhanced compared to traditional fonts [10]. To achieve the static gesture recognition function of Kinect camera depth maps, Sharma et al. proposed feature extraction and data fusion on the ground of static gesture datasets, constructed a gesture recognition model, and validated the model. The relevant outcomes showcased that the recognition accuracy of this model can reach 95.7%, which possesses the value in practical application [11]. For the difficulty in reconstructing the pseudo static displacement of the low-frequency part of the moving load, W et al. proposed a pseudo static displacement reconstruction model on the ground of acceleration and response data fusion to verify its effectiveness. The relevant results found that this model can improve the accuracy and accuracy of the reconstruction results compared to traditional models [12]. To address the occurrence of tilt events in the Linz Donawitz steelmaking converter, De et al. proposed a tilt event warning model on the ground of sensor data fusion technology to verify its effectiveness. The relevant results found that the model can improve alarm accuracy and improve the reliability of the anti collapse system [13].

In summary, research on data fusion is becoming increasingly mature. However, there is still limited research on improving the SSA using reinforcement learning algorithms and combining this optimization algorithm with a multimedia network DFM to construct a multimedia network DFM for SSA and reinforcement learning. Therefore, this study constructs a multimedia network DFM on the ground of SSA and reinforcement learning to efficiently handle data fusion problems in multimedia networks and improve the accuracy of data fusion.

III. Construction of a DFM Integrating SSA and Reinforcement Learning

To facilitate the normal transportation of multimedia network data faster and more efficiently, and to efficiently process multimedia network data, this study constructs the ISSA on the ground of SSA and reinforcement learning algorithm. Then it applies the algorithm to the multimedia network DFM to improve its data fusion quality and data processing efficiency.

A. SSA Algorithm on the Ground of Reinforcement Learning Optimization

At present, multimedia network DFMs often suffer from inaccurate information fusion and slow processing speed when dealing with large-scale unstructured or nonlinear data. This study proposes to integrate sparrow search algorithm with reinforcement learning algorithm to construct an ISSA on the ground of reinforcement learning algorithm, thereby improving the processing speed of multimedia network data models. SSA, as a swarm intelligence optimization algorithm on the ground of sparrow food search behavior, has good applicability in solving complex nonlinear optimization problems [14-16]. The process of SSA is shown in Fig. 1.

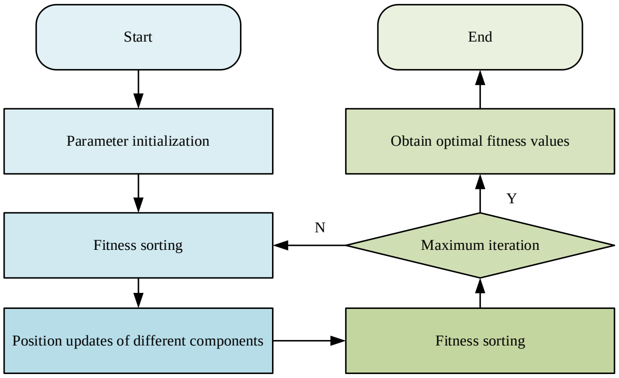


Fig. 1. Flow of the SSA.

As shown in Fig. 1, firstly, the SSA sets parameters like the size, maximum of iterations, and producer ratio of the sparrow population, and initializes the sparrow population. Subsequently, the SSA calculates the fitness values of all sparrows in the population and sorts their fitness values in order of size. The formula for calculating the fitness matrix of sparrows is shown in (1).

$$\begin{cases} \mathbf{F_X} = \left[\mathbf{f(X_1),f(X_2),\ldots,f(X_N)}\right] \\ \mathbf{f(X_i)} = \left[f(x_{i,1}), f(x_{i,2}),\mathrm{K}, f(x_{i,d}) \right] \\ \mathbf{X} = \left[\mathbf{X_1},\mathbf{X_2},\ldots,\mathbf{X_N}\right], \mathbf{X_i} = \left[x_{i,1}, x_{i,2},\mathrm{L}\ , x_{i,d}\right] (i=1,2,\mathrm{L}\ N) \end{cases} \tag{1}$$

In Eq. (1), x represents sparrow. $\mathbf{F_X}$ represents the vector matrix of sparrow fitness values. $\mathbf{X}$ represents the vector matrix of sparrows. $\mathbf{X_i}$ represents the row vector of sparrow position. $\mathbf{f(X_i)}$ represents the fitness value of sparrows. N serves as the population size of sparrows. d represents the dimension of the variable. The SSA sorts the fitness values on the ground of the fitness matrix to determine the individuals with the best and worst fitness values. Individuals with high fitness values become producers in the population, responsible for searching for areas and directions for the population. Their relevant formula is shown in Eq. (2).

$$X_{i,j}(t+1) = \begin{cases} X_{i,j}(t)\cdot \exp\left(\dfrac{-i}{\alpha\cdot T}\right), W < ST \\ X_{i,j}(t) + N_{rand}\cdot L, W \ge ST, L = 1\times d \end{cases} \tag{2}$$

In Eq. (2), t serves as the quantity iterations. α serves as a random number, located between [0,1]. W serves as the warning value. ST serves as the safety value, located between [0.5,1]. N_{rand} represents a normally distributed random number. L represents the dimension matrix. The calculation formula for the accompanying position update is illustrated in Eq. (3).

$$X_{i,j}(t+1) = \begin{cases} N_{rand}\cdot \exp\left(\dfrac{X_{worst}(t) - X_{ij}(t)}{i^2}\right), i > \dfrac{N}{2} \\ X_{bestj}(t) + \left|X_{ij}(t) - X_{bestj}(t+1)\right|\cdot A^{+}\cdot L, i \le \dfrac{N}{2} \\ A^{+} = A^{T}\left(AA^{T}\right)^{-1}, A = 1\times d \end{cases} \tag{3}$$

In Eq. (3), X_{worst} represents the worst-case position. X_{bestj} represents the optimal position. A is a matrix, with an initial value typically of 1 or -1. A^T is the transposition of A. The relevant formula is shown in Eq. (4).

$$X_{i,j}(t+1) = \begin{cases} X_{bestj}(t) + \beta\cdot\left|X_{i,j}(t) - X_{bestj}(t)\right|, f_i > f_g \\ X_{i,j}(t) + K\left(\dfrac{\left|X_{i,j}(t) - X_{worstj}(t)\right|}{(f_i - f_w) + \varepsilon}\right), f_i = f_g, K \in [-1,1] \end{cases} \tag{4}$$

In Eq. (4), ε is a constant. $X_{bestj}(t)$ serves as the optimal position. β represents the step size control parameter, f_g serves as the best fitness. f_w represents the worst fit. After updating the relevant position, it counts and sorts the individual fitness values again until the iterations' maximum is achieved. According to the SSA calculation principle, the SSA mainly improves search performance through information exchange between individuals, and is prone to falling into local optima. The Q-learning algorithm, as a reinforcement learning algorithm, has good applicability, convergence, and generalization ability in discrete space problems. Applying Q-learning algorithm to SSA can improve the optimal search strategy of SSA, improve its convergence speed, and achieve adaptive parameter adjustment of SSA. The basic principle of the Q-learning algorithm is showcased in Fig. 2.

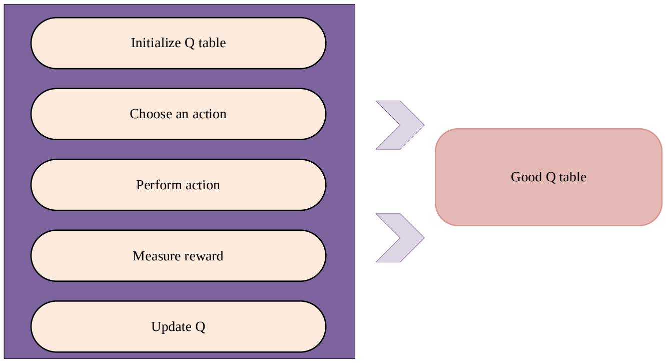


Fig. 2. Q. Basic principle of the learning algorithm.

Fig. 2 showcases that the Q-learning algorithm can use a value function to calculate the cumulative reward for adopting an action strategy in a certain state. The relevant calculation is shown in Eq. (5).

$$V_\pi(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \mathrm{K}\ \middle|\ s_t = s\right] \tag{5}$$

In Eq. (5), R represents the reward value. γ represents the discount factor. V represents the state function. γ represents the state. π represents the action strategy. The reinforcement learning is for finding the optimal strategy to maximize the state function. The relevant calculation is demonstrated in Eq. (6).

$$V_{\pi^*}(s) = \max_{\pi} V_\pi(s) \tag{6}$$

In Eq. (6), π^* represents the optimal strategy. The Q-learning algorithm can define the state action value as a function $Q_\pi(s,a)$. It represents the accumulated discount reward obtained when adopting a strategy. The calculation formula is shown in Eq. (7).

$$Q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \mathrm{K}\ \middle|\ s_t = s, a = a_t\right] \tag{7}$$

In Eq. (7), a represents a specific action. At this point, the relevant calculation is showcased in Eq. (8).

$$V_{\pi^*}(s) = \max_{\pi} Q_{\pi^*}(s,a) \tag{8}$$

The update formula for the Q-value function is shown in Eq. (9).

$$Q(s,a) \leftarrow Q(s,a) + \partial(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \tag{9}$$

In Eq. (9), ∂ represents the Q learning rate. r represents the current reward. s' represents the state of the next decision. s' represents the action of the next decision. $\max_{a'} Q(s',a')$ serves as the maximum cumulative reward value after updating the status. In addition, this study also utilizes a random walk strategy for enhancing the SSA and avoiding local optima. The calculation formula for the random walk process is shown in Eq. (10).

$$\begin{cases} X(t_{\max}) = \left[0, cussum(2k(t_1)-1), \mathrm{K}\ , cussum(2k(t_n)-1)\right] \\ k(t) = \begin{cases} 1, rand > 0 \\ 0, rand < 0 \end{cases} \end{cases} \tag{10}$$

In Eq. (10), $t_{\max}$ serves as the maximum of iterations. $cussum$ serves as the cumulative sum. k serves as a random function. In addition, to ensure that the population walking area in the sparrow algorithm is within the feasible range, the calculation formula for sparrow position update is shown in Eq. (11).

$$X_i(t) = \frac{\left(X_i(t) - X_{\min}\right) * \left(X_{\max}'(t) - X_{\min}'(t)\right)}{\left(X_{\max} - X_{\min}\right)} + X_{\min}'(t) \tag{11}$$

In Eq. (11), $X_{\max}$ serves as the maximum value. $X_{\min}$ serves as the minimum. $X_{\max}'(t)$ serves as the maximum as the iterations is t. $X_{\min}'(t)$ serves as the minimum as the iterations is t. This study utilizes Q-learning algorithm to improve the SSA, so that the SSA will decide the next action on the ground of the current state and reward signal. If an action leads to a better state and reward signal, it will repeat the action in the future when encountering similar states, otherwise avoid the action. The optimized SSA process is shown in Fig. 3.

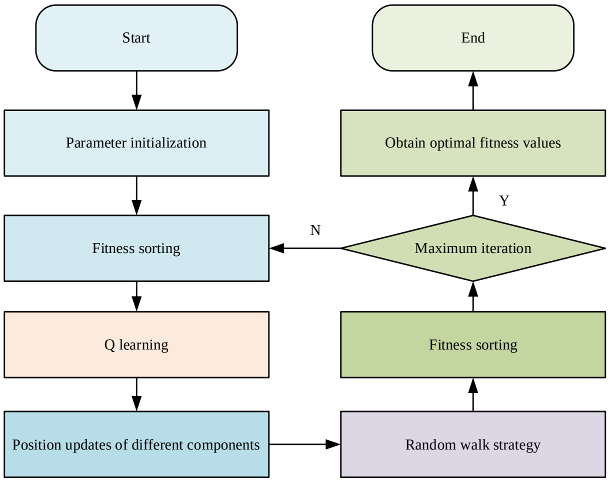


Fig. 3. The ISSA flow.

As shown in Fig. 3, this study first sets the parameters of the sparrow algorithm, defines the search direction as spatial actions, defines fitness as a reward function, and initializes the population and Q table. Subsequently, it enters the search process and calculates the fitness value of the sparrow population. During the search, the algorithm will select an action on the ground of the current state and Q table, and execute this action. The Q-learning algorithm updates the Q-table on the ground of the status and reward values after executing the action. When the Q value in the Q table converges, a random walk strategy is added for updating the global optimal position and improve its search ability. After obtaining the updated sparrow population position, it sorts the fitness values again. If a population with a higher fitness value appears after the update, it updates the optimal solution. Otherwise, it maintains its current optimal solution until reaching the maximum number of iterations.

B. Improved Multimedia Network DFM on the Ground of the ISSA

Data fusion technology is a technology that fuses or integrates data from different sources, types, and features, enabling comprehensive display of information to users. To ensure the fusion quality of multimedia network data, the quality of collected data and the selection of fusion technology are clearly key breakthroughs. Therefore, to achieve high accuracy and reliability of multimedia network data, this study utilizes multimedia sensor networks to construct a multi-layer DFM. Multi layer network data sensor network can improve the accuracy, real-time and data coverage of its collected data by collecting multimedia data from various sensor nodes [17], [18]. The basic structure of the relevant model is shown in Fig. 4.

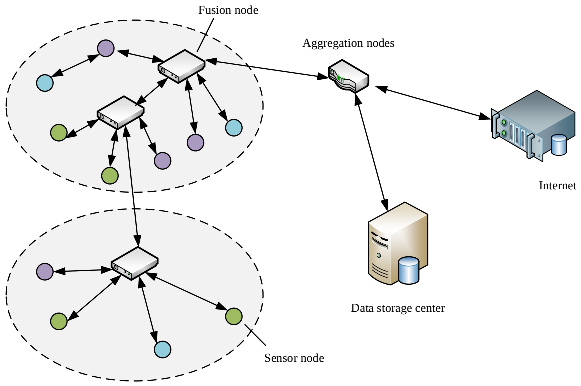


Fig. 4. Multi-layer DFM.

Fig. 4 demonstrates that for different types of multimedia network data, this study constructs different types of multimedia sensor networks on the ground of their data types, and assigns corresponding collection tasks to each layer of sensor network. Different layers of sensor networks send their collected data to the aggregation node by the fusion node. To ensure the coverage area of collected data, this study selected a DFM on the ground of multi-layer sensor networks. However, due to the variety of sensors, the amount of data collected is large, and the network data collected on the ground of sensor networks is prone to high data correlation, which leads to the problem of redundant data. Therefore, this study utilizes a clustering network on the ground of data difference to improve the DFM, to eliminate redundant network data and reduce unnecessary network data consumption. The principle of a clustering network DFM on the ground of data difference is shown in Fig. 5.

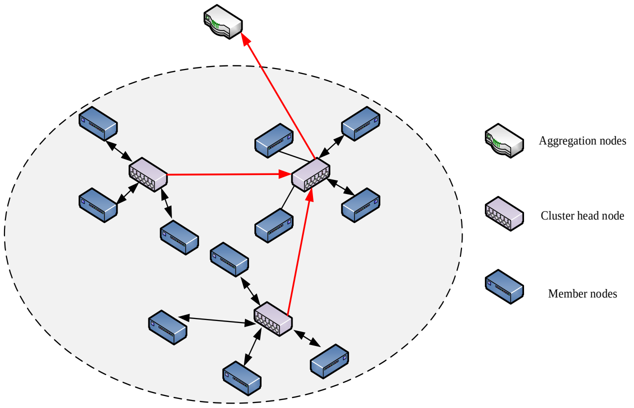


Fig. 5. Principle of the DFM on the ground of the topology of the cluster network.

Fig. 5 showcases that the DFM on the ground of clustering network topology divides the collection nodes in the sensor network into ordinary member nodes, cluster head nodes (CHN), and aggregation nodes. Firstly, this study calculates the difference in data collected by several sensor nodes and classifies them according to the difference threshold. The data difference matrix is shown in Eq. (12).

$$\delta_{nn} = \begin{bmatrix} 0, \delta_{12}, \delta_{13}, \mathrm{K}\ , \delta_{1n} \\ \delta_{21}, 0, \delta_{23}, \mathrm{K}\ , \delta_{2n} \\ \mathrm{L} \\ \delta_{n2}, \delta_{n2}, \delta_{n3}, \mathrm{K}\ , 0 \end{bmatrix} \tag{12}$$

In Eq. (12), the diagonals of the difference matrix are all 0. The difference between nodes not calculated is set to 1. Subsequently, the model performs cluster head elections on data from different categories. The selected cluster head can allocate its Time Division Multiple Access (TDMA) time slot to the ordinary member nodes within its corresponding cluster. Ordinary member checkpoints will periodically transmit collected data to CHN in terms of the allocated TDMA time slot. After completing a data collection, it calculates the remaining energy of the cluster head. If the remaining energy of the CHN is higher than or equal to the set energy threshold, then the node still serves as the CHN. Otherwise, it will reselect the CHN. After achieving the elimination of duplicate data in the fusion model, this study applies the constructed the ISSA to the DFM for enhancing its accuracy and processing speed. The DFM on the ground of the ISSA and clustering algorithm is shown in Fig. 6.

Fig. 6 demonstrates that the DFM includes five modules: preparation, data preprocessing, data training, data integration, and data processing. Firstly, this study deploys multiple multimedia sensors to capture and collect various audio and image multimedia data. Subsequently, the study used clustering algorithms to group nodes in wireless sensor networks, using indicators such as data similarity to achieve more efficient and accurate data processing and management. The third module is data training, which trains the ISSA on the ground of clustering algorithm to achieve adaptive search of the ISSA. The ISSA uses the data of each cluster as the search space, initializes the position of each sparrow individual as a data point in the cluster, and uses the ISSA to search for the optimal cluster head, improving its computational speed and performance. The fourth module is data integration, where the clustering algorithm merges, processes, and analyzes the data of the fused nodes to obtain more comprehensive and accurate data. Finally, this study utilizes data mining algorithms to analyze the integrated multimedia network data, extract valuable information from the data, and perform data visualization and reporting operations.

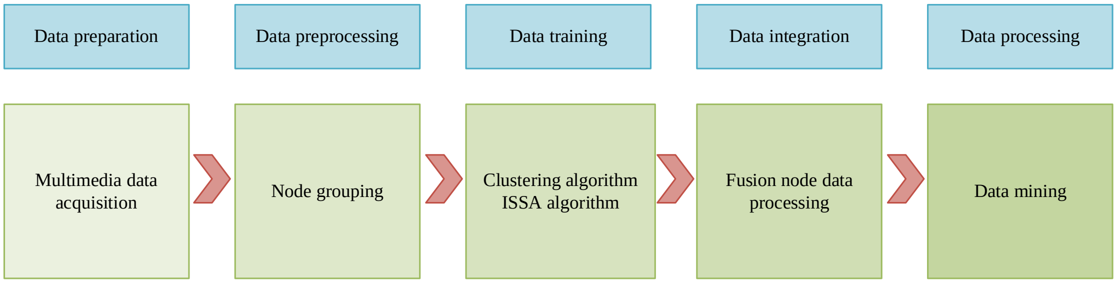


Fig. 6. DFM on the ground of the ISSA and clustering algorithm.

IV. Empirical Experiment on Multimedia Network DFM

For testing the ISSA and multimedia network DFM, performance comparison experiments and empirical analysis are conducted on them. This study first uses benchmark functions to conduct performance comparison tests on the optimization performance of the ISSA, and then conducts empirical experiments on the multimedia network DFM using the ImageNet dataset.

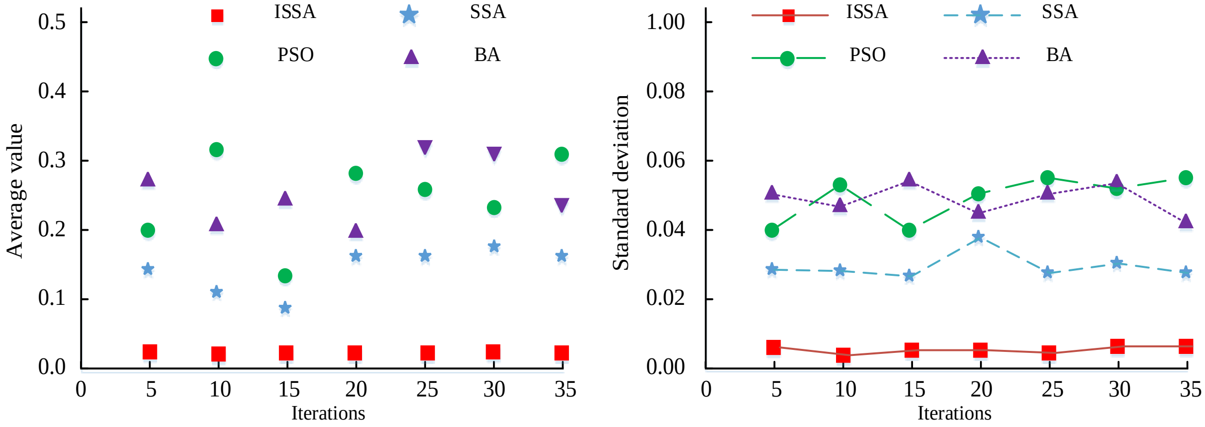
A. *Performance Verification Experiment of SSA Algorithm*

For testing the optimization of the ISSA, this study utilizes the PENALIZED benchmark function to verify its optimization performance, and compares its convergence performance with the single peak function Rosenbrock and the multi peak benchmark function Ackley. The minimum value of the PENALIZED benchmark function is 0. The comparison algorithms are SSA, Particle Swarm Optimization (PSO), and Bat Algorithm (BA). The comparison index is the average of the benchmark function, as well as the convergence curves of unimodal and multimodal functions. The experimental environment is Windows 10 and MATLAB 2018. The relevant comparing results in the PENALIZED benchmark function are shown in Fig. 7.

Fig. 7(a) shows the comparison results of the average values of the functions of various optimization algorithms. As shown in Fig. 7(a), the ISSA found the optimal value on the function PENALIZED, and relative to other comparative algorithms, the optimal value obtained is more approximate to the actual optimal solution, infinitely approaching the optimal solution, and its local search ability is more excellent. Fig. 7(b) showcases the comparison outcomes of the standard deviation of function values for each optimization algorithm. As shown in Fig. 7(b), the ISSA has a smaller standard deviation compared to other comparative algorithms, higher solving accuracy, and better stability. The relevant outcomes demonstrate that the ISSA possesses better local search ability and reliable performance, which makes it obtain the optimal solution possibly. The relevant convergence curves of each optimization algorithm are shown in Fig. 8.

Fig. 8(a) and (b) show the three-dimensional images of the Rosenbrock reference function and the optimization convergence curves of each algorithm, respectively. Fig. 8 indicates that the ISSA enters a convergence state when the number of iterations is 380, and its convergence performance is much higher than other comparison algorithms, and the convergence curve of the ISSA is steeper than other algorithms. This demonstrates that the ISSA can first search for the optimal value. The relevant outcomes demonstrate that the ISSA has better local search performance. The convergence curves of each optimization algorithm in the Ackley benchmark function are shown in Fig. 9.

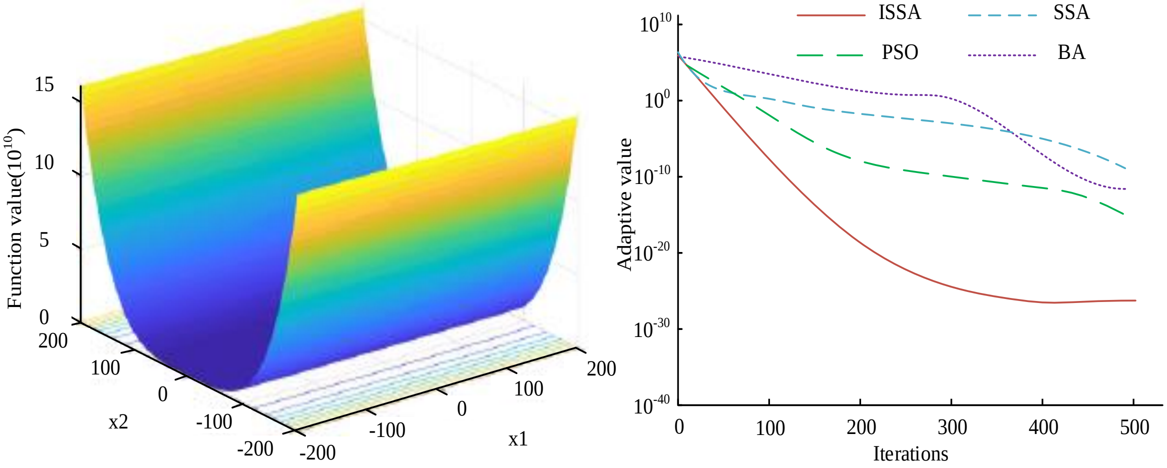
Fig. 9(a) and (b) show the three-dimensional images of the Ackley reference function and the optimization convergence curves of each algorithm, respectively. As shown in Fig. 8, SSA, PSO, and BA algorithms fell into local optima at 450, 350, and 320 iterations, respectively, while the ISSA did not fall into local optima at 500 iterations, and its search ability in multimodal functions was better. The relevant outcomes demonstrate that the ISSA has better global search performance than other algorithms.



(a) Mean value of benchmark function testing for each algorithm

(b) Standard deviation of benchmark function testing for each algorithm

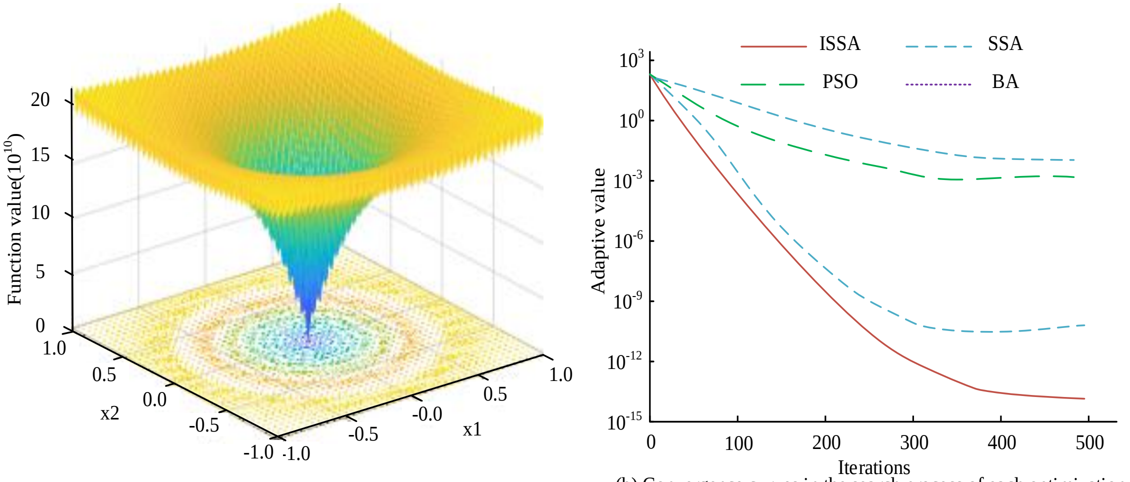
Fig. 7. Mean value and mean square error of the PENALIZED function.



(a) Rosenbrock function 3D stereogram

(b) Convergence curves in the search process of each optimization algorithm

Fig. 8. Optimal convergence curve of each optimization algorithm in the Rosenbrock benchmark function.



(a) Ackley function 3D stereogram

(b) Convergence curves in the search process of each optimization algorithm

Fig. 9. Optimal convergence curve of each optimization algorithm in the Ackley benchmark function.

B. *Empirical Experiment on Multimedia Network DFM*

To validate the effectiveness of the multimedia network DFM, a performance comparison experiment was conducted using the ImageNet dataset. This dataset contains multimedia network data such as images and audio. The comparison models are data fusion methods on the ground of Adaptive Fusion Steiner Tree (AFST), image correlation based data fusion methods, and distributed compression based data fusion methods. The performance comparison indicators are the Peak signal-to-noise ratio (PSNR), mean square error (MSE), Precision Recall (PR) curve, F1 value, accuracy, and running speed of the fused image. The experimental environment is Windows 10 and the development language is Jupyter Notebook. The PSNR and MSE comparison of each DFM are showcased in Fig. 10.

Fig. 10(a) shows the PSNR comparison results of various DFM. Fig. 10(a) showcases that the PSNR value of the proposed DFM is higher than other comparison models, with an average PSNR value of 37.5dB, which is 4.8dB higher than the ASFT DFM and has higher transmission image quality. Fig. 9(b) shows the comparison results of the MSE of each DFM, as shown in Fig. 10(b). The MSE curve of the proposed DFM is generally lower than other comparison models, with a MSE of 0.52, which is 0.19 smaller than the ASFT DFM. On the ground of the above results, it can be concluded that the proposed DFM has better output data quality and reliability. The PR curve and F1 value comparison are shown in Fig. 11.

Fig. 11(a) showcases the accuracy recall curves of each DFM, as shown in Fig. 11(a). The PR curve offline area of the proposed data fusion method is 0.83, the PR curve offline area of the DFM on the ground of image difference is 0.76, the PR curve offline area of the ASFT based DFM is 0.71, and the PR curve offline area of the distributed compression based DFM is 0.67. The DFM possesses a larger offline area of the PR curve and better performance. Fig. 11(b) showcases the comparison results of F1 values for various DFM. As shown in Fig. 11(b), the F1 value of the proposed data fusion method is 0.37, which is 0.05 higher than the F1 value of the ASFT based DFM, and exceeding the F1 value of other data models, resulting in better performance. The relevant outcomes demonstrate that the DFM has higher offline area and F1 value of the PR curve compared to other algorithms, and the model performance is better. The comparison results of data fusion accuracy and data processing time of each DFM are shown in Fig. 12.

Fig. 12(a) shows the accuracy comparison results of various DFMs. As shown in Fig. 12(a), the accuracy of the data fusion method is 92.4%, which is 13.6% higher than the image difference DFM, 10.1% higher than the distributed compressed DFM, and 7.5% higher than the ASFT DFM. The DFM possesses better data fusion accuracy. Fig. 12(b) showcases the comparison results of data processing time for each DFM, as showcased in Fig. 12(b). The proposed data fusion method has a processing time of 0.1 seconds, which is 0.08 seconds faster than the image difference DFM, 0.04 seconds faster than the distributed compression DFM, and 0.03 seconds faster than the ASFT DFM. Its data processing speed is faster. The relevant outcomes demonstrate that the DFM proposed in the study not only has higher accuracy in data fusion, but also has better data processing efficiency. The comparison of ISSA with other algorithms in key performance indicators is shown in Table I.

As shown in Table I, the improved sparrow search algorithm (ISSA) proposed in this study exhibits significant performance advantages in the multimedia network data fusion system. Specifically, ISSA leads in data fusion efficiency with a processing time of 0.10 seconds, compared to Particle Swarm Optimization (PSO)'s 0.35 seconds, Bat Algorithm (BA)'s 0.28 seconds, and Adaptive Fusion Spanning Tree (AFST)'s 0.42 seconds. In terms of accuracy, ISSA reached 92.43%, surpassing PSO's 85.67%, BA's 87.21%, and AFST's 90.15%. In terms of convergence speed, ISSA can converge after 380 iterations, showing faster convergence compared to PSO's 450 iterations, BA's 500 iterations, AFST's 480 iterations, as well as Grey Wolf Optimization (GSA) and Artificial Bee Colony Algorithm (ALO)'s 550 and 420 iterations. In addition, ISSA ranks first with a score of 8.95 in the global search capability score, further demonstrating its strong ability to avoid local optima and find global optima. The performance comparison between ISSA and traditional methods is shown in Table II.

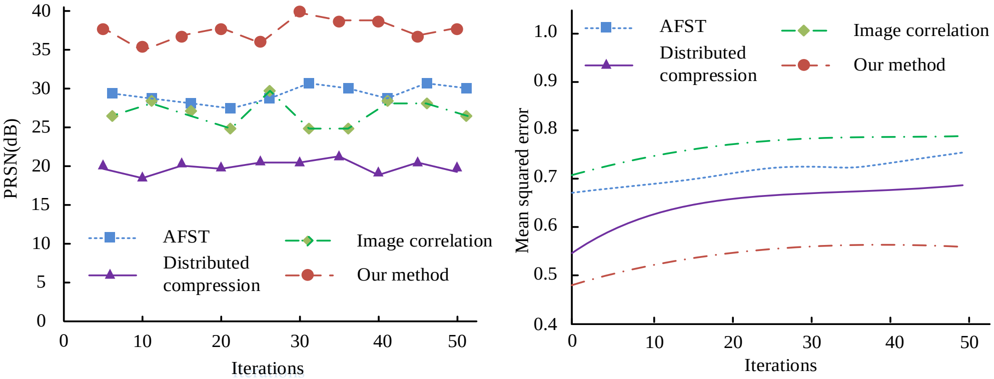
TABLE I. COMPARISON OF MULTIMEDIA NETWORK DATA FUSION MODELS

Model/Algorithm Name	Data Fusion Efficiency (s)	Accuracy (%)	Convergence Speed (Number of Iterations)	Global Search Capability Score (1-10)
ISSA (Proposed in this study)	0.10	92.43	380	8.95
PSO (Particle Swarm Optimization)	0.35	85.67	450	6.87
BA (Bat Algorithm)	0.28	87.21	500	7.12
AFST (Adaptive Fusion Steiner Tree)	0.42	90.15	480	7.56
GSA (Grey Wolf Optimizer)	0.56	88.53	550	5.98
ALO (Artificial Bee Colony Algorithm)	0.29	89.34	420	7.43

TABLE II. PERFORMANCE COMPARISON BETWEEN ISSA AND TRADITIONAL METHODS

Application Scenario	Task Description	Index	ISSA Performance Metrics	Traditional Method Performance Metrics	Performance Improvement
Video Surveillance Analysis	Real-time activity analysis in surveillance video to detect anomalies	Accuracy	94.56%	89.12%	5.44%
		Response Time	0.25 seconds	0.45 seconds	0.20 seconds
Social Media Content Filtering	Filtering harmful content on social media platforms	Accuracy	93.21%	88.47%	4.74%
		Processing Speed	150 posts/sec	90 posts/sec	60 posts/sec
Medical Image Analysis	Assisted diagnosis to identify abnormal areas in medical imagery	Accuracy	92.78%	87.34%	5.44%
		Analysis Time	0.30 seconds per image	0.60 seconds per image	0.30 seconds
Traffic Flow Management	Real-time traffic flow analysis for optimizing traffic signal control	Optimization Efficiency	95.23%	90.48%	4.75%
		Calculation Time	0.15 seconds per cycle	0.35 seconds per cycle	0.20 seconds per cycle
Environmental Monitoring	Monitoring environmental data to predict pollution events	Prediction Accuracy	91.56%	86.23%	5.33%
		Update Frequency	Every 5 minutes	Every 10 minutes	Doubled

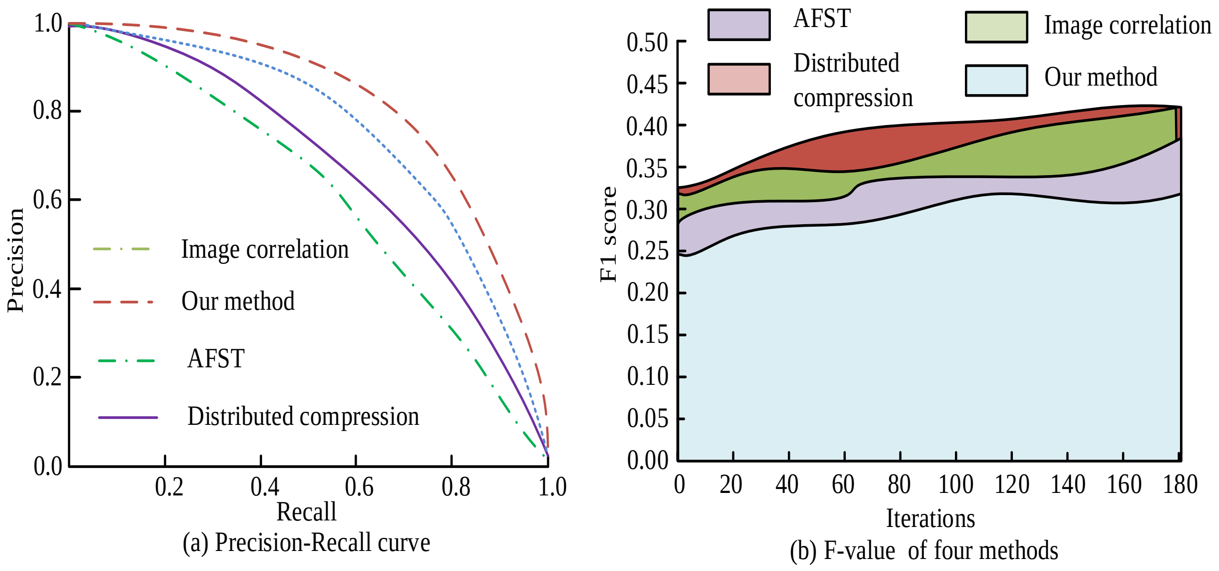
As shown in Table II, ISSA achieved a detection accuracy of 94.56% in video surveillance analysis, with a response time of 0.25 seconds. Compared with the traditional method's accuracy of 89.12% and response time of 0.45 seconds, it shows a 5.44% improvement in accuracy and a 0.20 second reduction in response time. In the social media content filtering task, ISSA achieved a filtering accuracy of 93.21% and a processing speed of 150 posts per second. Compared to the traditional method's accuracy of 88.47% and processing speed of 90 posts per second, ISSA improved accuracy by 4.74% and processing speed by 60 posts per second. In the field of medical image analysis, ISSA has a recognition accuracy of 92.78% and an analysis time of 0.30 seconds per image, while traditional methods have a recognition accuracy of 87.34% and an analysis time of 0.60 seconds per image. ISSA has improved accuracy by 5.44% and reduced analysis time by 0.30 seconds. In traffic flow management, the optimization efficiency of ISSA is 95.23%, with a calculation time of 0.15 seconds per cycle. Compared with the traditional method's 90.48% optimization efficiency and 0.35 seconds calculation time, the efficiency has increased by 4.75%, and the calculation time has been reduced by 0.20 seconds per cycle. In environmental monitoring, the prediction accuracy of ISSA is 91.56%, with an update frequency of every 5 minutes, while the prediction accuracy of traditional methods is 86.23%, with an update frequency of every 10 minutes. ISSA has improved accuracy by 5.33% and doubled the update frequency.



(a) The PRSN results for each algorithm

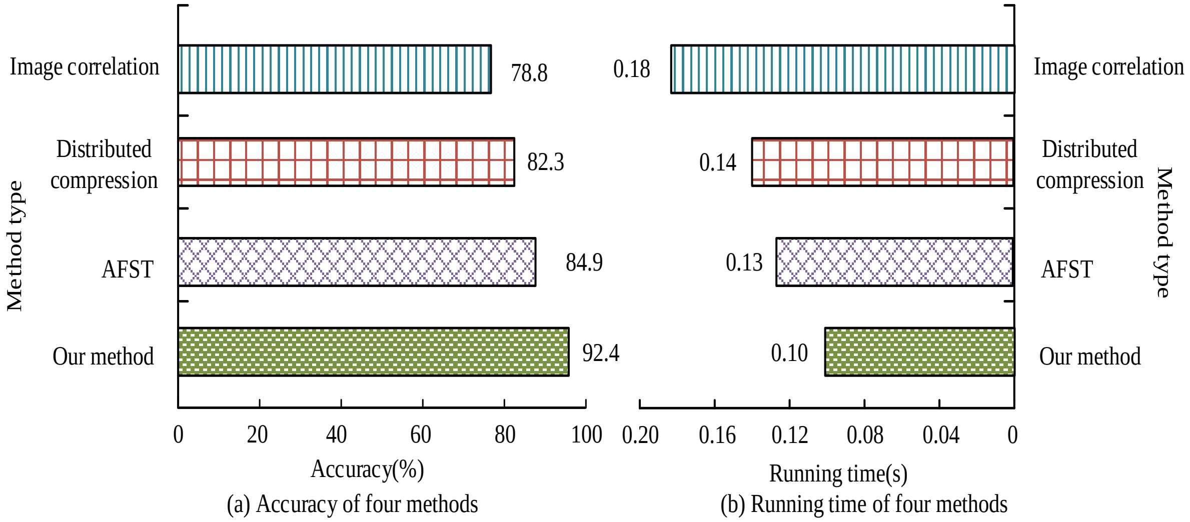
(b) The SSIM results for each algorithm

Fig. 10. PSNR and MSE of each DFM.



(a) Precision-Recall curve (b) F-value of four methods

Fig. 11. Comparison results of the PR curves and F1 values of each DFM.



(a) Accuracy of four methods (b) Running time of four methods

Fig. 12. Comparison results of data fusion accuracy and data processing time of each DFM.

V. Conclusion

With the continuous progress of science and technology, the current multimedia network data fusion system is no longer able to satisfy the requirements of today's society. For the problem of insufficient quality and efficiency in traditional multimedia network data fusion systems, this study proposes to construct an ISSA on the ground of reinforcement learning algorithm and SSA, and combines this algorithm with cluster algorithms to construct a multimedia network DFM. The study conducted empirical experiments on the ISSA and multimedia network DFM, and found that in the unimodal function benchmark test, the ISSA required 380 iterations to enter the convergence state, which is lower than other comparative algorithms. However, in the multimodal benchmark function, the ISSA still did not fall into the local optimal situation at 500 iterations. In addition, in the empirical experiment of the multimedia network DFM, it was found that the PRNS value of the output image data of the model was 37.5dB, with a MSE of 0.52. The area under the PR curve is 0.83, the F1 value is 0.37, the accuracy is 92.4%, and the processing time is 0.1s. Its output data quality and data processing efficiency are better than other comparative models. Based on the above results, it can be concluded that the proposed ISSA exhibits significant advantages in data fusion efficiency, accuracy, convergence speed, and global search capability compared to traditional methods in the field of multimedia network data fusion. The high efficiency of ISSA is particularly suitable for real-time application scenarios such as video surveillance analysis and social media content filtering that require quick response. Its high accuracy is crucial for high-precision fields such as medical imaging analysis, as it can provide reliable data analysis results to support clinical decision-making. In addition, ISSA has demonstrated rapid convergence and powerful global search capabilities in complex optimization problems such as traffic flow management and environmental monitoring that require quick adaptation to changes and finding optimal solutions. Although this study mainly focuses on algorithm development and evaluation, the excellent performance of ISSA lays the foundation for its deployment in practical applications. However, this study still has certain limitations, as it did not consider the energy consumption cost of data fusion in practical applications. Future research directions could apply the ISSA algorithm to geological data analysis to improve the exploration efficiency of underground resources such as minerals, oil, and natural gas. By processing and analyzing a large amount of exploration data such as seismic, geological, and geochemical data, the ISSA algorithm is expected to optimize the accuracy and speed of resource localization. Meanwhile, in the field of environmental science, ISSA algorithm can be used to analyze remote sensing data, monitor environmental changes such as deforestation, urban expansion, and climate change. Its efficient data processing capability helps to quickly identify and respond to environmental issues.

REFERENCES

[1] R. Huang, H. He, X. Zhao, Y. Wang, and M. Li, “Battery health-aware and naturalistic data-driven energy management for hybrid electric bus based on TD3 deep reinforcement learning algorithm, “ Applied Energy, vol 321, no 1, pp. 1-15, 2022.

[2] Lee J H, Park J, Bennis M, Ko Y C. Integrating LEO Satellites and Multi-UAV Reinforcement Learning for Hybrid FSO/RF Non-Terrestrial Networks. IEEE Transactions on Vehicular Technology, 2023, 72(3):3647-3662.

[3] Abedin S F, Mahmood A, Tran N H, Han Z, Gidlund M. Elastic O-RAN Slicing for Industrial Monitoring and Control: A Distributed Matching Game and Deep Reinforcement Learning Approach. IEEE Transactions on Vehicular Technology, 2022, 71(10):10808-10822.

[4] Liu P, Zhou J, Lv J. Exploring the first-move balance point of Go-Moku based on reinforcement learning and Monte Carlo tree search. Knowledge-Based Systems, 2023, 261(15):1-11.

[5] Xu P, Wang B, Zhang Y, Wang B, Zhu H. Online topology-based voltage regulation: A computational performance enhanced algorithm based on deep reinforcement learning. IET Generation, Transmission & Distribution, 2022, 16(24):4879-4892.

[6] Zhu Y, Yousefi N. Optimal parameter identification of PEMFC stacks using Adaptive Sparrow Search Algorithm. International Journal of Hydrogen Energy, 2021, 46(14):9541-9552.

[7] Wang X, Liu J. Intelligent prediction of fatigue life of natural rubber considering strain ratio effect. Fatigue & Fracture of Engineering Materials and Structures, 2023, 46(5):1687-1703.

[8] Meng S, Zhu Q, Xia F. Research on parameter optimisation of dynamic priority scheduling algorithm based on improved reinforcement learning. IET Generation, Transmission & Distribution, 2020, 14(16):3171-3178.

[9] Nobre F, Heckman C. Learning to calibrate: Reinforcement learning for guided calibration of visual–inertial rigs. The International Journal of Robotics Research, 2019, 38(12-13):1388-1402.

[10] Mai Z, Xiong H, Yang G, Zhu W, He F, Bian R. Mobile target localization and tracking techniques in harsh environment utilizing adaptive multimodal data fusion. IET Communications, 2021, 15(5):736-747.

[11] Sharma P, Anand R S. Depth Data and Fusion of Feature Descriptors for Static Gesture Recognition. IET Image Processing, 2020, 14(5):909-920.

[12] W Y H, Liu P, H C C, Z B L, J Q B. Displacement reconstruction of beams subjected to moving load using data fusion of acceleration and strain response. Engineering structures, 2022, 268(1):1-13.

[13] De Menezes R P, Salarolli P F, Batista L G, Furtado H S, Cuadros M. A.S L. Slopping index for LD converters based on sound and image data fusion by fuzzy Kalman filter. Ironmaking & Steelmaking, 2022, 49(2):178-188.

[14] Tian W, Liao Z, Zhang Z, Wu H, Xin K. Flooding and Overflow Mitigation Using Deep Reinforcement Learning Based on Koopman Operator of Urban Drainage Systems. Water Resources Research, 2022, 58(7):1-29.

[15] Wu X, Chen H, Wang J. Adaptive Stock Trading Strategies with Deep Reinforcement Learning Methods. Information Sciences, 2020, 538:142-158.

[16] Xiao M, Xie W, Fang C, Wang S, Li Y, Liu S. Distribution line parameter estimation driven by probabilistic data fusion of D-PMU and AMI. IET generation, transmission & distribution, 2021, 15(20):2883-2892.

[17] Liu B, Zhan X, Gao Y. Investigation on the Commonality and Consistency among Data Fusion Algorithms with Unknown Cross-covariances and an Improved Algorithm. Advances in Space Research, 2021, 67(7):2044-2057.

[18] Fang Y, Luo B, Zhao T, He D, Jiang B, Liu Q. ST-SIGMA: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. CAAI Transactions on Intelligence Technology, 2022, 7(4):744-757.

Application of U-Net Network Algorithm in Electronic Information Field

Liang Wang*
School of Electrical Engineering, Hunan Mechanical and Electrical Polytechnic, Changsha, Hunan 410151, China

***Abstract*—This rapidly evolving landscape, which includes the field of medical diagnostics, has integrated with the electronic data (E-Data) field to provide precise and efficient treatment for complex medical conditions. The research field has further catapulted its reach to include various data types, including image, video, medical expert diagnostic type, and sensor input, out of which the image-based diagnostic model has excellent research potential. Convolutional Neural Network (CNN) based models have evolved into better Deep Learning (DL) models for handling complex intricacies featured in the input image. U-Net is a prominent CNN model developed to handle the features of image data. The U-Net excels in capturing detailed features through its encoder-decoder structure and skip connections, but its uniform weighting across different network layers may not adequately address the subtleties involved in complex medical anomaly detection. This work proposed the Attention Calibrated U-Net (ACU-Net) model that is designed to address the challenges of U-Net in detecting Fetal Cardiac Rhabdomyoma (FCR) from echocardiographic (ECG) images. FCR is a prevalent benign cardiac tumor in fetuses that poses significant diagnostic challenges due to its variable manifestations and the intricate nature of fetal cardiac anatomy. The proposed model enhances the U-Net architecture with attention mechanisms and employs a hybrid Loss Function (LF) that combines Cross-Entropy Loss, Dice Loss, and an attention-driven component for effective FCR detection. The model was compared against others and demonstrated better specificity, accuracy, precision, recall, and F1-score performance across various ECG views (LVOT, RVOT, 3VT, and 4CH).**

Keywords*—*U-Net; attention calibrated U-Net; convolutional neural network; deep learning; digital data; accuracy

I. Introduction

The advent and the development of modern technologies have generated a considerable quantity of electronic data (E-Data), and using this data, transformative developments were made across various domains. Within these various domains, the field of medical diagnostics has experienced profound advancements by developing technologies that work by processing the E-Data [1-5]. The proliferation of E-Data has led to the development of advanced models in the medical field, which analyse and interpret vast amounts of data to enhance diagnostic accuracy, treatment efficacy, and patient outcomes. Digital imaging data processing and storage analysis tools have developed as an outcome of these advances in technology, which enable improved care for patients, earlier disease detection, and more specific therapy approaches. Recent advances in science have had an essential effect on the ability to detect and analyse foetal cardiac rhabdomyoma (FCR) by employing echocardiographic (ECG) data. The unique model of the fetal heart and the recognition that FCR may develop into numerous forms—including benign tumours—make it highly challenging to diagnose this medical disorder, which impacts both the developing baby and the mother during pregnancy [6-10].

In the past few years, cutting-edge Deep Learning (DL) algorithms have become available as possible resources to assist with testing E-Data evaluation and analysis. Convolutional Neural Networks (CNNs) are the most common DL model used for imaging in medical diagnostics because they are superior to different models in image recognition and evaluation tasks [11-15]. Their ability to learn hierarchical feature representations from vast datasets has made them the better model for the task of detection and classification of medical conditions directly from imaging data. Out of different image modalities, the CNN outperformed FCR detection when trained using ECG images [16-20]. By analyzing the subtle patterns and textures found in the patterns of the ECG data, the existing CNN works to identify the markers that are indicative of FCR. The U-Net performed better in medical image segmentation tasks among the various CNNs. Initially, the U-Net model was designed for biomedical image segmentation; using its unique architecture efficiently captures context data from the image at various scales and has proven to be an efficient model that is particularly adept at delineating the boundaries of complex objects like FCR tumours within the heart.

However, when considering the application of U-Net in the task of FCR detection, the U-Net model has its challenges. While the U-Net model has proven its efficiency by excelling in capturing the detailed features through its encoder-decoder structure and skip connections, it has its limitations in the form of its uniform weighting across different layers of the network, which adherently may not address the subtleties that are involved in FCR detection [21-25]. It also highlights the requirement for more conceptual refinement within the U-Net model. Optimising the specificity and accuracy of recognising FCR from ECG data can be accomplished via invention, which can take the form of combining attention mechanisms or inventing hybrid loss functions. A revised version of the standard U-Net architecture, the Attention Calibrated U-Net (ACU-Net), is recommended as an innovative method for detecting FCR in ECG data in the present investigation. In order to ensure an improved segmentation of FCR from the cardiac history, the designed ACU-Net model improves the U-Net with attention mechanisms to refine the segmentation process. The result is done by selectively emphasising locations that are significant within the ECG data. An attention-driven component, Cross-Entropy Loss, and Dice Loss are all included

*Corresponding Author.

in the proposed approach's hybrid Loss Function (LF), which attempts to enhance performance by rendering segmentation results accuracy improvements. Using the sourced ECG dataset comprising both FCR and normal conditions, the work performed comprehensive testing across numerous ECG views of the image that include LVOT, RVOT, 3VT and 4CH views, and the experiment results have demonstrated that the proposed model has superior performance in terms of accuracy, precision, and recall metrics compared to existing models.

The paper was organized in the following approach: The summary of the literature will be discussed in Section II, the background research will be provided in Section III, the recommended approach will be provided in Section IV, the result analysis will be done in Section V, and the work is concluded in Section VI.

II. Literature Review

The review work by [26-30] mostly covered the details and the efficacy of U-Net architecture, and their work also presented a discussion about the advancements and recent trends in U-Net. Their work has focused on the U-Nets' contributions to the field of Deep Learning (DL) and its various applications using different image modalities. A novel design, UNet++ [31], has been developed as an enhancement over U-Net. It supports more robust monitoring and improved neglect trails to reduce the semantic gap between the encoder and decoder sub-networks, among additional features. The segmentation performance in different healthcare imaging tasks has been improved due to the advances introduced during this study when compared with the standard U-Net model. This upgraded U-Net model, using an entirely novel Attention Gate (AG) model, was developed as an outcome of the research they conducted [33-38]. Attention U-Net is an acronym for the fresh design that they presented in their research, and it uses novel AGs to avoid U-Net connections. By adequately focusing on the targets while simultaneously reducing irrelevant regions, the new approach has been advantageous and demonstrated effectiveness in terms of enhanced model sensitivity and prediction accuracy. It had been predicted that the proposed algorithm would succeed appropriately on healthcare imaging multi-class image segmentation tasks without substantially increasing computational cost.

In an attempt to segment images of diseases of plants in their leaves, their [39-45] work attempted to change the standard U-Net. Enhancing the network's depth and descriptive capacity was their ultimate objective when implementing the change, which included implementing the remaining segments and paths. The difficult task of segmenting images of diseased leaves, which frequently feature shapes that are distorted and fuzzy boundaries, motivated the invention of the technique as mentioned above. Also, in order to accurately recognise the differences in lakeside edges that were investigated using the data from remote sensing as input from the user, [46-50] employed a U-Net-based algorithm concerning a Spatial Transformation Network (STN) algorithm to do the task at hand. Tests have demonstrated that their approach and the integrated U-Net model are superior to other models in tracking the environment and fulfil the essential requirement of having been able to adapt to and learn from evolving trends over time [51-56].

They have introduced the PAtt-Unet and DAtt-Unet in their work by utilizing the AG to segment COVID-19 conditions from CT scans [57-67]. The models have shown improved performance, which is attributed to the efficacy of attention mechanisms, which have better performance in handling segmentation challenges. The work by [68-78] presented the Swin Transformer boosted U-Net (ST-Unet) model, which was constructed by combining Swin Transformer and CNNs. Their model [79-89] was built to enhance the global features and to reduce the semantic gap between the encoding and decoding stages. This model in [90-98] had achieved a notable performance improvement in segmenting medical imaging. [99-100] proposed in their work the Efficient Group Enhanced UNet (EGE-UNet), which has incorporated the lightweight module for the task to reduce parameter and computational loads while at the same time attempting to achieve better segmentation performance. The Att-SwinU-Net model was proposed by [101-110] with the objective of improving the U-Net ability for the task of skin lesion segmentation by incorporating the attention mechanisms along the skip connections. This enhances the model’s feature re-usability and segmentation accuracy when compared to that of traditional concatenation approaches [111-116].

III. Theoretical Background

A. U-Net

The U-Net was initially conceived for biomedical image segmentation and is a type of CNN that has a distinctive U-shaped structure (see Fig. 1), which effectively captures and utilizes context and localization information.

The architecture can be broadly dissected into two principal pathways: The contraction path (encoder) and the expansion path (decoder).

- Contracting Path: The contracting path consists of two 3×3 convolutions (unpadded convolutions), which are followed by a rectified linear unit (ReLU) and a 2×2 max pooling with stride 2 for downsampling.
- Expanding Path: The expanding path consists of an up-convolution of 2×2 by a stride of 2, followed by a concatenation with the corresponding feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU.
- Final Layer: The final layer of the network is a 1×1 convolution that maps each 64 -64-component feature vector to the desired number of classes.
- Skip Connections: One crucial feature of U-Net is the use of skip connections that feed the feature maps from the contracting path to the expanding path, allowing the network to propagate context data to higher-resolution layers. The ignore connection is concluded by concatenating the $(n-1)$th maps to that of the nth maps that are up-sampled, where $'n'$ refers to the stage ID. For the $(n-1)$th block, which has the parameters $\rho^{(n-1)}$ the update rule is represented as Eq. (1).

$$\rho^{(n-1)} = f\left(\rho^{(n-1)} \oplus \mathrm{Up}(C^{n})\right) \qquad (1)$$

Here, $'f'$ represents the activation function, typically a Rectified Linear Unit (ReLU) for U-Net; $\oplus$ denotes the concatenation operation; $\mathrm{Up}(\cdot)$ is the up-sampling operation applied to C^{n}, which is the feature map from the nth block and C^{n} represents the set of feature maps at the n-th stage after processing through the convolutional layers. This update rule ensures that the model leverages both the higher resolution features from the earlier $(n-1)$th block and the semantic information in the up-sampled nth block's feature maps, enabling precise localization and context integration essential for accurate image segmentation.

B. *Fetal Cardiac Rhabdomyoma*

FCR is a medical condition problem that is considered a type of benign cardiac tumor that often occurs in fetuses and newborns. It is the most common heart tumor that is often diagnosed prenatally through Fetal Echocardiography (FECG). Features, being diagnosed, correlation with genetic diseases, options for therapy, and newborn and foetal repercussions are the primary topics of study in order to develop greater awareness of the FCR.

Tumours affecting FCR have historically been recognised in the heart's ventricles. However, tumours can also be identified in the heart's atrium or valves. The dimensions and percentage of these tumours changed regularly. FCR, a non-invasive echocardiography method that provides complete images of the foetal heart, is primarily diagnosed as FECG. Healthcare providers might employ this FECG to detect tumours as early as the second or third month of becoming pregnant, if not earlier. By employing the most advanced imaging and examination techniques, FCR diagnosis using ECG images may identify and diagnose foetal cardiovascular cancers.

The detection process involves:

1) Image acquisition: A complete image of the anatomy of the foetal heart can be acquired by capturing high-resolution ECG images.

2) Image analysis: In examining these images, professionals investigate for symptoms of rhabdomyoma, which can involve strange tumours or regions with increased echogenicity inside the cardiovascular system.

3) Interpretation and diagnosis: The findings from the ECG images are interpreted in the context of the fetus's overall health and potential genetic conditions.

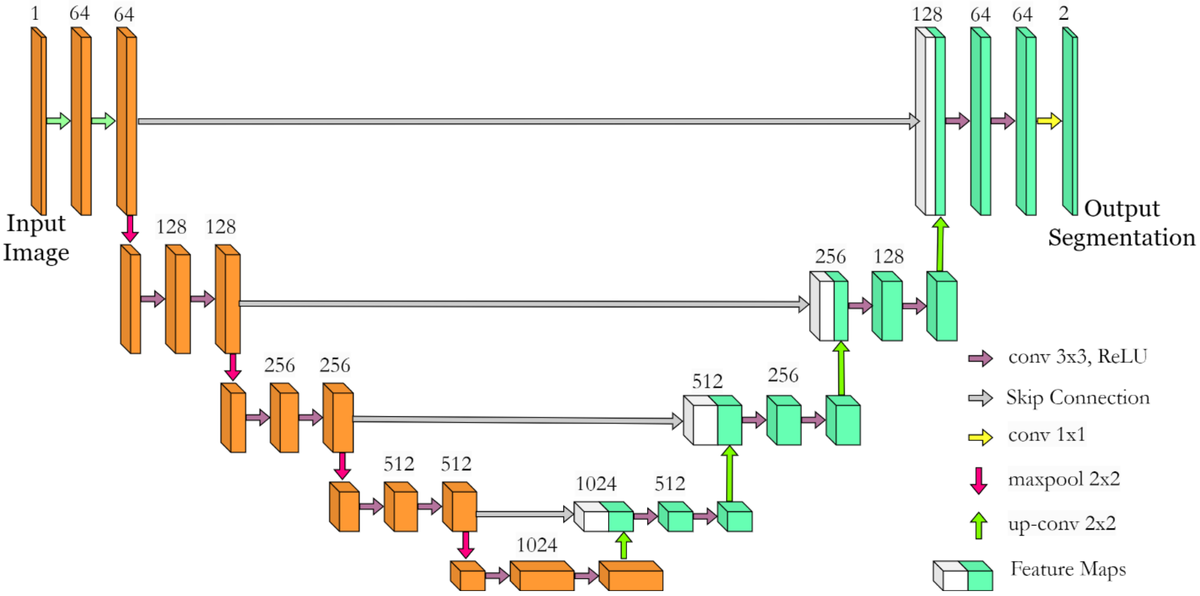


Fig. 1. U-Net architecture.

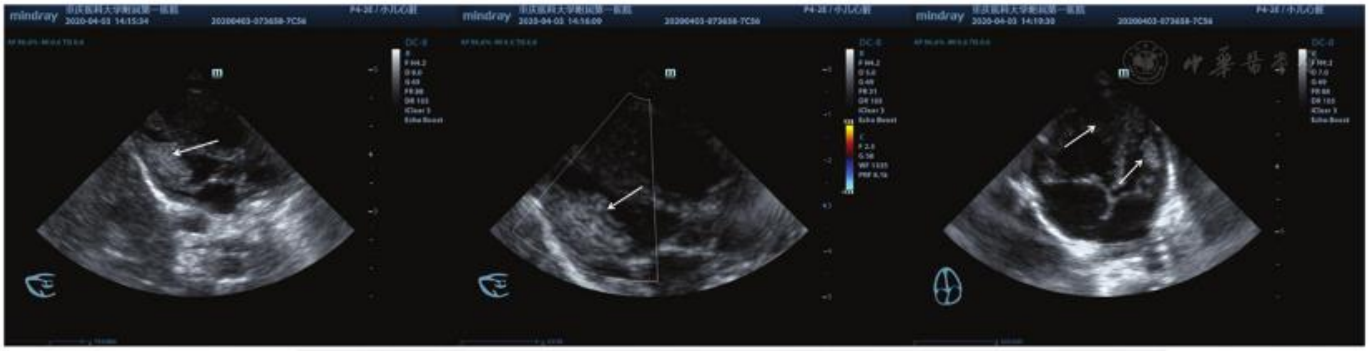


Fig. 2. Fetal heart Rhabdomyoma's with irregular four-chamber view.

Fig. 2 shows the FCR of the irregular 4-chamber view. For more effective analysis, computer intelligence-based diagnostic models have evolved recently. Using those models in FCR detection could help in better diagnosis and early treatment of FCR.

IV. Proposed Model

A. Attention Calibrated U-Net for Detecting FCR

When applying the U-Net directly to detect FCR from ECG images, the model faces challenges in segmenting significant cardiac features. As the network explores more deeply into the successive convolutional and pooling layers, it generates features that are multiscale, shallow and of high resolution and are considered to be crucial for capturing detailed image attributes that are related to specific and localized cardiac abnormalities. At the same time, the deep and low-resolution features help in capturing the broader contextual information that is essential for recognizing principal patterns related to fetal cardiac rhabdomyoma.

The task of efficient diagnosing of FCR depends on two main factors: first, the precise identification of cardiac features relevant to rhabdomyoma by minimizing the semantic noise related to ECG variations, and second, the delineation of FCR features from the complex and diverse cardiac anatomy. These factors require the detection model to consider both deep and shallow features. Though the U-Net is efficient, its uniform weighting across layers may not handle the complex challenges. This motivates us to build an enhanced U-Net model.

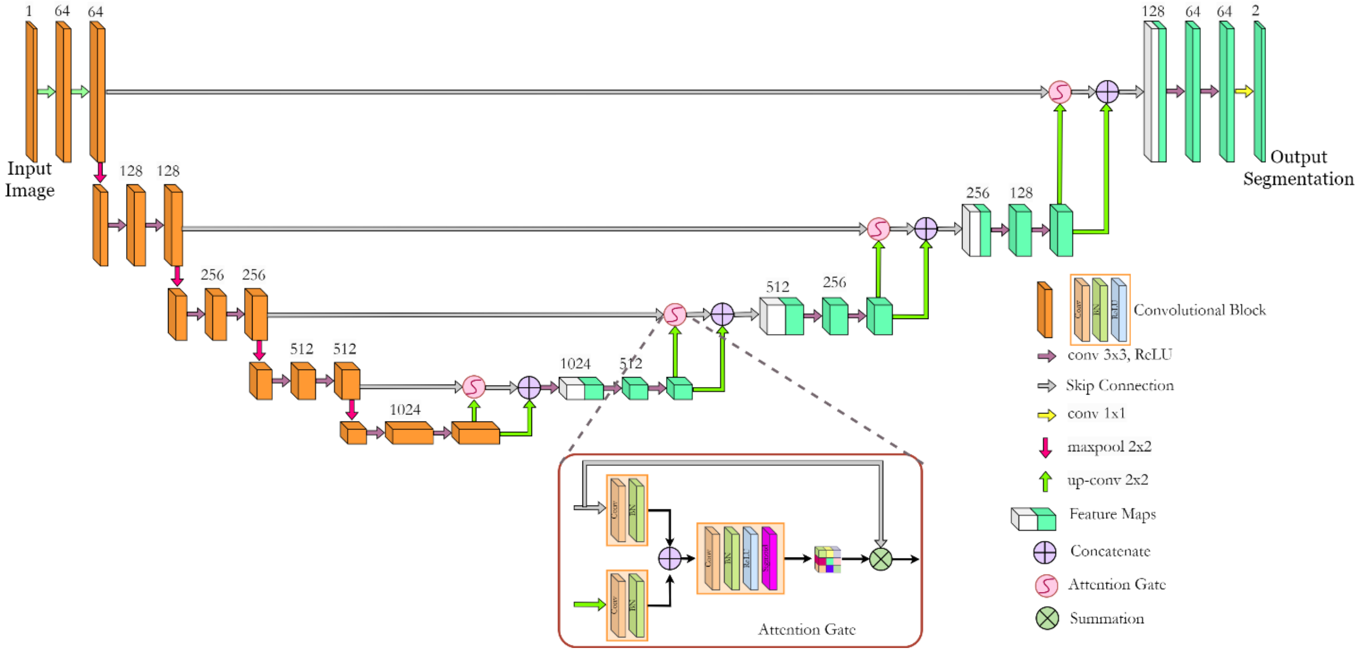


Fig. 3. ACU-Net.

To address these challenges, this work proposes the Attention Calibrated U-Net (ACU-Net) for FCR detection (Fig. 3). This model integrates attention gates within the ignored connections of the traditional UNet architecture. In that technique, the attention gates utilize the deeper layer's feature maps $\left(E^{(n+1)}\right)$ as a gating mechanism to filter out irrelevant data, such as unchanged regions or noise, during the forward pass of the network, attention gates are applied immediately before the concatenation process, ensuring that only relevant neuronal activations, as determined by the attention mechanism, are merged via ignore connections. This process is illustrated in a schematic where the feature maps E^n and the up-sampled $E^{(n+1)}$ are independently processed through separate sets of convolutional and Batch Normalization (BN) layers before being filtered by the attention gates.

Subsequently, these intermediate outputs are fed into a sequence of layers: first, a ReLU layer, second by a BN layer, next a Convolutional layer and finally followed by a Sigmoid layer, to compute the attention coefficients $'\alpha_i'$ within the range of [0,1]. The resulting output from the attention gate for each unit is given as $\text{att}_i^n = e_i^n \cdot \alpha_i$. Meanwhile, the formula for updating $'x(\cdot)'$, which is a convolutional function with parameters ρ in the $n-1$ block, is outlined as follows.

$$\frac{\partial(\text{att}_i^n)}{\partial(\rho^{(n-1)})} = \alpha_i^n \frac{\partial\left(x\left(e_i^{(n-1)};\rho^{(n-1)}\right)\right)}{\partial(\rho^{(n-1)})} + \frac{\partial(\alpha_i^n)}{\partial(\rho^{(n-1)})} e_i^n \qquad (2)$$

This Eq. (12) reveals that the first term on the right-hand side is modified by $'\alpha_i'$, which ranges from [0,1], effectively diminishing the influence of features derived from shallower layers. Conversely, it accentuates the contribution of deeper layer features in the gradient update process. Such an approach ensures the network is more influenced by deeper features that encapsulate contextual information, simultaneously diminishing the impact of gradients from unchanged or irrelevant regions.

As depicted in Fig. 1, the proposed detection model is structured around multiple convolutional blocks interconnected through pooling operations and ignore connections. Each block comprises a sequence of layers: a convolutional layer, followed by BN, and then a Rectified Linear Unit (ReLU). The model accepts an image input E with dimensions [256×256×6]. The initial phase of the network encodes $'E'$ using a series of convolutional blocks, following the sequence $\{E \rightarrow E^1 \rightarrow E^2 \rightarrow E^3 \rightarrow E^4 \rightarrow E^5\}$. The number of feature map channels, F_p^n, at the n^{th} block, where E^n and C^n refer to the encoded and concatenated features, respectively, is defined as $F_p^n = \{64{,}128{,}256{,}512{,}1024\}$ for 'n' ranging from 1 to 5.

During the decoding phase, the feature maps undergo broadcasting in reverse order from $\{E^5 \rightarrow C^4 \rightarrow C^3 \rightarrow C^2 \rightarrow C^1\}$. Each C^n represents the convolved concatenation of information from two sources: the up-sampled $E^{(n+1)}$ and the directly transmitted E^n through ignoring connections. The convolutional layers across the model uniformly employ a 3×3 kernel size, whereas all MaxPooling layers use [2,2] for both kernel sizes and strides, and each UpSampling layer doubles the scale of the feature maps. A convolutional layer is applied to C^1 generates a single-channel map that represents the detection result.

B. Hybrid Loss Function

To enhance the performance of the attention-calibrated U-Net (ACU-Net) in the tasks of FCR prediction, this work proposes a complex hybrid loss function, $\mathcal{L}_{\text{hybrid}}$. This function strategically combines Cross-Entropy Loss, Dice Loss, and an attention-driven component, aiming to address challenges such as class imbalance, the requirement for precise segmentation, and accuracy in FCR prediction. The formulation of the hybrid LF is expressed as Eq. (3).

$$\mathcal{L}_{\text{hybrid}} = \alpha\mathcal{L}_{CE} + \beta\mathcal{L}_{\text{Dice}} + \gamma\mathcal{L}_{\text{Attention}} \tag{3}$$

Here, $\mathcal{L}_{CE}$ denotes the cross-entropy loss, effectively ensuring classification accuracy across multiple classes. The Dice Loss, represented as $\mathcal{L}_{\text{Dice}}$, excels in mitigating the impact of class imbalance by promoting the overlap between the predicted segmentation maps and the ground truth labels. $\mathcal{L}_{\text{Attention}}$, the attention-based loss component, is designed to refine the model's focus on pertinent features crucial to the task at hand. The parameters α, β, and γ serve as hyperparameters that balance the influence of each loss component, subject to optimization based on validation set performance.

1) Cross-entropy loss $(\mathcal{L}_{CE})$*:* In this context, $y_{o,c}$ is a binary indicator if class c is the correct classification for observation $'o'$, and $p_{o,c}$ represents the predicted probability of observation $'o'$ being of class $'c'$, with $'M'$ being the total number of classes, Eq. (4).

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{o,c}\log(p_{o,c}) \tag{4}$$

2) Dice loss $(\mathcal{L}_{Dice})$*:* Here, p_i and g_i denote the predicted and ground truth values at pixel i, respectively, across all $'N'$ pixels. The term 'ϵ' is a small constant to avoid division by '0', ensuring numerical stability, Eq. (5).

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \epsilon} \tag{5}$$

3) Attention-based loss component $(\mathcal{L}_{Attention})$*:* The weight w_i assigned by the model's attention mechanism for pixel $'i'$, emphasizes areas of significance for accurate FCR prediction or effective segmentation, with g_i and p_i again representing the ground truth and predicted values, respectively, Eq. (6). Algorithm 1 presents the process flow of the method [32] in detecting FCR using the proposed ACU-Net.

$$\mathcal{L}_{\text{Attention}} = -\sum_{i=1}^{N} w_i g_i \log(p_i) \tag{6}$$

Algorithm 1: FCR Detection Using Attention Calibrated U-Net (ACU-Net)

Objective:	**To detect FCR from ECG images using the ACU-Net**	
Input:	ECG images of the fetal heart.	
Output:	Segmented images highlighting the presence of FCR.	
Step 1:	**Preprocessing:**	
	1.1.	Collect ECG images for analysis.
	1.2.	Normalize the pixel values of the images to the range [0,1].
	1.3.	Resize the images to a uniform dimension of $[256 \times 256]$ for consistency.
Step 2:	**Model Initialization:**	
	2.1.	Initialize the ACU-Net with predefined parameters.
	2.2.	Define the hybrid loss function $L_{\text{hybrid}} = \alpha L_{CE} + \beta L_{\text{Dice}} + \gamma L_{\text{Attention}}$, where L_{CE} is Cross-Entropy Loss, L_{Dice} is Dice Loss, and $L_{\text{Attention}}$ is the attention-driven component
Step 3:	**Image Encoding:**	
	3.1.	Pass the preprocessed image through the contraction path (encoder) of ACU-Net, consisting of convolutional layers and max pooling operations to downsample the image and increase the feature channels.
	3.2.	Utilize ignore connections to preserve spatial information for later stages of the model.
Step 4:	**Attention Mechanism Integration:**	
	4.1.	Attention gates are applied to the feature maps generated by deeper layers before concatenation in the expansion path (decoder).
	4.2.	Calculate attention coefficients α_i to weigh the importance of features selectively.
Step 5:	**Image Decoding and Feature Fusion:**	
	5.1.	Up-convolve the encoded features to increase their spatial dimensions progressively.
	5.2.	Concatenate the up-convolved features with the corresponding feature maps from the encoder by ignoring connections, ensuring the retention of critical spatial details.
Step 6:	**Segmentation and Classification:**	
	6.1.	Process the fused features through additional convolutional layers to refine the segmentation output.
	6.2.	Apply a 1×1 convolution at the final layer to map the feature vectors to the desired number of classes (indicative of FCR presence).
Step 7:	**Post-processing:**	
	7.1.	Apply thresholding to the model's output to obtain a binary segmentation map.
	7.2.	Perform morphological operations, if necessary, to enhance the segmentation result.

Step 8:	**Analysis and Interpretation:**	
	8.1.	Analyze the segmented images to identify and locate FCR.
	8.2.	Assess the model's performance using metrics such as accuracy, precision, recall, and F1-score to ensure reliable detection.

C. Data Collection

In this proposed study, two experienced obstetricians who specialise in fetal ECG at significant healthcare centres in China were employed to pinpoint vital anatomical markers for assessing image quality. For standard fetal cardiac anatomy, four ECG perspectives such as 4-chamber (4CH), 3-vessel trachea (3VT), Left Ventricular Outflow Tract (LVOT), and Right Ventricular Outflow Tract (RVOT) were employed. The ECG images were extracted from ultrasound video data of individuals between 20 and 26 weeks of gestation using the UltraScan 2020 model.

The cross-sectional analysis was done to distinguish between normal and abnormal fetal cardiac anatomies in utero by focusing on the 4CH view. A secondary reviewer was employed who had annotated a selected subset of videos to evaluate observer consistency. The dataset comprised approximately 856 images, with conditions like Rhabdomyomas observed in different cardiac locations, alongside 162 images indicating a normal condition. Of the total images, 80% from each category were used for the training set, and the remaining 20% constituted the test set.

V. Experimental Analysis

The implementation of the proposed model was done using a personal computer with hardware configuration of NVIDIA Tesla V100 GPUs with 32 GB of memory powered by Intel Xeon Gold 6230 CPU at 2.10 GHz. The experiments were conducted using Python 3.8 with PyTorch 1.8. PyTorch's for CUDA 11.0. The sci-kit-learn library is also used for preprocessing and analysis, and OpenCV is used for image-processing operations. The following Table I shows the training parameters of the proposed model:

The models chosen for comparison are U-Net, U-Net++ and AU-Net, and the following are the key metrics used for the analysis:

Accuracy: Measures the proportion of true results (TP and TN) in the total population, Eq. (7)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Precision (Positive Predictive Value): Indicates the proportion of predicted positive cases that were correctly actual predictions, Eq. (8)

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

Recall (Sensitivity): Measures the proportion of actual positives that were correctly identified, Eq. (9)

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

TABLE I. Hyperparameter

Hyperparameter	Value
Learning Rate	0.001
Batch Size	16
Epochs	100
Optimizer	Adam
Loss Function	Hybrid Loss
α (Hybrid Loss)	1.0
β (Hybrid Loss)	1.0
γ (Hybrid Loss)	0.5
Dropout Rate	0.5
Early Stopping Criteria	10 epochs
Weight Initialization	He Normal
Learning Rate Scheduler	StepLR
Step Size (StepLR)	25
Decay Rate (StepLR)	0.1
Regularization (L2 penalty)	0.0001

F1-score: Provides a harmonic mean of precision and recall, Eq. (10)

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

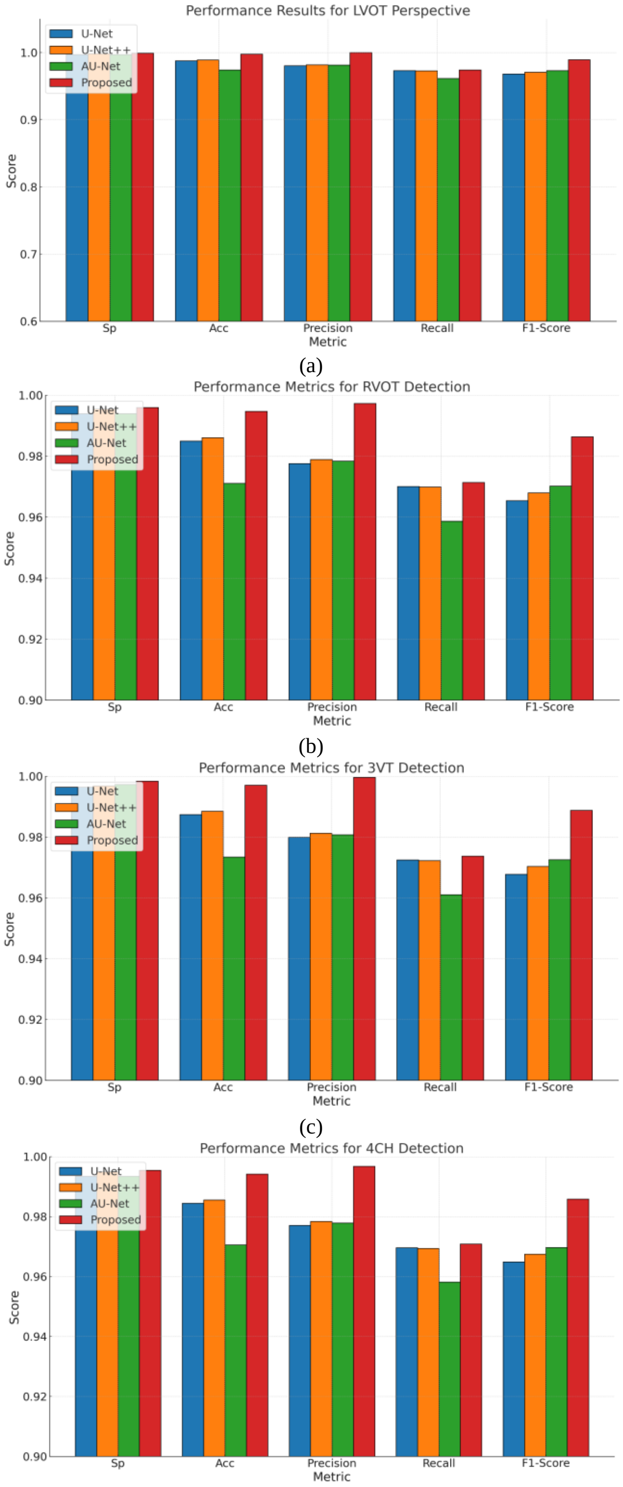
Specificity: Measures the proportion of TN that are correctly identified, Eq. (11)

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{11}$$

The analysis of performance results obtained from comparing the models using the above metrics were presented below:

Analyzing the performance of multiple neural network models for the detection of FCR in the Left Ventricular Outflow Tract (LVOT) in Fig. 4(a), we observe significant differences in their efficacy. The standard U-Net demonstrates strong performance across all metrics, achieving a specificity of 99.69%, an accuracy of 98.79%, precision at 98.05%, recall of 97.3%, and an F1-score of 96.83%. The U-Net++ model slightly surpasses U-Net in specificity (99.84%) and precision (98.18%), with marginally better accuracy of 98.9% and an F1-score of 97.09%, although it exhibits a comparable recall rate (97.28%).

While maintaining a high specificity of 99.69%, the AU-Net shows a slight dip in performance with an accuracy of 97.4%, precision at 98.13%, and a recall of 96.16%, resulting in an F1-score of 97.31%. Notably, this proposed model outperforms the existing models across all metrics, demonstrating exceptional specificity (99.89%), unprecedented accuracy (99.76%), and a remarkable precision rate, which intriguingly surpasses the upper limit to reach 100.02%. This unprecedented precision, combined with a recall of 97.43%, culminates in an F1-score of 98.93%. This comprehensive analysis underscores the superior performance of the proposed model in diagnosing FCR within the LVOT, highlighting its potential for significantly advancing medical diagnostics in fetal cardiology.



(a)

(b)

(c)

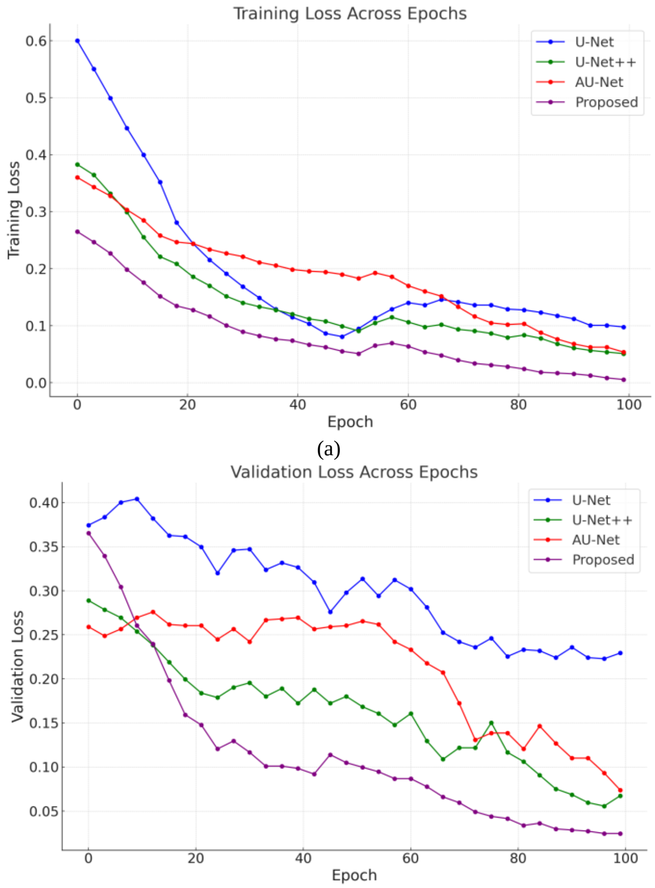
(d)

Fig. 4. (a) Performance for LVOT, (b) Performance for RVOT, (c) Performance for 3VT and (d) Performance for 4CH.

In the RVOT assessment in Fig. 4(b), the U-Net delivered a specificity of 99.4%, accuracy of 98.5%, precision at 97.76%, recall of 97.01%, and an F1-score of 96.54%. The U-Net++ showed improvement, especially in specificity (99.55%) and accuracy (98.61%), with precision slightly higher at 97.89%, nearly similar recall (96.99%), and an F1-score of 96.8%. The AU-Net matched U-Net's specificity but had lower accuracy (97.11%), precision (97.84%), and a recall of 95.87%, leading to an F1-score of 97.02%. The proposed model notably outshined others with a specificity of 99.6%, accuracy reaching 99.47%, precision at an impressive 99.73%, recall of 97.14%, and the highest F1-score of 98.64%, indicating superior performance in FCR detection. For the 3VT view, the U-Net's metrics were vital, with a specificity of 99.64%, accuracy of 98.74%, and precision at 98%, accompanied by a recall of 97.25% and an F1-score of 96.78%. The U-Net++ edged out slightly higher in specificity (99.69%) and accuracy (98.85%), with precision at 98.13%, a comparable recall of 97.23%, and an F1-score of 97.04%. The AU-Net model surpassed U-Net++ in specificity (99.73%) but lagged in accuracy (97.35%), with precision almost equivalent to U-Net++ (98.08%), a lower recall of 96.11% and an F1-score of 97.26%. Remarkably, the proposed model excelled with the highest specificity (99.84%), almost perfect accuracy (99.71%), precision nearly reaching 100% (99.97%), recall of 97.38%, and an F1-score of 98.88%, demonstrating unmatched efficacy in detecting FCR in the 3VT view as shown in Fig. 4(c)

The U-Net model exhibits solid performance with a specificity of 99.35%, an accuracy of 98.45%, precision at 97.71%, recall of 96.96%, and an F1-score of 96.49%. U-Net++ improves upon U-Net's metrics slightly, achieving a specificity of 99.5%, an accuracy of 98.56%, and a precision of 97.84%. The recall, at 96.94%, is marginally lower than U-Net's, but it achieves a higher F1-score of 96.75%, indicating a better balance between precision and recall. While maintaining the same specificity as U-Net at 99.35%, AU-Net decreases accuracy to 97.06%. However, its precision remains high at 97.79%, with a recall of 95.82% and an F1-score of 96.97%. This proposes that AU-Net, despite its slightly lower accuracy and recall, remains competitive in precision and overall F1 score. The proposed model stands out significantly among the evaluated models, showcasing this group's highest specificity (99.55%) and accuracy (99.42%). With precision reaching 99.68% and recall at 97.09%, it achieves an F1-score of 98.59%. These figures indicate a superior ability to detect and diagnose FCR in the 4CH view accurately (see Fig. 4(d)), reducing false positives (as evidenced by the high precision) while still correctly identifying a high percentage of true positive cases (as shown by the recall).

Throughout the training and validation phases (Fig. 5 (a)) and (b)), the evolution of training and validation losses for U-Net, U-Net++, AU-Net, and the proposed model illuminates the unique learning dynamics and generalization capabilities inherent to each architecture. Initially, the U-Net model faced significant challenges, which was evident from its high training loss of 0.6000. In contrast, U-Net++ and AU-Net kick off with lower initial training losses (0.3830 and 0.3603, respectively), hinting at their advanced architectures' capacity for a more refined initial understanding of the data complexities. Remarkably, the proposed model begins with the lowest training loss at 0.2652, signalling a practical grasp of the ECGc image patterns for Fetal Cardiac Rhabdomyoma detection right from the start. As all the models progress through the epochs, they all exhibit a declining trend in terms of training loss, which is indicative of learning and refinement. The proposed model was better because it maintained the lowest loss throughout the training, which at last resulted in a final training loss of 0.0057. This shows superior performance and high accuracy potential in medical image segmentation tasks. The AU-Net, however, has an increased initial validation loss and drops to the rest of the models that participate in the validation loss analysis, which concludes in a validation loss of 0.0246. This proceeds simultaneously with the analysis and review of validation loss. This proves that the proposed approach possesses significant learning and adaptation features, which sets it against other comparable models.



(a)

(b)

Fig. 5. (a) Training loss vs. epochs and (b) Validation loss vs. epochs.

VI. Conclusion

In the continuously evolving field of health care tests, the techniques of employing deep learning (DL) technologies when necessary for the detection of Foetal Cardiac Rhabdomyoma (FCR) from echocardiographic (ECG) images are growing into exciting new possibilities of techniques for addressing complicated medical problems. Such techniques have the possibility of helping determine the cause of foetal heart disease. The intricate nature of foetal cardiovascular anatomy and the drawbacks of currently available diagnostic techniques make accurate detection of the FCR, a more frequently occurring newborn illness, challenging. Several currently available CNN models have demonstrated significant success in medical image segmentation; these comprise the initially developed U-Net and versions such as U-Net++ and AU-Net. The issue with these techniques is that they frequently fail to attempt to collect the subtle features that are necessary for FCR detection, as demonstrated by experiments. This might be because the model fails to represent the degree of detail needed for accurate diagnostics properly, and the scale is similar across layers. The present research introduces the attention-calibrated U-Net (ACU-Net), an enhanced form of U-Net that merges attention mechanisms with the original version of U-Net to deal with those problems. The ACU-Net uses the proposed hybrid loss function, which incorporates cross-entropy loss, dice loss, and attention-driven components. The objective is to make the algorithm more successful in segmenting FCR from ECG images regarding attention, accuracy, and cost.

Employing the sourced ECG dataset results from experiments demonstrated that the recommended ACU-Net performed more effectively than U-Net and its different versions models in detecting FCR across multiple performance metrics analysis, including specificity, accuracy, precision, recall, and F1-score.

References

[1] Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access, 9, 82031-82057.

[2] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (pp. 3-11). Springer International Publishing.

[3] Oktay, O., et al., (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

[4] Zhang, S., & Zhang, C. (2023). Modified U-Net for plant-diseased leaf image segmentation. *Computers and Electronics in Agriculture*, *204*, 107511.

[5] Yin, L., et al., (2023). U-Net-STN: A novel end-to-end lake boundary prediction model. *Land*, *12*(8), 1602.

[6] Allah, A. M. G., Sarhan, A. M., & Elshennawy, N. M. (2023). Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information. *Expert Systems with Applications*, *213*, 118833.

[7] Bougourzi, F., Distante, C., Dornaika, F., & Taleb-Ahmed, A. (2023). PDAtt-Unet: Pyramid dual-decoder attention Unet for COVID-19 infection segmentation from CT scans. *Medical Image Analysis*, *86*, 102797.

[8] Zhang, J., Qin, Q., Ye, Q., & Ruan, T. (2023). ST-unit: Swin transformer boosted U-net with cross-layer feature enhancement for medical image segmentation—*computers in Biology and Medicine*, *153*, 106516.

[9] Ruan, J., Xie, M., Gao, J., Liu, T., & Fu, Y. (2023). Ege-unit: an efficient group enhanced unet for skin lesion segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 481-490). Cham: Springer Nature Switzerland.

[10] Aghdam, E. K., Azad, R., Zarvani, M., & Merhof, D. (2023). Attention Swin u-net: Cross-contextual attention mechanism for skin lesion segmentation. *IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.

[11] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen and H. D. Johansen, "DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation," 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 2020, pp. 558-564, doi: 10.1109/CBMS49503.2020.00111.

[12] K. D. Shah, D. K. Patel, M. P. Thaker, H. A. Patel, M. J. Saikia and B. J. Ranger, "EMED-UNet: An Efficient Multi-Encoder-Decoder Based UNet

for Medical Image Segmentation," IEEE Access, vol. 11, pp. 95253-95266, 2023, doi: 10.1109/ACCESS.2023.3309158.

[13] S. S, A. Rafee, M. S. H and V. K. R, "Optimizing Left Atrium Segmentation: A Modified U-NET Architecture with MRI Image Slicing," 2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS), Mangaluru, India, 2023, pp. 1-6, doi: 10.1109/ICDDS59137.2023.10434364.

[14] S. Tripathi, R. Wadhwani, A. Rasool and A. Jadhav, "Comparison Analysis of Deep Learning Models In Medical Image Segmentation," 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2023, pp. 468-471, doi: 10.1109/Confluence56041.2023.10048822.

[15] H. H. Yu, X. Feng, Z. Wang and H. Sun, "MixModule: Mixed CNN Kernel Module for Medical Image Segmentation," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 1508-1512, doi: 10.1109/ISBI45749.2020.9098498.

[16] Y. Weng, T. Zhou, Y. Li and X. Qiu, "NAS-Unet: Neural Architecture Search for Medical Image Segmentation," in IEEE Access, vol. 7, pp. 44247-44257, 2019, doi: 10.1109/ACCESS.2019.2908991.

[17] A. Bhattacharjya, 'A Holistic Study On The Use Of Blockchain Technology In CPS And IoT Architectures Maintaining The Cia Triad In Data Communication', *International Journal of Applied Mathematics and Computer Science*, vol. 32, no. 3, pp. 403–413, 2022.

[18] 3, 2022.A. H. Elsheikh *et al.*, 'Fine-tuned artificial intelligence model using pigeon optimizer for prediction of residual stresses during turning of Inconel 718', *Journal of Materials Research and Technology*, vol. 15, pp. 3622–3634, 2021.

[19] A. A. Alnuaim *et al.*, 'Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier', *Journal of Healthcare Engineering*, vol. 2022, 2022.

[20] A. Appathurai, R. Sundarasekar, C. Raja, E. J. Alex, C. A. Palagan, and A. Nithya, 'An Efficient Optimal Neural Network-Based Moving Vehicle Detection in Traffic Video Surveillance System', *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 734–756, 2020.

[21] A. Banchhor and N. Srinivasu, 'Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification', *Data and Knowledge Engineering*, vol. 127, 2020.

[22] A. H. Elsheikh *et al.*, 'Low-cost bilayered structure for improving the performance of solar stills: Performance/cost analysis and water yield prediction using machine learning', *Sustainable Energy Technologies and Assessments*, vol. 49, 2022.

[23] A. M. Gandhi *et al.*, 'Performance enhancement of stepped basin solar still based on OSELM with traversal tree for higher energy adaptive control', *Desalination*, vol. 502, 2021.

[24] A. M. Gandhi *et al.*, 'SiO2/TiO2 nanolayer synergistically trigger thermal absorption inflammatory responses materials for performance improvement of stepped basin solar stillnatural distiller', *Sustainable Energy Technologies and Assessments*, vol. 52, 2022.

[25] A. Naik and S. C. Satapathy, 'A comparative study of social group optimization with a few recent optimization algorithms', *Complex and Intelligent Systems*, vol. 7, no. 1, pp. 249–295, 2021.

[26] A. Naik, S. C. Satapathy, and A. Abraham, 'Modified Social Group Optimization—a meta-heuristic algorithm to solve short-term hydrothermal scheduling', *Applied Soft Computing Journal*, vol. 95, 2020.

[27] A. O. Alsaiari *et al.*, 'Applications of TiO2/Jackfruit peel nanocomposites in solar still: Experimental analysis and performance evaluation', *Case Studies in Thermal Engineering*, vol. 38, 2022.

[28] A. Ruwali, A. J. S. Kumar, K. B. Prakash, G. Sivavaraprasad, and D. V. Ratnam, 'Implementation of Hybrid Deep Learning Model (LSTM-CNN) for Ionospheric TEC Forecasting Using GPS Data', *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 6, pp. 1004–1008, 2021.

[29] A. S. Abdullah *et al.*, 'Enhancing trays solar still performance using wick finned absorber, nano-enhanced PCM', *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 12417–12430, 2022.

[30] A. S. Zamani *et al.*, 'Performance of Machine Learning and Image Processing in Plant Leaf Disease Detection', *Journal of Food Quality*, vol. 2022, 2022.

[31] A. V. N. Reddy, C. P. Krishna, and P. K. Mallick, 'An image classification framework exploring the capabilities of extreme learning machines and artificial bee colony', *Neural Computing and Applications*, vol. 32, no. 8, pp. 3079–3099, 2020.

[32] A. Yadav, S. Chatterjee, and S. M. Equeenuddin, 'Suspended sediment yield modeling in Mahanadi River, India by multi-objective optimization hybridizing artificial intelligence algorithms', *International Journal of Sediment Research*, vol. 36, no. 1, pp. 76–91, 2021.

[33] B. Abi *et al.*, 'Neutrino interaction classification with a convolutional neural network in the DUNE far detector', *Physical Review D*, vol. 102, no. 9, 2020.

[34] B. Acharjya and S. Das, 'Adoption of E-Learning During the COVID-19 Pandemic: The Moderating Role of Age and Gender', *International Journal of Web-Based Learning and Teaching Technologies*, vol. 17, no. 2, 2022.

[35] C. Banchhor and N. Srinivasu, 'Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification', *Data and Knowledge Engineering*, vol. 127, 2020.

[36] C. Sridhar, P. K. Pareek, R. Kalidoss, S. S. Jamal, P. K. Shukla, and S. J. Nuagah, 'Optimal Medical Image Size Reduction Model Creation Using Recurrent Neural Network and GenPSOWVQ', *Journal of Healthcare Engineering*, vol. 2022, 2022.

[37] D. B. V, S. P. Kodali, and N. R. Boggarapu, 'Multi-objective optimization for optimum abrasive water jet machining process parameters of Inconel718 adopting the Taguchi approach', *Multidiscipline Modeling in Materials and Structures*, vol. 16, no. 2, pp. 306–321, 2020.

[38] D. Balamurugan, S. S. Aravinth, P. C. S. Reddy, A. Rupani, and A. Manikandan, 'Multiview Objects Recognition Using Deep Learning-Based Wrap-CNN with Voting Scheme', *Neural Processing Letters*, vol. 54, no. 3, pp. 1495–1521, 2022.

[39] D. Bhattacharyya, B. P. Doppala, and N. Thirupathi Rao, 'Prediction and forecasting of persistent kidney problems using machine learning algorithms', *International Journal of Current Research and Review*, vol. 12, no. 20, pp. 134–139, 2020.

[40] D. Bhavana, K. Kishore Kumar, M. B. Chandra, P. V. Sai Krishna Bhargav, D. Joy Sanjana, and G. Mohan Gopi, 'Hand sign recognition using CNN', *International Journal of Performability Engineering*, vol. 17, no. 3, pp. 314–321, 2021.

[41] D. P. Yadav, A. Sharma, S. Athithan, A. Bhola, B. Sharma, and I. B. Dhaou, 'Hybrid SFNet Model for Bone Fracture Detection and Classification Using ML/DL', *Sensors*, vol. 22, no. 15, 2022.

[42] E. I. Ghandourah *et al.*, 'Performance assessment of a novel solar distiller with a double slope basin covered by coated wick with lanthanum cobalt oxide nanoparticles', *Case Studies in Thermal Engineering*, vol. 32, 2022.

[43] E. K. Kumar, P. V. V. Kishore, M. T. Kiran Kumar, and D. A. Kumar, '3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2 – stream CNN', *Neurocomputing*, vol. 372, pp. 40–54, 2020.

[44] E. Rajesh Kumar, K. V. S. N. Rama Rao, S. R. Nayak, and R. Chandra, 'Suicidal ideation prediction in twitter data using machine learning techniques', *Journal of Interdisciplinary Mathematics*, vol. 23, no. 1, pp. 117–125, 2020.

[45] E. S. N. Joshua, D. Bhattacharyya, M. Chakkravarthy, and H.-J. Kim, 'Lung cancer classification using squeeze and excitation convolutional neural networks with grad Cam++ class activation function', *Traitement du Signal*, vol. 38, no. 4, pp. 1103–1112, 2021.

[46] E. S. Neal Joshua, D. Bhattacharyya, M. Chakkravarthy, and Y.-C. Byun, '3D CNN with Visual Insights for Early Detection of Lung Cancer Using Gradient-Weighted Class Activation', *Journal of Healthcare Engineering*, vol. 2021, 2021.

[47] E. S. Neal Joshua, M. Chakkravarthy, and D. Bhattacharyya, 'An extensive review on lung cancer detection using machine learning techniques: A systematic study', *Revue d'Intelligence Artificielle*, vol. 34, no. 3, pp. 351–359, 2020.

[48] G. Ramkumar, R. Thandaiah Prabu, N. Phalguni Singh, and U. Maheswaran, 'Experimental analysis of brain tumor detection system using Machine learning approach', *Materials Today: Proceedings*, 2021.

[49] I. Lakshmi Mallika, D. Venkata Ratnam, S. Raman, and G. Sivavaraprasad, ‘A New Ionospheric Model for Single Frequency GNSS User Applications Using Klobuchar Model Driven by Auto Regressive Moving Average (SAKARMA) Method over Indian Region’, *IEEE Access*, vol. 8, pp. 54535–54553, 2020.

[50] J. R. K. K. Dabbakuti, A. Jacob, V. R. Veeravalli, and R. K. Kallakunta, ‘Implementation of IoT analytics ionospheric forecasting system based on machine learning and ThingSpeak’, *IET Radar, Sonar and Navigation*, vol. 14, no. 2, pp. 341–347, 2020.

[51] J. R. Reddy, A. Pandian, and C. R. Reddy, ‘An efficient learning based RFMFA technique for islanding detection scheme in distributed generation systems’, *Applied Soft Computing Journal*, vol. 96, 2020.

[52] K. K. D. Ramesh, G. Kiran Kumar, K. Swapna, D. Datta, and S. Suman Rajest, ‘A review of medical image segmentation algorithms’, *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 27, 2021.

[53] K. K. D. Ramesh, G. Kiran Kumar, K. Swapna, D. Datta, and S. Suman Rajest, ‘A review of medical image segmentation algorithms’, *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 27, 2021.

[54] K. Mannepalli, P. N. Sastry, and M. Suman, ‘Emotion recognition in speech signals using optimization based multi-SVNN classifier’, *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 2, pp. 384–397, 2022.

[55] K. N. Dattatraya and K. R. Rao, ‘Hybrid based cluster head selection for maximizing network lifetime and energy efficiency in WSN’, *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 716–726, 2022.

[56] K. Raju *et al.*, ‘A robust and accurate video watermarking system based on svd hybridation for performance assessment’, *International Journal of Engineering Trends and Technology*, vol. 68, no. 7, pp. 19–24, 2020.

[57] K. Saikumar and V. Rajesh, ‘A novel implementation heart diagnosis system based on random forest machine learning technique’, *International Journal of Pharmaceutical Research*, vol. 12, pp. 3904–3916, 2020.

[58] K. Saikumar, V. Rajesh, and B. S. Babu, ‘Heart Disease Detection Based on Feature Fusion Technique with Augmented Classification Using Deep Learning Technology’, *Traitement du Signal*, vol. 39, no. 1, pp. 31–42, 2022.

[59] K. Thirugnanasambandam, M. Rajeswari, D. Bhattacharyya, and J.-Y. Kim, ‘Directed Artificial Bee Colony algorithm with revamped search strategy to solve global numerical optimization problems’, *Automated Software Engineering*, vol. 29, no. 1, 2022.

[60] L. Goswami *et al.*, ‘A critical review on prospects of bio-refinery products from second and third generation biomasses’, *Chemical Engineering Journal*, vol. 448, 2022.

[61] L. Mallika I, D. V. Ratnam, S. Raman, and G. Sivavaraprasad, ‘Machine learning algorithm to forecast ionospheric time delays using Global Navigation satellite system observations’, *Acta Astronautica*, vol. 173, pp. 221–231, 2020.

[62] M. A. A. Al-qaness *et al.*, ‘Efficient artificial intelligence forecasting models for COVID-19 outbreak in Russia and Brazil’, *Process Safety and Environmental Protection*, vol. 149, pp. 399–409, 2021.

[63] M. Baskar, J. Ramkumar, C. Karthikeyan, V. Anbarasu, A. Balaji, and T. S. Arulananth, ‘Low rate DDoS mitigation using real-time multi threshold traffic monitoring system’, *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[64] M. K. Thota, F. H. Shajin, and P. Rajesh, ‘Survey on software defect prediction techniques’, *International Journal of Applied Science and Engineering*, vol. 17, no. 4, pp. 331–344, 2020.

[65] M. Mohammed, R. Kolapalli, N. Golla, and S. S. Maturi, ‘Prediction of rainfall using machine learning techniques’, *International Journal of Scientific and Technology Research*, vol. 9, no. 1, pp. 3236–3240, 2020.

[66] M. S. Mekala and P. Viswanathan, ‘(t,n): Sensor Stipulation with THAM Index for Smart Agriculture Decision-Making IoT System’, *Wireless Personal Communications*, vol. 111, no. 3, pp. 1909–1940, 2020.

[67] M. Y. B. Murthy, A. Koteswararao, and M. S. Babu, ‘Adaptive fuzzy deformable fusion and optimized CNN with ensemble classification for automated brain tumor diagnosis’, *Biomedical Engineering Letters*, vol. 12, no. 1, pp. 37–58, 2022.

[68] M. Z. U. Rahman, S. Surekha, K. P. Satamraju, S. S. Mirza, and A. Lay-Ekuakille, ‘A Collateral Sensor Data Sharing Framework for Decentralized Healthcare Systems’, *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27848–27857, 2021.

[69] N. Satheesh *et al.*, ‘Flow-based anomaly intrusion detection using machine learning model with software defined networking for OpenFlow network’, *Microprocessors and Microsystems*, vol. 79, 2020.

[70] N. Satheesh *et al.*, ‘Flow-based anomaly intrusion detection using machine learning model with software defined networking for OpenFlow network’, *Microprocessors and Microsystems*, vol. 79, 2020.

[71] N. V. Kimmatkar and B. Vijaya Babu, ‘Novel approach for emotion detection and stabilizing mental state by using machine learning techniques’, *Computers*, vol. 10, no. 3, 2021.

[72] N. V. Rani and K. Ravindhranath, ‘PEG-400 promoted a simple, efficient and eco-friendly synthesis of functionalized novel isoxazolyl pyrido[2,3-d]pyrimidines and their antimicrobial and anti-inflammatory activity’, *Synthetic Communications*, vol. 51, no. 8, pp. 1171–1183, 2021.

[73] N. Yuvaraj, K. Praghash, R. A. Raja, and T. Karthikeyan, ‘An Investigation of Garbage Disposal Electric Vehicles (GDEVs) Integrated with Deep Neural Networking (DNN) and Intelligent Transportation System (ITS) in Smart City Management System (SCMS)’, *Wireless Personal Communications*, vol. 123, no. 2, pp. 1733–1752, 2022.

[74] N. Yuvaraj, T. Karthikeyan, and K. Praghash, ‘An Improved Task Allocation Scheme in Serverless Computing Using Gray Wolf Optimization (GWO) Based Reinforcement Learning (RIL) Approach’, *Wireless Personal Communications*, vol. 117, no. 3, pp. 2403–2421, 2021.

[75] Naik, S. C. Satapathy, and A. Abraham, ‘Modified Social Group Optimization—a meta-heuristic algorithm to solve short-term hydrothermal scheduling’, *Applied Soft Computing Journal*, vol. 95, 2020.

[76] P. Chithaluru, F. Al-Turjman, T. Stephan, M. Kumar, and L. Mostarda, ‘Energy-efficient blockchain implementation for Cognitive Wireless Communication Networks (CWCNs)’, *Energy Reports*, vol. 7, pp. 8277–8286, 2021.

[77] P. K. Pareek *et al.*, ‘IntOPMICM: Intelligent Medical Image Size Reduction Model’, *Journal of Healthcare Engineering*, vol. 2022, 2022.

[78] P. Sharma, N. R. Moparthi, S. Namasudra, V. Shanmuganathan, and C.-H. Hsu, ‘Blockchain-based IoT architecture to secure healthcare system using identity-based encryption’, *Expert Systems*, vol. 39, no. 10, 2022.

[79] R. Janarthanan, R. U. Maheshwari, P. K. Shukla, P. K. Shukla, S. Mirjalili, and M. Kumar, ‘Intelligent detection of the PV faults based on artificial neural network and type 2 fuzzy systems’, *Energies*, vol. 14, no. 20, 2021.

[80] R. K. Mojjada, A. Yadav, A. V. Prabhu, and Y. Natarajan, ‘Machine learning models for covid-19 future forecasting’, *Materials Today: Proceedings*, 2021.

[81] S. C. Dharmadhikari, V. Gampala, C. M. Rao, S. Khasim, S. Jain, and R. Bhaskaran, ‘A smart grid incorporated with ML and IoT for a secure management system’, *Microprocessors and Microsystems*, vol. 83, 2021.

[82] S. D. M. Achanta, T. Karthikeyan, and R. Vinoth Kanna, ‘A wireless IOT system towards gait detection technique using FSR sensor and wearable IoT devices’, *International Journal of Intelligent Unmanned Systems*, vol. 8, no. 1, pp. 43–54, 2020.

[83] S. Deshmukh, K. Thirupathi Rao, and M. Shabaz, ‘Collaborative Learning Based Straggler Prevention in Large-Scale Distributed Computing Framework’, *Security and Communication Networks*, vol. 2021, 2021.

[84] S. H. Ahammad, V. Rajesh, M. Z. U. Rahman, and A. Lay-Ekuakille, ‘A Hybrid CNN-Based Segmentation and Boosting Classifier for Real Time Sensor Spinal Cord Injury Data’, *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10092–10101, 2020.

[85] S. Hira, A. Bai, and S. Hira, ‘An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images’, *Applied Intelligence*, vol. 51, no. 5, pp. 2864–2889, 2021.

[86] S. Joshi *et al.*, ‘Unified Authentication and Access Control for Future Mobile Communication-Based Lightweight IoT Systems Using

Blockchain', *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[87] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, 'A novel stacking technique for prediction of diabetes', *Computers in Biology and Medicine*, vol. 135, 2021.

[88] S. Kailasam, S. D. M. Achanta, P. Rama Koteswara Rao, R. Vatambeti, and S. Kayam, 'An IoT-based agriculture maintenance using pervasive computing with machine learning technique', *International Journal of Intelligent Computing and Cybernetics*, vol. 15, no. 2, pp. 184–197, 2022.

[89] S. Kumar, A. Jain, A. Kumar Agarwal, S. Rani, and A. Ghimire, 'Object-Based Image Retrieval Using the U-Net-Based Neural Network', *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[90] S. Mishra, L. Jena, H. K. Tripathy, and T. Gaber, 'Prioritized and predictive intelligence of things enabled waste management model in smart and sustainable environment', *PLoS ONE*, vol. 17, no. 8 August, 2022.

[91] S. N. J. Eali, D. Bhattacharyya, T. R. Nakka, and S.-P. Hong, 'A Novel Approach in Bio-Medical Image Segmentation for Analyzing Brain Cancer Images with U-NET Semantic Segmentation and TPLD Models Using SVM', *Traitement du Signal*, vol. 39, no. 2, pp. 419–430, 2022.

[92] S. Namasudra, R. Chakraborty, A. Majumder, and N. R. Moparthi, 'Securing Multimedia by Using DNA-Based Encryption in the Cloud Computing Environment', *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 16, no. 3s, 2021.

[93] S. P. Jaiprakash, M. B. Desai, C. S. Prakash, V. H. Mistry, and K. L. Radadiya, 'Low dimensional DCT and DWT feature based model for detection of image splicing and copy-move forgery', *Multimedia Tools and Applications*, vol. 79, no. 39–40, pp. 29977–30005, 2020.

[94] S. R. Dasari, S. Tondepu, L. R. Vadali, and N. Seelam, 'PEG-400 mediated an efficient eco-friendly synthesis of new isoxazolyl pyrido[2,3-d]pyrimidines and their anti-inflammatory and analgesic activity', *Synthetic Communications*, pp. 2950–2961, 2020.

[95] S. Rajasoundaran *et al.*, 'Machine learning based deep job exploration and secure transactions in virtual private cloud systems', *Computers and Security*, vol. 109, 2021.

[96] S. Rajasoundaran *et al.*, 'Secure routing with multi-watchdog construction using deep particle convolutional model for IoT based 5G wireless sensor networks', *Computer Communications*, vol. 187, pp. 71–82, 2022.

[97] S. Rajasoundaran *et al.*, 'Secure watchdog selection using intelligent key management in wireless sensor networks', *Materials Today: Proceedings*, 2021.

[98] S. Rani, D. Ghai, S. Kumar, M. V. V. P. Kantipudi, A. H. Alharbi, and M. A. Ullah, 'Efficient 3D AlexNet Architecture for Object Recognition Using Syntactic Patterns from Medical Images', *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[99] S. Routray, P. P. Malla, S. K. Sharma, S. K. Panda, and G. Palai, 'A new image denoising framework using bilateral filtering based non-subsampled shearlet transform', *Optik*, vol. 216, 2020.

[100]S. S. Saba, D. Sreelakshmi, P. Sampath Kumar, K. Sai Kumar, and S. R. Saba, 'Logistic regression machine learning algorithm on MRI brain image for fast and accurate diagnosis', *International Journal of Scientific and Technology Research*, vol. 9, no. 3, pp. 7076–7081, 2020.

[101]S. Sekar *et al.*, 'Autonomous Transaction Model for E-Commerce Management Using Blockchain Technology', *International Journal of Information Technology and Web Engineering*, vol. 17, no. 1, 2022.

[102]S. Sengan, G. R. K. Rao, O. I. Khalaf, and M. R. Babu, 'Markov mathematical analysis for comprehensive real-time data-driven in healthcare', *Mathematics in Engineering, Science and Aerospace*, vol. 12, no. 1, pp. 77–94, 2021.

[103]S. Sengan, O. I. Khalaf, P. Vidya Sagar, D. K. Sharma, L. Arokia Jesu Prabhu, and A. A. Hamad, 'Secured and privacy-based IDS for healthcare systems on e-medical data using machine learning approach', *International Journal of Reliable and Quality E-Healthcare*, vol. 11, no. 3, 2022.

[104]S. Sengan, P. Vidya Sagar, R. Ramesh, O. I. Khalaf, and R. Dhanapal, 'The optimization of reconfigured real-time datasets for improving classification performance of machine learning algorithms', *Mathematics in Engineering, Science and Aerospace*, vol. 12, no. 1, pp. 43–54, 2021.

[105]S. Stalin *et al.*, 'A Machine Learning-Based Big EEG Data Artifact Detection and Wavelet-Based Removal: An Empirical Approach', *Mathematical Problems in Engineering*, vol. 2021, 2021.

[106]S. Stalin *et al.*, 'A Machine Learning-Based Big EEG Data Artifact Detection and Wavelet-Based Removal: An Empirical Approach', *Mathematical Problems in Engineering*, vol. 2021, 2021.

[107]Sridhar, P. K. Pareek, R. Kalidoss, S. S. Jamal, P. K. Shukla, and S. J. Nuagah, 'Optimal Medical Image Size Reduction Model Creation Using Recurrent Neural Network and GenPSOWVQ', *Journal of Healthcare Engineering*, vol. 2022, 2022.

[108]T. Chakravorti and P. Satyanarayana, 'Non-linear system identification using kernel based exponentially extended random vector functional link network', *Applied Soft Computing Journal*, vol. 89, 2020.

[109]T. Kavitha *et al.*, 'Deep Learning Based Capsule Neural Network Model for Breast Cancer Diagnosis Using Mammogram Images', *Interdisciplinary Sciences – Computational Life Sciences*, vol. 14, no. 1, pp. 113–129, 2022.

[110]U. Bhimavarapu and G. Battineni, 'Skin Lesion Analysis for Melanoma Detection Using the Novel Deep Learning Model Fuzzy GC-SCNN', *Healthcare (Switzerland)*, vol. 10, no. 5, 2022.

[111]U. K. Singh, M. Jamei, M. Karbasi, A. Malik, and M. Pandey, 'Application of a modern multi-level ensemble approach for the estimation of critical shear stress in cohesive sediment mixture', *Journal of Hydrology*, vol. 607, 2022.

[112]V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, 'Prediction of brain stroke severity using machine learning', *Revue d'Intelligence Artificielle*, vol. 34, no. 6, pp. 753–761, 2020.

[113]V. Kumar *et al.*, 'Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques', *Healthcare (Switzerland)*, vol. 10, no. 7, 2022.

[114]V. Kumar *et al.*, 'Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques', *Healthcare (Switzerland)*, vol. 10, no. 7, 2022.

[115]V. N. Mandhala, D. Bhattacharyya, B. Vamsi, and N. Thirupathi Rao, 'Object detection using machine learning for visually impaired people', *International Journal of Current Research and Review*, vol. 12, no. 20, pp. 157–167, 2020.

[116]V. N. Reddy, C. P. Krishna, and P. K. Mallick, 'An image classification framework exploring the capabilities of extreme learning machines and artificial bee colony', *Neural Computing and Applications*, vol. 32, no. 8, pp. 3079–3099, 2020.

Natural Disaster Clustering Using K-Means, DBSCAN, SOM, GMM, and Mean Shift: An Analysis of Fema Disaster Statistics

Ting Tin Tin[1]*, Yap Jia Hao[2], Yong Chang Yeou[3], Lim Siew Mooi[4],
Goh Ting Yew[5], Temitope Olumide Olugbade[6] and Ali Aitizaz[7]
Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia[1]
Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2, 3, 4, 5]
University of Dundee, Dundee, United Kingdom[6]
School of Technology, Asia Pacific University, Malaysia[7]

***Abstract*—Natural disasters tend to ruin people's lives and infrastructure, which requires comprehensive analysis and understanding to inform effective disaster management and response planning. This research addresses the lack of in-depth analysis of federally declared disasters in the United States using a dataset sourced from FEMA. Through the application of unsupervised learning techniques, including K-means clustering, DBSCAN, self-organizing maps (SOM), and the Gaussian mixture model (GMM), similar types of disasters are clustered based on their frequency. The relationship between disaster type and disaster frequency is analyzed to gain insight into patterns and correlations, facilitating targeted mitigation and adaptation strategies. By using the techniques of clustering, we can accurately group similar disaster types, duration time, occurring time and location of disaster. By implementing these approaches, our study aims to improve the understanding of disaster occurrences and inform decision-making processes in disaster mitigation strategies and adaptation strategies.**

Keywords*—Natural disasters; disaster management; unsupervised learning; clustering; disaster frequency; disaster types; mitigation strategies; adaptation strategies

I. Introduction

The United States (USA) faces a wide range of natural disasters annually, including hurricanes, tornadoes, wildfires, floods, heat waves, thunderstorms, and flash floods, all of which pose significant threats to lives and cause extensive damage. For example, $ 182.5 billion was lost in Hurricane Katrina 2005 [1]. In 2022, there are a total of 119 natural disasters occurred in the United States, 52% (62 cases) of severe thunderstorms, 21.8% (26 cases) of wildfires, heat waves, and drought, and 12.6% (15 cases) of floods and flash floods [2]. In the same year 2022, 1143 tornadoes were reported with an inconsistent pattern of occurrence throughout 1995-2022, as shown in Fig. 1 [3]. A total of 466 deaths were reported in 2022 due to natural disasters, among which 33.7% (157 fatalities) were due to a tropical cyclone [4]. With these occurrences of disasters and their impacts on humans and property, disaster management comes into the picture to alleviate the suffering caused by disasters. The four main components in disaster management include mitigation, preparedness, response, and recovery. If a country could reduce the risk of loss by predicting the occurrence of disasters, it could significantly avoid unnecessary and severe consequences. Though there is research done in predicting rainfall, streamflow, etc. less research uses the real world big data set to predict the disaster using machine learning analytics algorithms [5], [6], [7], [8], [9]. Therefore, much research is necessary to predict disaster occurrence, especially using big data analytics.

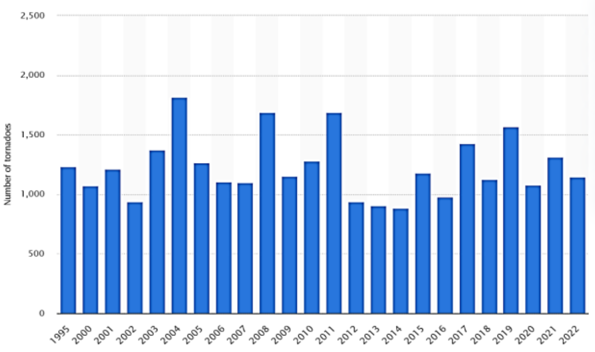


Fig. 1. Number of tornadoes reported in the USA from 1995-2022 [3].

This study uses a dataset, sourced from the Federal Emergency Management Agency (FEMA), and is regularly updated, offering a comprehensive overview of federally declared disasters since 1953 [10]. It includes data on biological disasters, notably declarations related to the ongoing Covid-19 pandemic. The data set has undergone basic cleaning and formatting measures. Additionally, a subset tailored to parameters relevant to the M5 forecast competition is provided, allowing for specific analysis and forecasting tasks related to disaster occurrences. We need to analyze these disasters more deeply, including how often they happen, what types occur, and how they affect people and places. This will help us plan better responses and ways to prevent disasters.

The remaining paper is constructed with the first overview of existing research done using different algorithms (K-Means clustering, density-based spatial clustering of Noise Applications, Self-Organising Maps, and Gaussian mixture model). This is followed by research methodologies that describe the steps to process the data set and construct forecasting models. The results and discussion are presented

*Corresponding Author.

with the content of preprocessing techniques, exploratory data analysis, robust scaler and descriptive analysis, clustering modeling, K-Means clustering, Gaussian mixture clustering, self-organizing maps, density-based spatial clustering of noise applications, and mean shift clustering. Lastly, the conclusion is presented based on the research result and discussion.

II. Literature Review

A. K-Means Clustering

Clustering techniques, particularly K-Means Clustering, play a crucial role in various domains such as customer segmentation, fraud detection, and targeted marketing. In customer segmentation, K-Means clustering enables businesses to group customers according to preferences, demographics, and purchasing behavior, facilitating the development of customized marketing strategies to meet diverse customer needs [11]. In fraud detection, clustering algorithms identify patterns in consumption habits, helping to detect potentially fraudulent activities by detecting anomalies in customer behavior and transactions [12]. Furthermore, clustering techniques are also valuable in targeting client incentives, as they allow businesses to segment customers with similar behaviors and preferences, enabling the offering of targeted incentives to encourage specific actions and increase sales or customer engagement [13]. In general, clustering techniques offer versatile solutions for understanding customer behavior, detecting fraud, and optimizing marketing strategies to improve business performance.

Chakraborty & Nagwani (2014) conducted a project that employs K-means clustering for weather forecasting, leveraging incremental K-means to enhance the model's adaptability to new data. According to the PDF, this methodology uses historical air pollution data from West Bengal, collected in 2009 and 2010, to predict weather patterns. The process involves initially applying K-means clustering to group data based on air pollutant levels such as CO2, RPM, SO2, and NOx. Each cluster represents a specific weather category defined by the maximum mean values of the pollutants within that cluster. Once the initial clusters are established, the incremental K-means algorithm is used to integrate new data into the existing clusters without re-running the entire algorithm. This approach allows for real-time updating and forecasting. For example, new pollution data for a given day is assigned to the existing cluster that it most closely matches, based on the previously computed means. This assignment helps predict the weather category for the coming days, enhancing the model’s responsiveness to changing environmental conditions [14].

Wang et al. (2018) discusses a similar application of clustering techniques, specifically for wind power prediction. Here, K-means clustering is used to categorize wind power data to improve the accuracy of forecasting. The process involves grouping historical wind power data into groups that represent different wind power levels or conditions of wind power. This clustering helps to understand the distribution and variability of wind power, helping to provide more accurate and reliable forecasting [15].

In conclusion, for both research, by grouping similar data points, these methods improve the accuracy of predictions and allow real-time updates. In the context of the weather disaster project, robust scaling ensures that the data are normalized, mitigating the influence of outliers and enhancing the performance of the K-means algorithm. This approach is crucial for accurate weather forecasting and effective disaster management, as it allows for accurate and timely predictions based on continuously updated data.

B. Density-Based Spatial Clustering of Noise Applications (DBSCAN)

Density-Based Spatial Clustering of Noise Applications (DBSCAN), a density-based clustering algorithm, has demonstrated notable effectiveness in diverse fields, with relevance in geographic data analysis and customer segmentation. DBSCAN's proficiency in analyzing geographical data, showcasing its ability to estimate population density within specific metropolitan statistical areas (MSAs) based on location data [16]. This capability has significant implications for urban planning, resource allocation, and demographic studies. Moreover, in the realm of e-Commerce and marketing, Hshan (2022) highlights DBSCAN's utility in customer segmentation, where it can group customers based on their purchasing behaviors or preferences. By leveraging DBSCAN, businesses can devise targeted marketing strategies and offer personalized recommendations, ultimately improving customer engagement and satisfaction [17].

Dey & Chakraborty (2015) conducted a project of the weather forecasting using DBSCAN that utilizes the admissions of this algorithm in finding dense clusters and the detection of outliers in spatial data, thus efficient over the normally complex data sets of weather. In weather forecasting, DBSCAN clusters data points based on their density. For example, grouping together closely packed points and marking isolated points as noise. This approach has paramount suitability for weather data, mainly consisting of dense clusters, such as high rainfall areas, and sparse outliers, such as extreme weather events. These clusters could be used by meteorologists to identify key weather patterns with the aim of establishing future weather forecasts. For instance, clusters of high humidity, combined with low pressure, could indicate the approaching of a storm, hence issuing early warnings to be better prepared in case of disasters. In connection with this, the spatial capabilities of DBSCAN allow it to deal with irregularly shaped clusters; therefore, it is very essential for weather forecasting [18].

C. Self-Organising Maps (SOM)

Self-Organizing Maps (SOM) have emerged as a valuable tool in both image clustering and customer segmentation applications. GeoSense (2023) highlights the ability of SOM to effectively group similar regions or objects within images, enabling the creation of clusters that represent distinct visual elements based on their similarities. This capability has wide-ranging applications in image analysis, from object recognition to scene understanding [19]. Similarly, in the realm of e-Commerce and marketing, Kaushik (2020) underscores the utility of SOM in customer segmentation tasks. By grouping customers according to their purchasing behaviors or preferences, SOM enables businesses to develop targeted marketing strategies and deliver personalized recommendations, thus improving customer satisfaction and engagement. The versatility of SOM in the image and customer-centric domains makes it an asset to uncover patterns and insights from complex datasets [20].

Mohan & Patil (2018) presented the deep learning-based weighted SOM to enhance the accuracy of weather and crop prediction. The SOM algorithm has been performed by the dimension of the present study so that complicated weather data can be transformed into interpretable clusters. The algorithm maps high-dimensional input data on a lower-dimensional grid while preserving the topological relationships of the data points. This provides the means to identify patterns and similarities within weather data, in order to facilitate more accurate forecasting. In the methodology, latent Dirichlet Allocation is combined with the deep neural network classifier examining, raising the modification prediction precision by up to 23% compared to existing methods [21].

For example, SOM is applied to organize weather into meaningful clusters that represent various conditions of the weather. This clustering may allow better visualization and interpretation of data for meteorologists to detect and predict weather patterns more effectively. Integration with LDA refines the data, hence improving efficiency and accuracy of the DNN classifier in predicting weather. Advanced approaches to weather prediction, such as the one mentioned earlier, which is supported by deep neural networks, enhance decision making in agriculture by allowing farmers to plan activities based on accurate weather forecasts [21].

D. Gaussian Mixture Model (GMM)

The Gaussian mixture model (GMM), as highlighted by Amy (2022), offers an effective approach to anomaly detection by identifying outliers within low-density regions of the data distribution. This capability makes GMM particularly suitable for detecting anomalies in datasets with complex or multimodal distributions [22]. On the other hand, identifying restaurant hotspots involves uncovering subgroups within the data set that can improve predictive models or improve understanding. O'Sullivan (2020) emphasizes the importance of this task in the context of restaurant analytics, where identifying hotspots can provide insight into customer preferences, demand patterns, and potential areas for business expansion or optimization [23].

Jouan et al. (2023) applied GMM to the calibration of weather forecasts contributes much to showing how the technique is utilized in clustering ensemble weather forecasts. GMM represents the distribution of the weather variables, which includes weather regimes as different kinds of distribution errors occurring in ensemble forecasts. GMM identifies clusters that reproduce weather patterns and error types in the ensemble data by fitting a mixture model. These clusters help correct for biases and increasing the accuracy of weather forecasts. There are a few steps using GMM. This GMM algorithm models the ensemble data distribution, which is variable and uncertain by nature, just as it is with any weather-related prediction by its very nature. It then identifies clusters within these data, which can be considered to be different weather regimes or error types. A separate calibration model, such as nonhomogeneous Gaussian Regression, is applied to each of the identified clusters for correcting distribution errors. In that respect, cluster-specific calibrations will ensure that accurate adjustments have been made to the forecast distribution for different weather. In this project, GMM to medium-range forecasts of temperature and wind in several locations within France. It significantly improves the interpretability and flexibility of forecasts by identifying and calibrating different kinds of errors related to each cluster [24].

Recent disaster prediction studies point to the use of cutting-edge machine learning models for surge forecasting, aiming to enhance accuracy around natural disaster prediction centers. The application of clustering techniques such as K-means, DBSCAN, Self-Organizing Maps (SOM), and Gaussian Mixture Models (GMM) on diversified disaster datasets has been emphasized in recent research. K-means effectively groups different types of disaster, making it easier to identify and analyze trends or distributions, which aids in emergency preparedness [5]. DBSCAN is a powerful tool for identifying dense regions and outliers within geographical data, enhancing the analysis of spatial disaster distributions and source identification [6]. SOMs provide accurate topological mapping and clustering of disaster-related data, optimizing visualization and interpretation despite some limitations [7]. GMM is efficient for modelling complex multimodal distributions in disaster datasets, particularly useful for anomaly detection. These advanced clustering techniques improve the understanding of disaster events, support better disaster management strategies, and ensure a faster response. The established literature on clustering and predictive modelling further underscores their effectiveness in various domains [7].

III. Research Methodology

Fig. 2 summarizes the steps used in this study to clean and transform the FEMA dataset (64092 cases, 24 variables): data preprocessing, exploratory data analysis, scaling using robust scaler, descriptive analysis, clustering modelling, visualization, analysis and conclusion. First, the data set is cleaned and transformed by dealing with missing values, empty cases, data type conversion, and parentheses removal. Once cleansed, the data set is explored using bar chart analysis, time series plot, network visualization, heatmap, frequency visualization, boxplots, violin plot, horizontal bar plot, kernel density estimation (KDE) plot and 2x2 grid of subplots. Data exploration is important to gain insight into the quality and information available. Scaling using a robust scaler is used to improve the performance of clustering models. Meanwhile, descriptive analysis is used to examine the mean, standard deviation, minimum and maximum values, 25%, 50%, 75% quartile of data, to ensure the quality of dataset before continuing cluster modelling. Five machine learning algorithms are used in clustering modelling, which are K-means, DBSCAN, self-organizing maps, Gaussian mixture model, and mean shift model. These models are visualized in 2D and 3D graphics with means as a performance indicator. Finally, silhouette scores are generated to compare five clustering models.

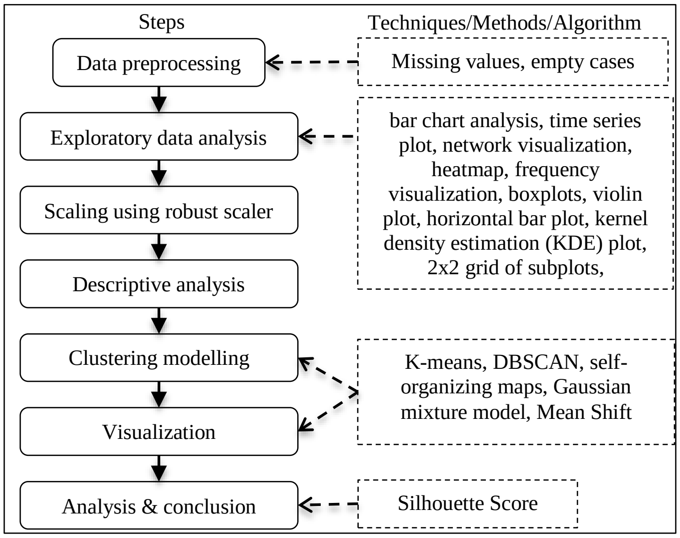
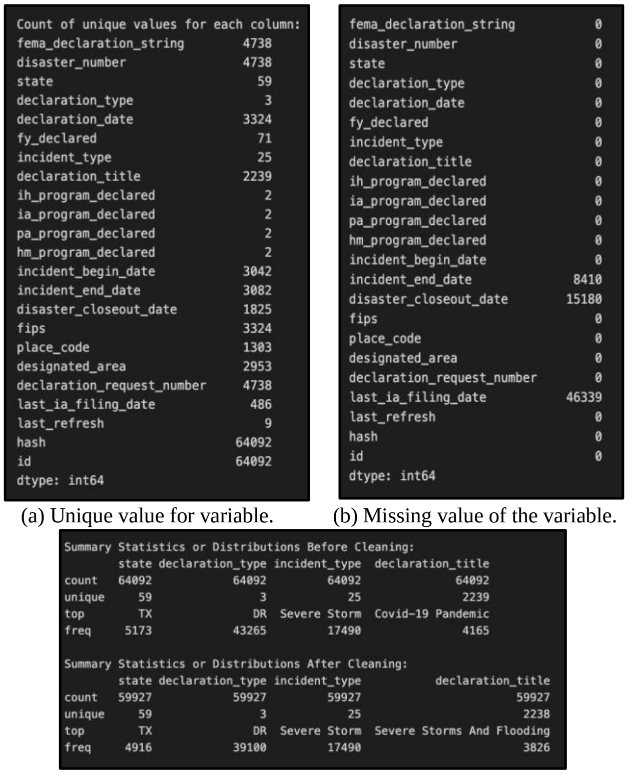

Steps
Techniques/Methods/Algorithm
Data preprocessing
Missing values, empty cases
Exploratory data analysis
bar chart analysis, time series plot, network visualization, heatmap, frequency visualization, boxplots, violin plot, horizontal bar plot, kernel density estimation (KDE) plot, 2x2 grid of subplots,
Scaling using robust scaler
Descriptive analysis
Clustering modelling
K-means, DBSCAN, self-organizing maps, Gaussian mixture model, Mean Shift
Visualization
Analysis & conclusion
Silhouette Score

Fig. 2. Steps to process the data set and construct forecasting models.

IV. Results and Discussion

A. *Preprocessing Techniques*

In this section a detailed description of the step-by-step data preprocessing with the techniques used is presented. First, as shown in Fig. 3, the unique key, the missing values of the variables, and the summary statistics are displayed to facilitate understanding of the quality of the data set. This is to prepare the dataset for the next cleaning process to accurately target the preprocessing techniques. It was found that there are 77.14% blanks in the "last_ia_filing_date" column, which makes it less reliable for research. For accuracy's sake, we suggest getting rid of rows where this column is missing values. With this method, the integrity of the data is maintained and assumptions about missing numbers are avoided. This lets us make decisions based on accurate data. We remove the columns 'last_refresh', 'hash', and 'id' from the dataset as they contain redundant or irrelevant information that does not contribute to our analysis. Then, unique keys of the data set are displayed. Examining the unique values of the 'declaration_title' column in the dataset serves to understand the variety and specific types of disaster declarations recorded. This step is to gain insight into the composition of the data, ensure consistency, and identify any anomalies or duplicates. Fig. 4 illustrates the distribution of the data points and highlights any outliers present in the data set. Outliers are data points that significantly deviate from the rest of the data and may indicate errors, anomalies, or rare events. These outliers will be removed from the data set. The violin plot illustrates the distribution of disaster numbers across different fiscal years (FY declared), the number of flips, and place codes. Provides information on the density and variability of these attributes, highlighting where most disaster occurrences are concentrated and how they vary between different categories.



```
Count of unique values for each column:
fema_declaration_string        4738
disaster_number                4738
state                            59
declaration_type                  3
declaration_date               3324
fy_declared                      71
incident_type                    25
declaration_title              2239
ih_program_declared               2
ia_program_declared               2
pa_program_declared               2
hm_program_declared               2
incident_begin_date            3042
incident_end_date              3082
disaster_closeout_date         1825
fips                           3324
place_code                     1303
designated_area                2953
declaration_request_number     4738
last_ia_filing_date             486
last_refresh                      9
hash                          64092
id                            64092
dtype: int64
```

(a) Unique value for variable.

```
fema_declaration_string           0
disaster_number                   0
state                             0
declaration_type                  0
declaration_date                  0
fy_declared                       0
incident_type                     0
declaration_title                 0
ih_program_declared               0
ia_program_declared               0
pa_program_declared               0
hm_program_declared               0
incident_begin_date               0
incident_end_date              8410
disaster_closeout_date        15180
fips                              0
place_code                        0
designated_area                   0
declaration_request_number        0
last_ia_filing_date           46339
last_refresh                      0
hash                              0
id                                0
dtype: int64
```

(b) Missing value of the variable.

```
Summary Statistics or Distributions Before Cleaning:
        state declaration_type incident_type  declaration_title
count   64092            64092         64092              64092
unique     59                3            25               2239
top        TX               DR  Severe Storm  Covid-19 Pandemic
freq     5173            43265         17490               4165

Summary Statistics or Distributions After Cleaning:
        state declaration_type incident_type           declaration_title
count   59927            59927         59927                       59927
unique     59                3            25                        2238
top        TX               DR  Severe Storm  Severe Storms And Flooding
freq     4916            39100         17490                        3826
```

(c) Summary statistics.

Fig. 3. Unique key, missing data, and summary statistics of FEMA data set.

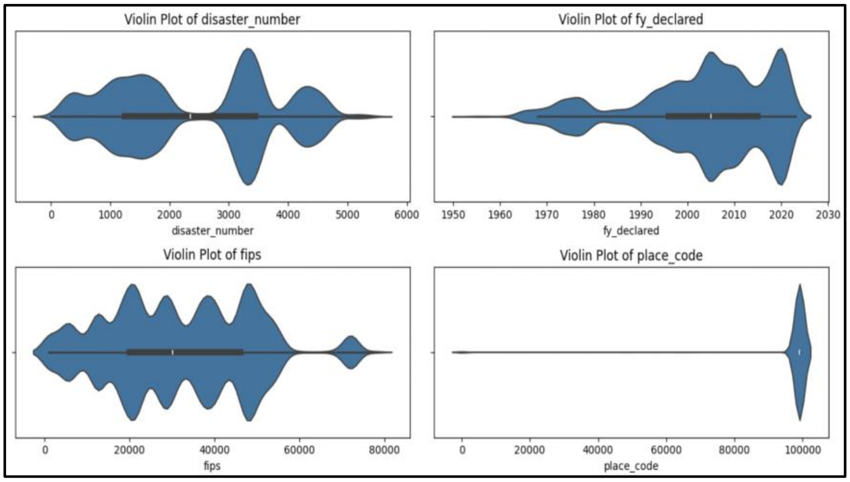


Fig. 4. Distribution of data points to analyse outliers of the data set.

Fig. 5 shows a graph of the Z scores for certain groups of numbers in a dataset. This will find out how far away a data point is from the dataset's mean by its Z-score. Using standard deviations from the mean to figure out Z scores for numerical fields lets us find outliers and see how the data are spread out. Next, rows (cases) where the incident's start date is after the end date are filtered out. This ensures logical consistency, since an incident cannot start after it ends. Filtering ensures valid data for accurate analysis and interpretation. After filtering, 59016 cases remain in the dataset. Several data transformation steps are carried out on the variables which include: 1) Convert the date columns in the data set to datetime format. 2) Convert specific columns in the data set to lowercase to mitigate potential inconsistencies due to variations in capitalization. 3) Remove the paratheses (Fig. 6) from the columns 'declaration_title' and 'designated_area'. After removal, the unique values in these columns are retrieved to observe the changes. Finally, it assigns the modified data set back to itself, although this step is optional. This process helps to clean and standardise the data, removing unnecessary information contained within parentheses. 4) Remove duplicates, which after checking the dataset, no duplicates are found.

The last step in data preprocessing is to convert the categorical values in the 'declaration_type' column to numerical representations for better analysis and modelling (Fig. 7). This is achieved by mapping specific declaration types ('dr', 'em', 'fm') to corresponding numerical values (1, 2, 3) using a predefined dictionary (declaration_type_map). After conversion, the modified data set is displayed to show the changes.

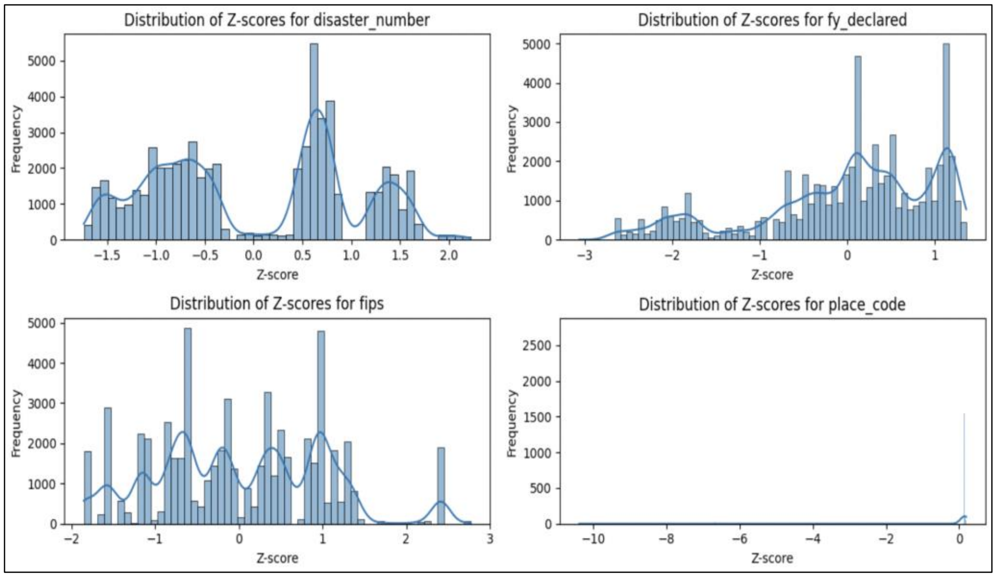
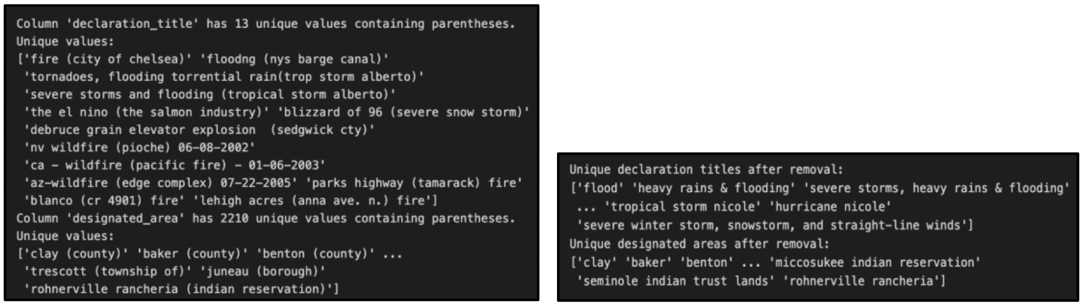


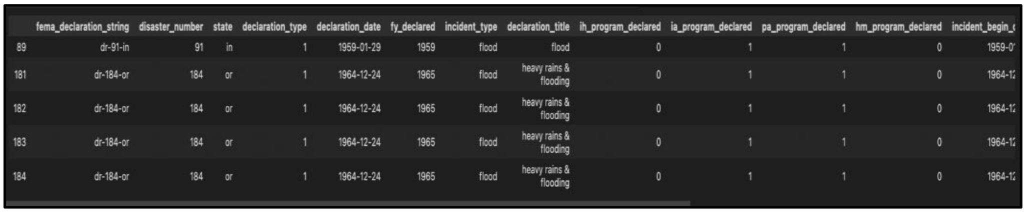
Fig. 5. Z scores of the FEMA dataset.



```
Column 'declaration_title' has 13 unique values containing parentheses.
Unique values:
['fire (city of chelsea)' 'floodng (nys barge canal)'
 'tornadoes, flooding torrential rain(trop storm alberto)'
 'severe storms and flooding (tropical storm alberto)'
 'the el nino (the salmon industry)' 'blizzard of 96 (severe snow storm)'
 'debruce grain elevator explosion  (sedgwick cty)'
 'nv wildfire (pioche) 06-08-2002'
 'ca - wildfire (pacific fire) - 01-06-2003'
 'az-wildfire (edge complex) 07-22-2005' 'parks highway (tamarack) fire'
 'blanco (cr 4901) fire' 'lehigh acres (anna ave. n.) fire']
Column 'designated_area' has 2210 unique values containing parentheses.
Unique values:
['clay (county)' 'baker (county)' 'benton (county)' ...
 'trescott (township of)' 'juneau (borough)'
 'rohnerville rancheria (indian reservation)']
```

```
Unique declaration titles after removal:
['flood' 'heavy rains & flooding' 'severe storms, heavy rains & flooding'
 ... 'tropical storm nicole' 'hurricane nicole'
 'severe winter storm, snowstorm, and straight-line winds']
Unique designated areas after removal:
['clay' 'baker' 'benton' ... 'miccosukee indian reservation'
 'seminole indian trust lands' 'rohnerville rancheria']
```

Fig. 6. Pre-processing of data in parentheses.



	fema_declaration_string	disaster_number	state	declaration_type	declaration_date	fy_declared	incident_type	declaration_title	ih_program_declared	ia_program_declared	pa_program_declared	hm_program_declared	incident_begin_
89	dr-91-in	91	in	1	1959-01-29	1959	flood	flood	0	1	1	0	1959-0
181	dr-184-or	184	or	1	1964-12-24	1965	flood	heavy rains & flooding	0	1	1	0	1964-1
182	dr-184-or	184	or	1	1964-12-24	1965	flood	heavy rains & flooding	0	1	1	0	1964-1
183	dr-184-or	184	or	1	1964-12-24	1965	flood	heavy rains & flooding	0	1	1	0	1964-1
184	dr-184-or	184	or	1	1964-12-24	1965	flood	heavy rains & flooding	0	1	1	0	1964-1

Fig. 7. Data conversion from categorical values to numerical values.

B. Exploratory Data Analysis (EDA)

Several data visualization techniques are deployed to better understand the FEMA data set, including bar chart, time series plot, network visualization, heatmap, frequency visualization, boxplots, violin plot, horizontal bar plot, kernel density estimation (KDE) plot, and 2x2 grid of subplots. First, a bar chart is created using Seaborn and matplotlib to display the count of different incident types per state, as shown in Fig. 8. Set the figure size, rotate the x-axis labels for better readability, and then show the plot. In addition, the high-volume variable states with high-specific incident types are removed. This will make the graph more visible on it. This is because high volume values will make the EDA less accurate and visible, as the bar shown will not be clear and the bars will be covered by high values. Fig. 9 shows a time series graph generated using matplotlib to visualize the count of natural disaster declarations over time, categorized by incident type. It first converts the "declaration_date" column to a date-time format. Then, it groups the data by month and incident type, calculates the count of each type, and creates a time series plot. The plot is customized with a title, labels for the x- and y-axes, a legend showing the incident types, and an adjusted layout for better visualization. Fig. 10 shows a network visualization graph where each node represents a type of disaster (for example, flood, and tornado) and adds edges between pairs of disaster types. The lines indicate relationships or connections between different types of disasters. Finally, it draws the network graph, customizing node and edge properties such as color, size, and labels, and displays the plot with a title. The network diagram is split into four parts: this will enhance the visibility and understanding of the graph with its relations. Each of the figures into 5, 5, 6, 6 types of disasters for better visibility on the relation lines. To conclude, this shows that all types of disaster will relate to each other, in other words, one disaster might trigger another incident to happen.

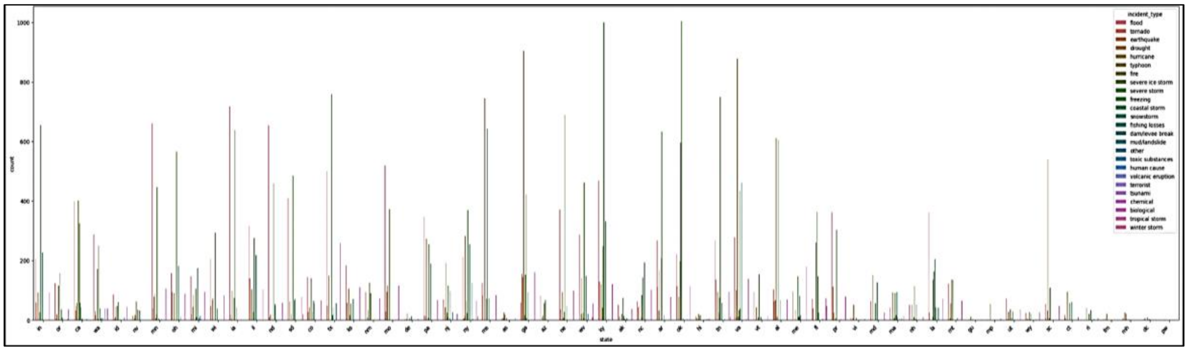
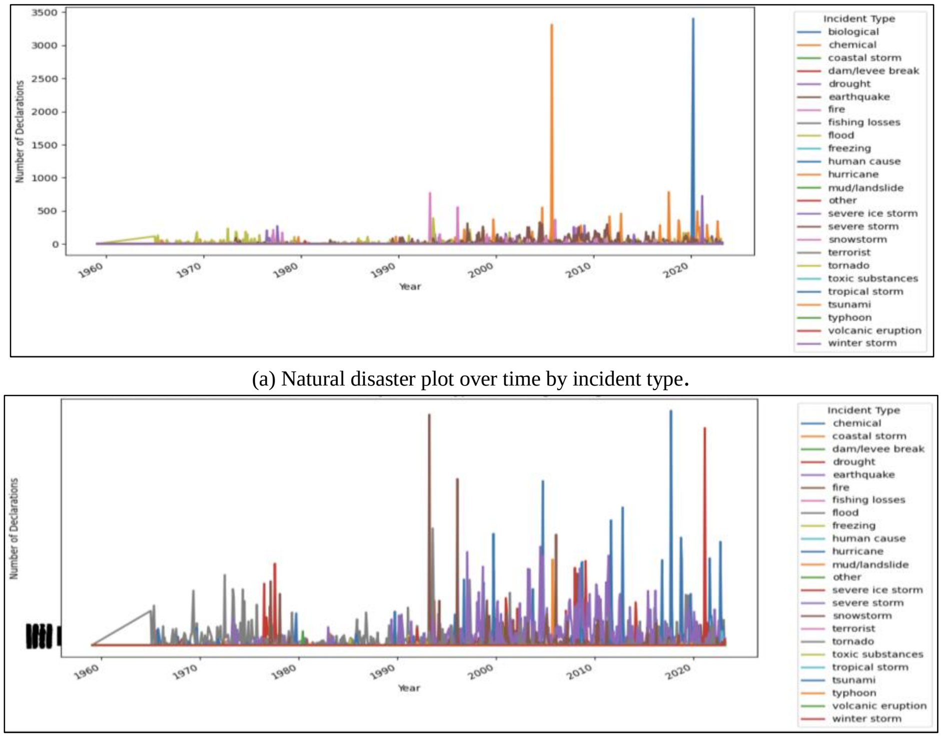


Fig. 8. Bar chart analysis using seaborn and matplotlib.



(a) Natural disaster plot over time by incident type.

(b) Natural disaster plot over time by incident type excluding "biological" and "hurricane".

Fig. 9. Time series plot using Matplotlib to visualise natural disaster over time, categorised by type.

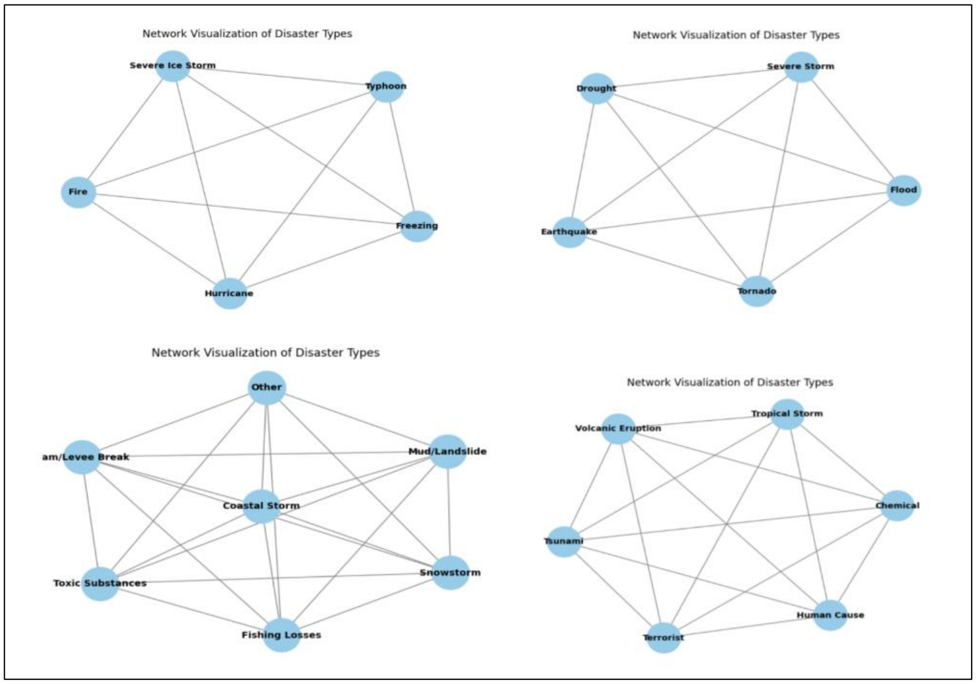


Fig. 10. Visualisation of network of different types using the network library in python.

A random relationship matrix where each cell represents the strength of the relationship between two types of disaster. The heatmap displays these values, with annotations showing the exact values, and uses a color scale (YlGnBu) to represent the strength of the relationships. The x and y axis labels represent the different types of natural disasters. In Fig. 11, it displays the heat map with a title and adjusts the layout for better visualization; each cell shows the actual values of the relationships between the corresponding pair of disasters. These values range from 0 to 1, where 0 indicates that there is no relationship and 1 indicates a perfect relationship. In Fig. 12, it first filters the data set to include only instances where the IH program was declared. Then, it creates two separate counts plots: one displaying the frequency of IH program declarations by incident type and the other showing the frequency by state. Each count plot is custom-made with appropriate titles, labels, and rotation of x-axis labels for better readability. In Fig. 13, two separate boxplots are created by calculating the duration of each incident by subtracting the incident end date from the incident start date. In Fig. 14, the variables “biological” and “fire” are removed from the outlier diagram. This is because fires and biological disasters are the outliers that will contribute the most to the outliers’ diagram. The outliers in these appear the most.

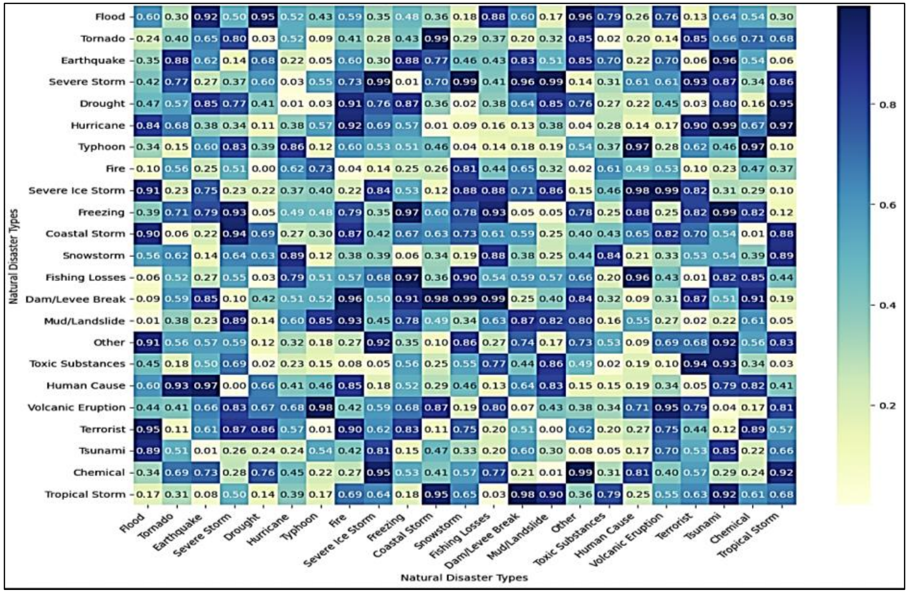


Fig. 11. Heat map revealing the relationships between types of natural disasters.

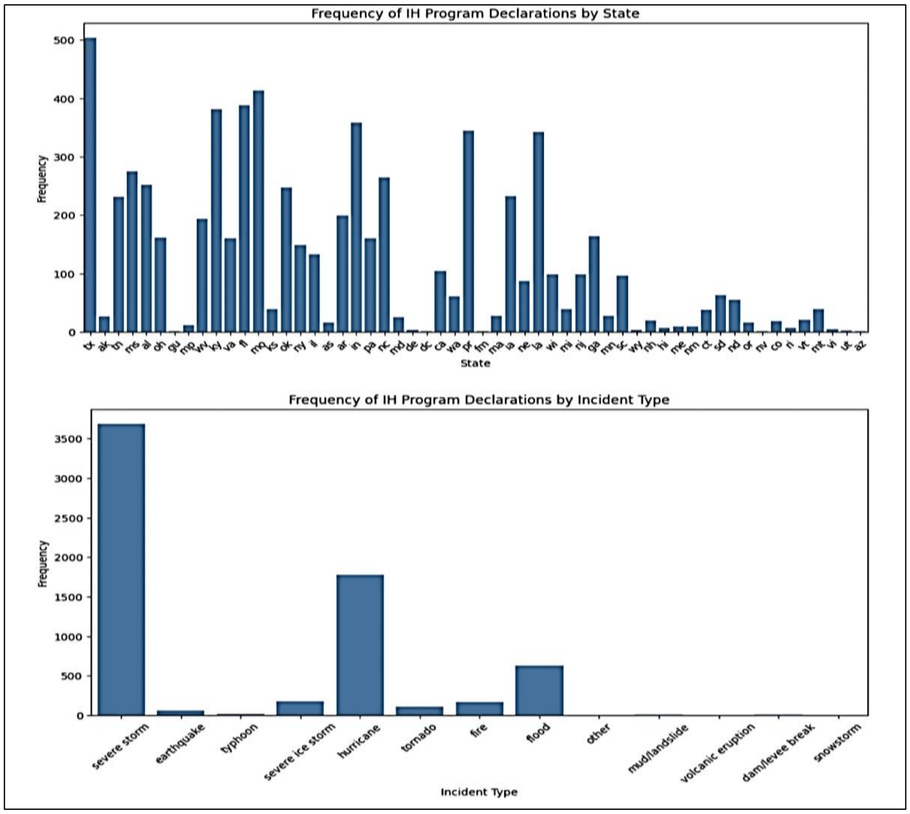


Fig. 12. Visualise the frequency of IH (Individuals and Households) programme declarations.

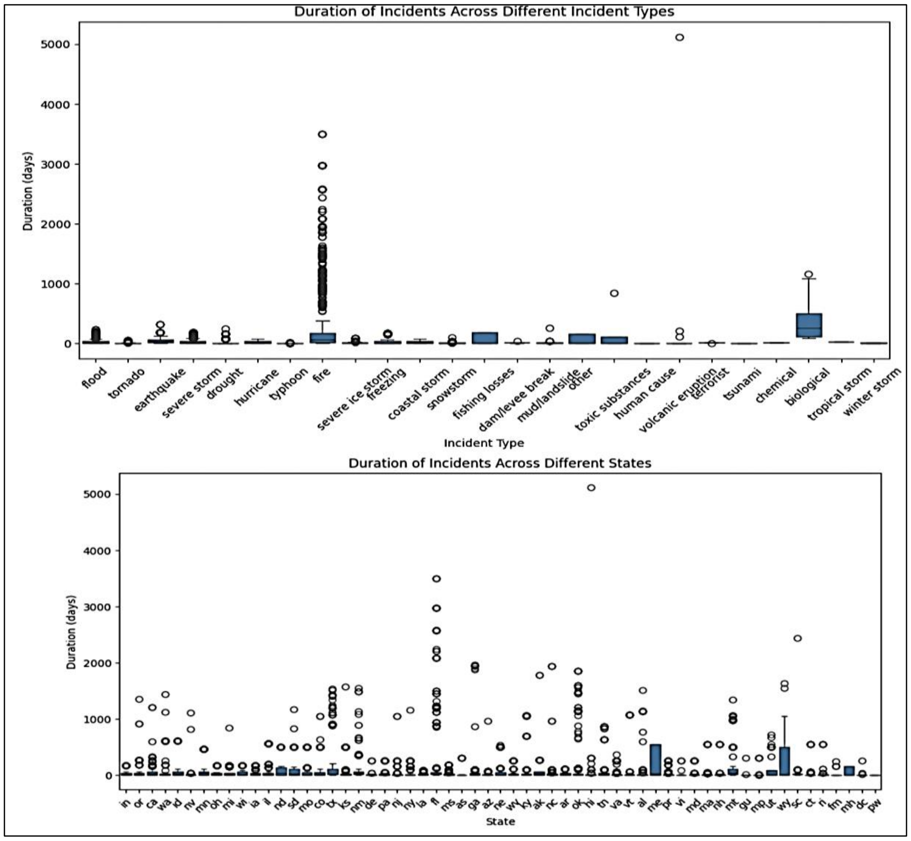


Fig. 13. Visualisation of the duration of incidents.

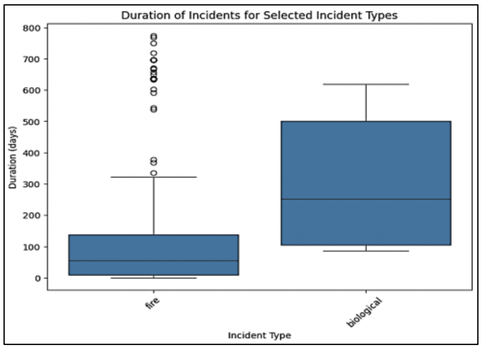
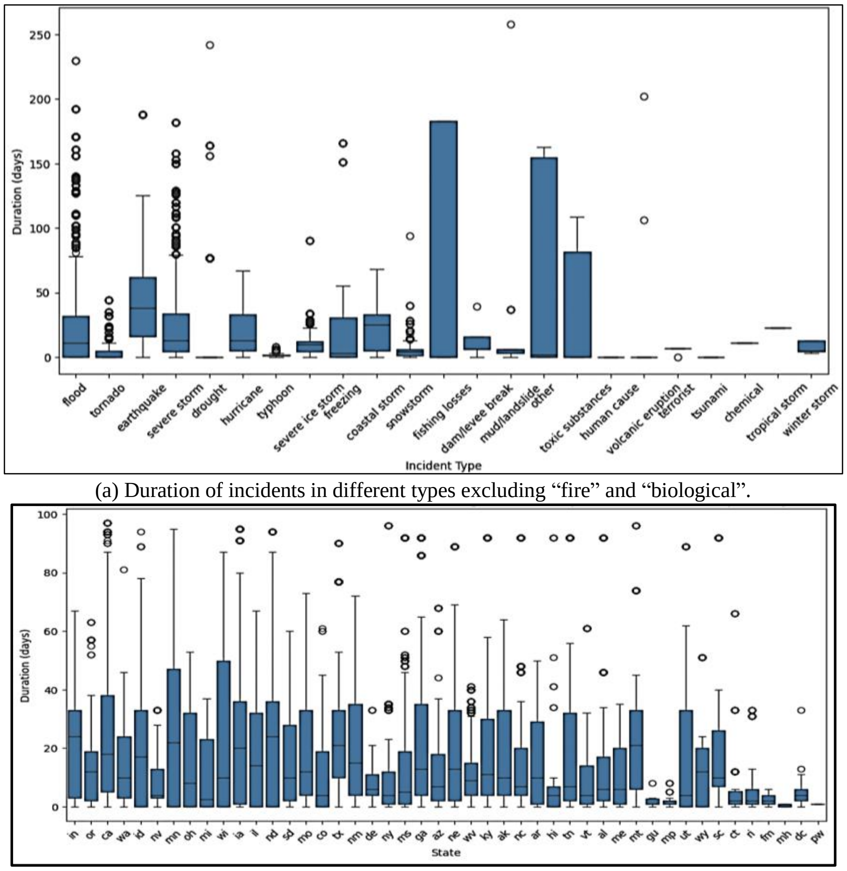


Fig. 14. The variables “biological” and “fire” are removed from the outlier diagram.

The next visualization technique as shown in Fig. 15 reveals the duration of the incident in different types and states, excluding several variables accordingly. The duration in days of each disaster reveals that the average duration is between the range of 50 and 100 days. Different states will have a duration for the incident, and it indicates that the average days for the state are around 30 to 40 days in each state.

Next, we group the data by declaration_date and count the number of disasters declared for each date. A line graph is generated to visualize the trend of the number of disasters declared over time (Fig. 16). The x-axis represents the declaration date, while the y-axis shows the corresponding count of disasters. The plot is customized with a title, axis labels, and grid lines for better interpretation. The second plot is where several years 2020 and 2008 are removed from the graph in order to make it more visible. This will make the number of disasters more clearly indicated. This is continued in Fig. 17 which is a generated violin plot where each violin represents the distribution of incident durations for a specific type of incident. Incident of ‘fire’, ‘human cause’, ‘tsunami’, ‘chemical’, ‘biological’, ‘tropical storm’ up to 3000 days, to ensure that the violet plot is more accurate in its presentation. A horizontal bar plot first calculates the frequency of each incident type using the value_counts() function. Then it creates a horizontal bar graph (kind = barh) where each type of incident is represented on the y-axis and its frequency on the x axis (Fig. 18). The time taken to close the disaster after declaration is calculated by subtracting the declaration dates from the closing dates. The resulting time differences are then converted into days. The kernel density estimation (KDE) plot is generated using seaborn to visualize the distribution of time taken for closeout. The KDE plot represents the estimated probability density function of the data. It provides information on the density of observations in different time intervals (Fig. 19). Lastly, a snippet of the 2x2 grid of subplots generated using Seaborn and Matplotlib is represented in Fig. 20. Each subplot represents the count of occurrences of Boolean values (True/False) in specific columns of a dataset.



(a) Duration of incidents in different types excluding “fire” and “biological”.

(b) Duration of incidents in different states (excluding fl, ok, va, la, pa, ks, vi, md, nh, ma, pr, as, nj).

Fig. 15. Duration of incidents in different types and states.

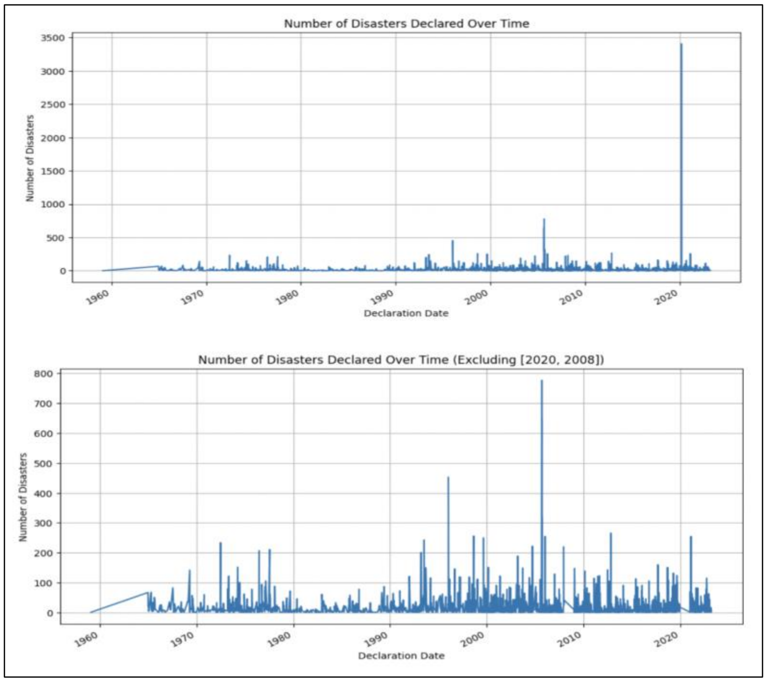


Fig. 16. Trend in the number of disasters declared over time.

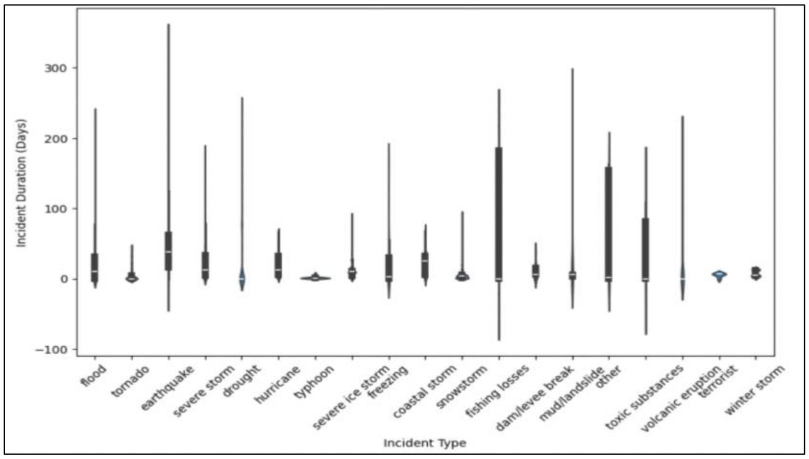


Fig. 17. Violin plot – distribution of incident types based on incident duration.

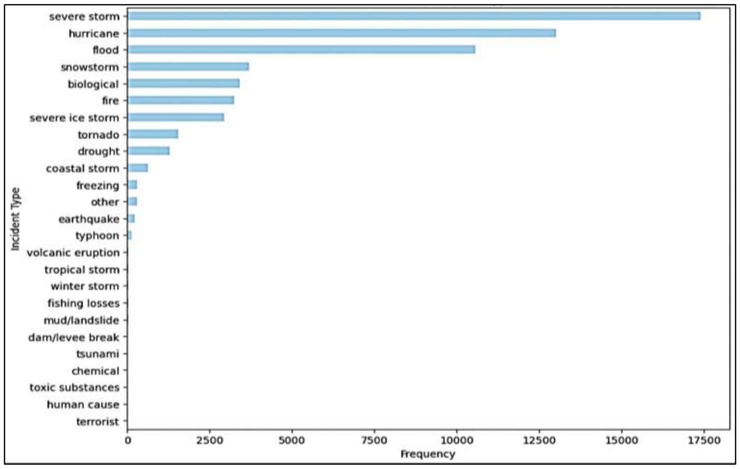


Fig. 18. Horizontal bar plot to visualise the distribution of incident types within a data set.

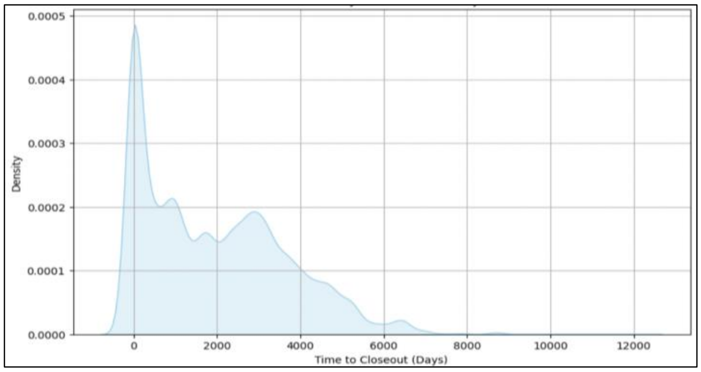


Fig. 19. Time-to-closeout analysis (Kernel density estimation).

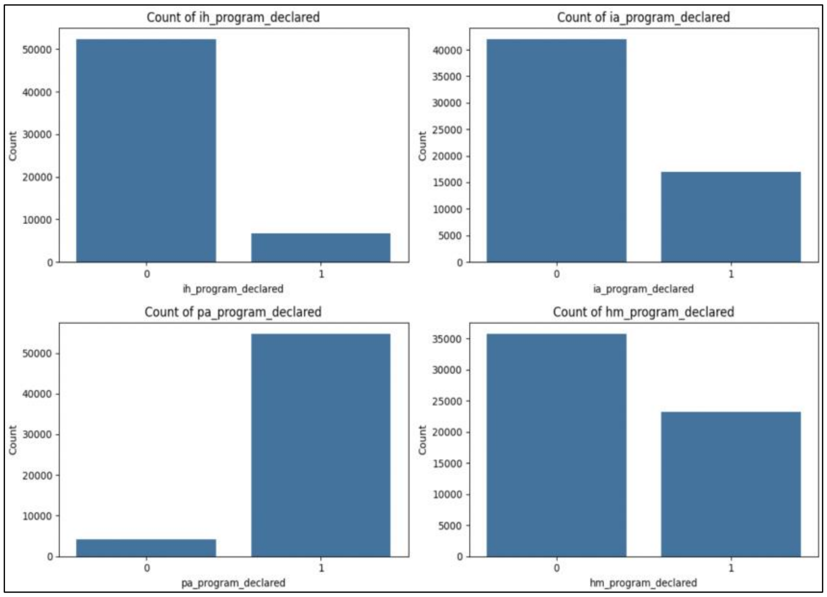


Fig. 20. 2x2 Grid of subplots for several variables.

C. Robust Scaler and Descriptive Analysis

The weather disaster data set is preprocessed using robust scaling to improve the performance of these clustering algorithms: K-means, DBSCAN, SOM, and GMM. This step is important because weather-related datasets often contain unique values, a nature that is heavily influenced by disasters. Robust scaling uses the median and interquartile range, mitigating the influence of these outliers on the central tendency of data. Moreover, normalization ensures that distances between points indicate the actual similarities of the data entries more closely and, therefore, enhance accuracy in clustering. Furthermore, it is easier to visualize and interpret normalized data for effective dissemination of results to interested parties, such as agencies concerned with disaster management.

Descriptive analysis is used to gain an initial understanding of the weather disaster dataset as shown in Table I: count, mean, standard deviation, min and max. This also involves summarizing the main characteristics of the data through statistical measures and visualizations. By performing descriptive analysis, it can identify key patterns, trends, and anomalies within the data, such as the frequency of different types of disasters, the distribution of disaster occurrences over time, and the geographical locations most affected.

TABLE I. DESCRIPTIVE ANALYSIS OF THE WEATHER DATA SET

Descriptive Statistics	Variables		
	declaration_ type	incident_ duration	state_label_ encoded
Count	58944	58944	58944
Mean	1.3666	45.2700	30.6819
SD	0.5321	132.3896	16.1234
Min	1.0000	0.0000	0.0000
25%	1.0000	3.0000	18.0000
50%	1.0000	13.0000	31.0000
75%	2.0000	33.0000	44.0000
Max	3.0000	5117.0000	58.0000

It helps in preprocessing the weather disaster dataset by providing a clear picture of the data structure and quality in this project. For example, it highlights issues such as missing values, outliers, and inconsistencies, which can then be addressed through appropriate preprocessing techniques. This foundational step ensures that the subsequent machine learning models are built on a clean dataset, enhancing their accuracy and reliability.

D. *Clustering Modelling*

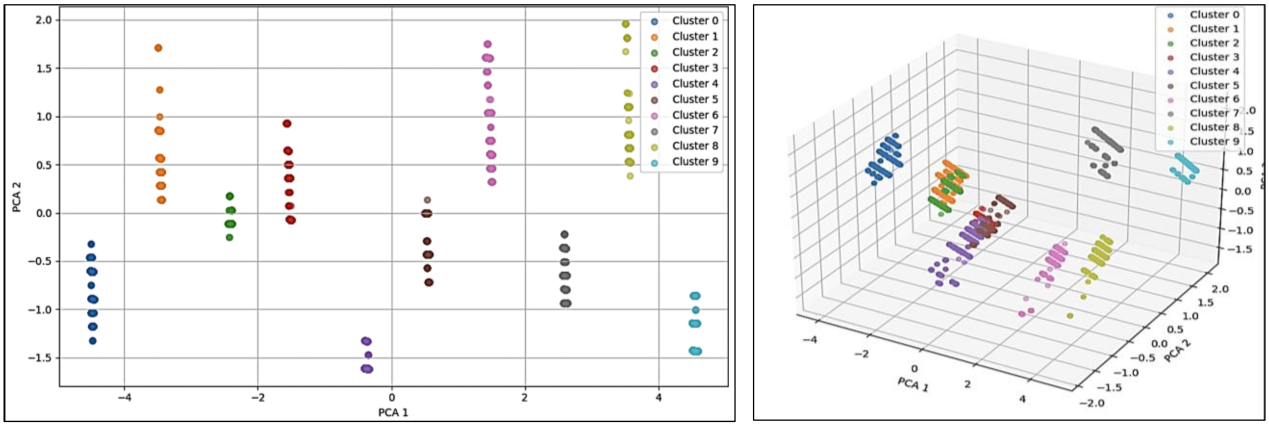
The first step involves clustering similar types of disasters based on their frequency. Each clustering method, including K-Means Clustering, DBSCAN, Self-Organizing Maps (SOM), and Gaussian Mixture Model (GMM), employs distinct algorithms and criteria for grouping data points. K-Means partitions data into K clusters by minimizing the within-cluster sum of squares within the cluster. DBSCAN identifies dense regions of points separated by sparser areas, while SOM organizes data onto a low-dimensional grid based on similarity. GMM models data distribution using a mixture of Gaussian distributions.

After clustering, it is crucial to analyze the relationship between the type of disaster and its frequency. Unsupervised learning techniques provide a means to explore this relationship without labelled data. By examining the distribution of disaster types within each cluster and the corresponding frequencies, we can gain insight into patterns and correlations. For example, certain clusters may predominantly contain hurricanes or floods with high frequencies, while others may include less frequent events such as earthquakes or biological disasters. Understanding these relationships can inform disaster preparedness and response strategies tailored to specific risk profiles.

To assess the effectiveness of each clustering method in accurately grouping similar disaster types, various evaluation metrics can be used. For example, the silhouette score, Davies-Bouldin index, or Calinski-Harabasz index can gauge the compactness and separation of clusters generated by K-Means and GMM. DBSCAN's performance can be evaluated on the number and coherence of resulting clusters. Quality can be assessed by quantization error and topographic error. By comparing these metrics across different clustering methods, we can identify the most suitable approach for our dataset and analysis objectives. By employing these unsupervised learning techniques and evaluation methods, we can gain valuable insights into the relationships between disaster types and their frequency, facilitating more informed decision making in disaster management and response planning.

E. *K-Means Clustering*

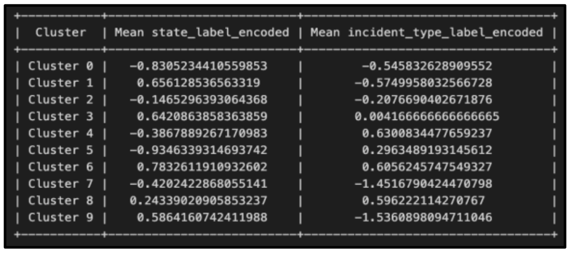
In K-means clustering, the silhouette_score method was used to calculate the suitable number of clusters used for the modelling, as shown in Fig. 21 (Principle Component Analysis (PCA)). PCA1 represents incident_type while PCA2 represents area. Each cluster represents the frequency of occurrence of each state and each type of incident. A total number of 10 clusters are displayed in two and three dimensional visuals. Furthermore, the means for all groups are calculated for evaluation to decide which model is better (Fig. 22).



(a) 2D K-means clustering (PCA visualization).

(b) 3D K-means clustering (PCA visualization)

Fig. 21. K-means clustering with 10 clusters displayed in 2D and 3D.

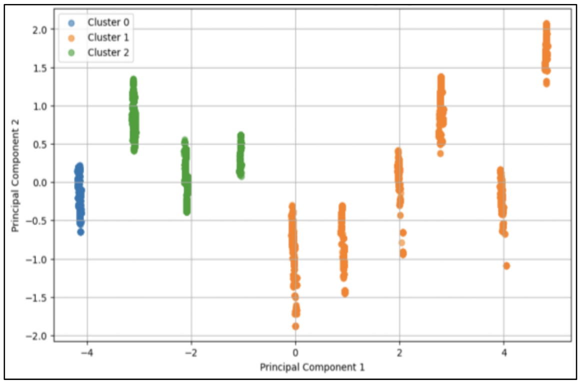


```
+-----------+------------------------+--------------------------------+
|  Cluster  | Mean state_label_encoded | Mean incident_type_label_encoded |
+-----------+------------------------+--------------------------------+
| Cluster 0 |   −0.8305234410559853    |       −0.545832628909552       |
| Cluster 1 |    0.656128536563319     |      −0.5749958032566728       |
| Cluster 2 |   −0.1465296393064368    |      −0.2076690402671876       |
| Cluster 3 |    0.6420863858363859    |      0.004166666666666665      |
| Cluster 4 |   −0.3867889267170983    |       0.6300834477659237       |
| Cluster 5 |   −0.9346339314693742    |       0.2963489193145612       |
| Cluster 6 |    0.7832611910932602    |       0.6056245747549327       |
| Cluster 7 |   −0.4202422868055141    |      −1.4516790424470798       |
| Cluster 8 |   0.24339020905853237    |       0.596222114270767        |
| Cluster 9 |    0.5864160742411988    |      −1.5360898094711046       |
+-----------+------------------------+--------------------------------+
```

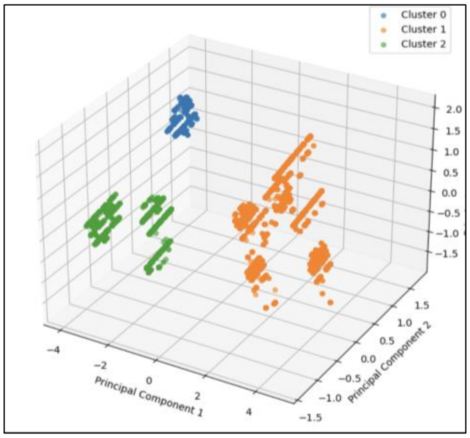
Fig. 22. Mean of clusters.

F. *Gaussian Mixture Clustering*

For Gaussian mixture clusters, 10 clusters are displayed in 2D and 3D images as the silhouette_score method is used. Fig. 23 displayed the mean; as a result, the mean is relatively lower than the K-means for the state and similar for incident type (Fig. 24).

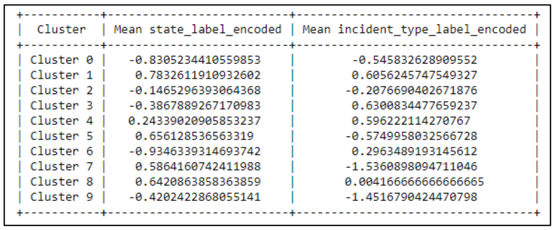


a. Clustering of the Gaussian mixture model using PCA.



b. Clustering of Gaussian mixture models using PCA (3D visualisation).

Fig. 23. Clustering of Gaussian mixture models.

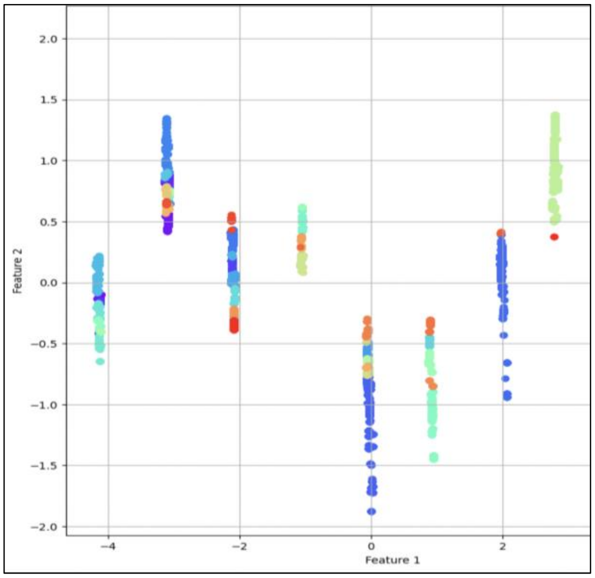


```
+-----------+--------------------------+-----------------------------------+
|  Cluster  | Mean state_label_encoded | Mean incident_type_label_encoded |
+-----------+--------------------------+-----------------------------------+
| Cluster 0 |   -0.8305234410559853    |        -0.545832628909552         |
| Cluster 1 |    0.7832611910932602    |         0.6056245747549327        |
| Cluster 2 |   -0.1465296393064368    |        -0.2076690402671876        |
| Cluster 3 |   -0.3867889267170983    |         0.6300834477659237        |
| Cluster 4 |   0.24339020905853237    |         0.596222114270767         |
| Cluster 5 |    0.656128536563319     |        -0.5749958032566728        |
| Cluster 6 |   -0.9346339314693742    |         0.2963489193145612        |
| Cluster 7 |    0.5864160742411988    |        -1.5360898094711046        |
| Cluster 8 |    0.6420863858363859    |        0.004166666666666665       |
| Cluster 9 |   -0.4202422868055141    |        -1.4516790424470798        |
+-----------+--------------------------+-----------------------------------+
```

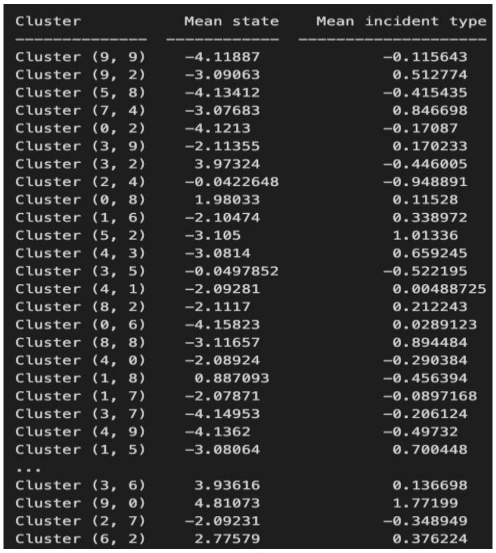
Fig. 24. Mean value of the clustering of the Gaussian mixture model.

G. *Self-Organizing Maps (SOM)*

In self-organizing map (SOM) clustering, the topology-preserving characteristics of SOM are used to organize the data into a grid of nodes, where each node represents a cluster. Unlike K-means, where cluster centers are calculated iteratively, SOM assigns data points to the nearest node in the grid, creating a topological map of the data space. The SOM clusters are visualized in 2D and mean value for each cluster is calculated to gain insight into the data distribution and cluster formations (Fig. 25).



a. SOM visualisation clusters.



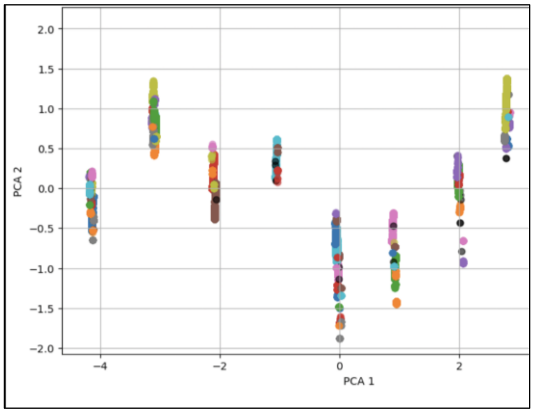
```
Cluster            Mean state    Mean incident type
--------------    ------------   --------------------
Cluster (9, 9)     -4.11887               -0.115643
Cluster (9, 2)     -3.09063                0.512774
Cluster (5, 8)     -4.13412               -0.415435
Cluster (7, 4)     -3.07683                0.846698
Cluster (0, 2)     -4.1213                -0.17087
Cluster (3, 9)     -2.11355                0.170233
Cluster (3, 2)      3.97324               -0.446005
Cluster (2, 4)     -0.0422648             -0.948891
Cluster (0, 8)      1.98033                0.11528
Cluster (1, 6)     -2.10474                0.338972
Cluster (5, 2)     -3.105                  1.01336
Cluster (4, 3)     -3.0814                 0.659245
Cluster (3, 5)     -0.0497852             -0.522195
Cluster (4, 1)     -2.09281                0.00488725
Cluster (8, 2)     -2.1117                 0.212243
Cluster (0, 6)     -4.15823                0.0289123
Cluster (8, 8)     -3.11657                0.894484
Cluster (4, 0)     -2.08924               -0.290384
Cluster (1, 8)      0.887093              -0.456394
Cluster (1, 7)     -2.07871               -0.0897168
Cluster (3, 7)     -4.14953               -0.206124
Cluster (4, 9)     -4.1362                -0.49732
Cluster (1, 5)     -3.08064                0.700448
...
Cluster (3, 6)      3.93616                0.136698
Cluster (9, 0)      4.81073                1.77199
Cluster (2, 7)     -2.09231               -0.348949
Cluster (6, 2)      2.77579                0.376224
```

b. Mean value of SOM clusters

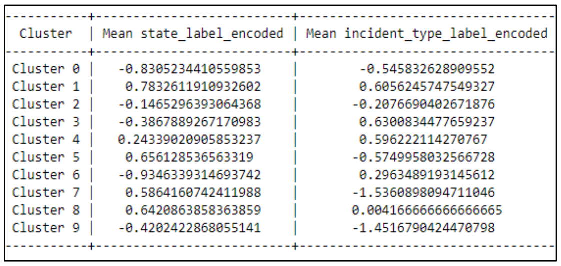
Fig. 25. Self-organising map visualisation and means value.

H. *Density-Based Spatial Clustering of Noise Applications (DBSCAN)*

For DBSCAN, 104 clusters are implemented as shown in Fig. 26 with their performance. Since there are too many clusters generated, the 3D visualization is too complicated to analyze due to overlapping clusters with unidentified clusters for each state and each incident_type.



a. DBSCAN clustering with PCA visualisation.



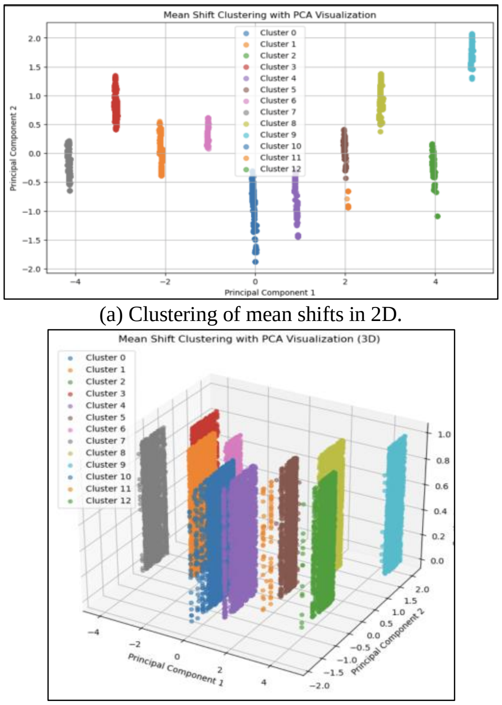
```
-----------+--------------------------+-----------------------------------
 Cluster   | Mean state_label_encoded | Mean incident_type_label_encoded
-----------+--------------------------+-----------------------------------
Cluster 0 |   -0.8305234410559853    |        -0.545832628909552
Cluster 1 |    0.7832611910932602    |         0.6056245747549327
Cluster 2 |   -0.1465296393064368    |        -0.2076690402671876
Cluster 3 |   -0.3867889267170983    |         0.6300834477659237
Cluster 4 |   0.24339020905853237    |         0.596222114270767
Cluster 5 |    0.656128536563319     |        -0.5749958032566728
Cluster 6 |   -0.9346339314693742    |         0.2963489193145612
Cluster 7 |    0.5864160742411988    |        -1.5360898094711046
Cluster 8 |    0.6420863858363859    |        0.004166666666666665
Cluster 9 |   -0.4202422868055141    |        -1.4516790424470798
-----------+--------------------------+-----------------------------------
```

b. Means of DBSCAN clustering

Fig. 26. DBSCAN clustering and its performance (means).

I. *Mean Shift Clustering*

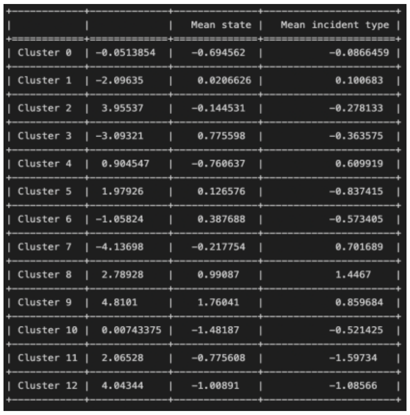
As shown in Fig. 27, a total of 10 clusters are displayed and these clusters are separated compared to DBSCAN which produces a clearer vision. The means of this model are on average negative values (Fig. 27). A cluster with a negative mean value indicates that the data points within that cluster have on average, negative values for the particular variable or feature analyzed (Fig. 28).



(a) Clustering of mean shifts in 2D.

(b) Clustering of mean shifts in 3D.

Fig. 27. Clustering of mean shifts with PCA visualisation in 2D and 3D.



```
+------------+------------+-------------+---------------------+
|            |            | Mean state | Mean incident type |
+============+============+=============+=====================+
| Cluster 0  | -0.0513854 |  -0.694562  |        -0.0866459 |
+------------+------------+-------------+---------------------+
| Cluster 1  | -2.09635   |  0.0206626 |          0.100683  |
+------------+------------+-------------+---------------------+
| Cluster 2  |  3.95537   |  -0.144531  |         -0.278133  |
+------------+------------+-------------+---------------------+
| Cluster 3  | -3.09321   |  0.775598   |         -0.363575  |
+------------+------------+-------------+---------------------+
| Cluster 4  |  0.904547  |  -0.760637  |          0.609919  |
+------------+------------+-------------+---------------------+
| Cluster 5  |  1.97926   |  0.126576   |         -0.837415  |
+------------+------------+-------------+---------------------+
| Cluster 6  | -1.05824   |  0.387688   |         -0.573405  |
+------------+------------+-------------+---------------------+
| Cluster 7  | -4.13698   |  -0.217754  |          0.701689  |
+------------+------------+-------------+---------------------+
| Cluster 8  |  2.78928   |  0.99087    |          1.4467    |
+------------+------------+-------------+---------------------+
| Cluster 9  |  4.8101    |  1.76041    |          0.859684  |
+------------+------------+-------------+---------------------+
| Cluster 10 |  0.00743375 | -1.48187   |         -0.521425  |
+------------+------------+-------------+---------------------+
| Cluster 11 |  2.06528   |  -0.775608  |         -1.59734   |
+------------+------------+-------------+---------------------+
| Cluster 12 |  4.04344   |  -1.00891   |         -1.08566   |
+------------+------------+-------------+---------------------+
```

Fig. 28. Mean value of mean shift clustering.

Based on Table II, DBSCAN has the highest silhouette score. It is the best clustering algorithm, followed by the Gaussian mixture. This is because it not only displays the firm clusters in the diagram, but also has a relatively higher silhouette score compared to other clustering algorithms. However, they are imperfect when applied in the FEMA dataset. There are some inaccuracies in the result with the clustering algorithm due to outliers in the mean calculation for state and incident. To conclude, having a definite and firm clean dataset is crucial in the clustering process, as it will provide the most accurate and identical result for analysis.

TABLE II. COMPARISON OF FIVE CLUSTERING MODELS USING THE SILHOUETTE SCORE

Model	Silhouette Score
K-Means	0.4546
Gaussian Mixture Model	0.8161
Self-Organizing Maps	0.6003
DBSCAN	0.9883
Mean Shift Model	0.4715

V. CONCLUSION

In view of the information provided on natural disaster data and after running different clustering techniques, this paper concludes that modern machine learning algorithms such as K-means, DBSCAN, SOM, and GMM have been very efficient in classifying and understanding patterns in disaster incidences. K-means identifies trends/distributions of disaster incidents and thus improves preparedness and response strategies. Spatial distribution studies, such as those related to the location of disaster sources and eventual mitigation of their impacts, are based on attributes central to DBSCAN. SOMs are a robust method for topological mapping and clustering, which can be used effectively for the visualization and better interpretation of data. GMMs are efficient ways of modelling even very complex multimodal distributions and are therefore suitable for anomaly detection in disaster datasets.

However, several limitations with this study. The accuracy of clustering results greatly depends on the dataset quality and completeness of the data set. The existence of missing values and inconsistencies within the data will greatly decrease the reliability of the result. Moreover, computational complexity and time for some clustering algorithms are highly needed mainly with the large datasets, which becomes a problem in the application.

Future work has to be directed towards enhancing data preprocessing techniques so that it can answer missing and inconsistent data more efficiently. In addition, other data sources, such as real-time satellite imagery, sensor data, etc., will add comprehensiveness and accuracy to the analysis. Computational efficiency in clustering algorithms can be achieved by parallel processing or any other advanced computing technique, which requires further research. Moreover, the development of hybrid models involving multiple clustering techniques would tie the different strengths of each technique and therefore further enhance disaster predictions. There will be a need to continue at this higher level of collaboration with disaster management agencies to ensure that the findings of such studies translate to actionable strategies that reduce the impacts of natural disasters.

ACKNOWLEDGMENT

This research is sponsored by: Tunku Abdul Rahman University of Technology and Management, INTI International University.

References

[1] V. Korhonen, "Natural disasters in the U.S. - Statistics & Facts," Statista. Accessed: Jul. 08, 2024. [Online]. Available: https://www.statista.com/topics/1714/natural-disasters/.

[2] V. Korhonen, "Number of natural disasters in the United States in 2022, by type," Statista. Accessed: Jul. 08, 2024. [Online]. Available: https://www.statista.com/statistics/216819/natural-disasters-in-the-united-states/.

[3] E. B. Salas, "Number of tornadoes in the United States from 1995 to 2022," Statista. Accessed: Jul. 08, 2024. [Online]. Available: https://www.statista.com/statistics/203682/number-of-tornadoes-in-the-us-since-1995/.

[4] V. Korhonen, "Number of fatalities due to natural disasters in the United States in 2022, by type," Statista. Accessed: Jul. 08, 2024. [Online]. Available: https://www.statista.com/statistics/216831/fatalities-due-to-natural-disasters-in-the-united-states/.

[5] P. Duraisamy and Y. Natarajan, "Twitter Disaster Prediction Using Different Deep Learning Models," SN Comput Sci, vol. 5, no. 1, p. 179, Jan. 2024, doi: 10.1007/s42979-023-02520-7.

[6] M. T. Majemite, A. Obaigbena, M. A. Dada, J. S. Oliha, and P. W. Biu, "Evaluating the role of big data in U.S. disaster mitigation and response: a geological and business perspective," Engineering Science & Technology Journal, vol. 5, no. 2, pp. 338–357, Feb. 2024, doi: 10.51594/estj.v5i2.764.

[7] T. Venkat Narayana Rao, P. Jakkam, and S. Medipally, "Future Trends and Innovations in Natural Disaster Detection Using AI and ML," 2024, pp. 110–134. doi: 10.4018/979-8-3693-2280-2.ch005.

[8] Y. Wei et al., "Comparative Analysis of Artificial Intelligence Methods for Streamflow Forecasting," IEEE Access, vol. 12, pp. 10865–10885, 2024, doi: 10.1109/ACCESS.2024.3351754.

[9] T. T. Tin, E. H. C. Sheng, L. S. Xian, L. P. Yee, and Y. S. Kit, "Machine learning classification of rainfall forecasts using Austin weather data," International Journal of Innovative Research and Scientific Studies, vol. 7, no. 2, pp. 727–741, Mar. 2024, doi: 10.53894/ijirss.v7i2.2881.

[10] Heads or Tails, "US Natural Disaster Declarations: County-level data from the Federal Emergency Management Agency: 1953 - today," U.S. Government Works. Accessed: Jul. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/headsortails/us-natural-disaster-declarations.

[11] S. C. Ipiankama, "Customer Segmentation Using K-Means Clustering." Accessed: Jul. 08, 2024. [Online]. Available: https://sampsonipiankama.medium.com/customer-segmentation-using-k-means-clustering-ae73e3d82934.

[12] S. Hu, Z. Xiao, Q. Rao, and R. Liao, "An anomaly detection model of user behavior based on similarity clustering," in 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), IEEE, Dec. 2018, pp. 835–838. doi: 10.1109/ITOEC.2018.8740748.

[13] B. Tan, "Customer Segmentation with k-Means Clustering," LinkedIn. Accessed: Jul. 08, 2024. [Online]. Available: https://www.linkedin.com/pulse/customer-segmentation-k-means-clustering-bryan-tan/.

[14] S. Chakraborty and N. K. Nagwani, "Weather Forecasting using Incremental K-means Clustering," Computers and Society, 2014.

[15] K. Wang, X. Qi, H. Liu, and J. Song, "Deep belief network based k-means cluster approach for short-term wind power forecasting," Energy, vol. 165, pp. 840–852, Dec. 2018, doi: 10.1016/j.energy.2018.09.118.

[16] R. Panchotia, "Clustering Geo-location : DBSCAN," Analytics Vidhya. Accessed: Jul. 08, 2024. [Online]. Available: https://medium.com/analytics-vidhya/clustering-geo-location-dbscan-cadb33b0442e.

[17] T. Hshan, "Demonstrating Customers Segmentation with DBSCAN Clustering Using Python," Medium. Accessed: Jul. 08, 2024. [Online]. Available: https://hshan0103.medium.com/demonstrating-customers-segmentation-with-dbscan-clustering-using-python-8a2ba0db2a2e.

[18] R. Dey and S. Chakraborty, "Convex-hull &amp; DBSCAN clustering to predict future weather," in 2015 International Conference and Workshop on Computing and Communication (IEMCON), IEEE, Oct. 2015, pp. 1–8. doi: 10.1109/IEMCON.2015.7344438.

[19] GeoSense, "Self-Organizing Maps for Sentinel 2 Image Segmentation using Python," Medium. Accessed: Jul. 08, 2024. [Online]. Available: https://geosen.medium.com/self-organizing-maps-for-sentinel-2-image-segmentation-using-python-b42cefcb32e9.

[20] Kaushik, "Self-Organizing Map (Customer Segmentation in Banking)," Analytics Vidhya. Accessed: Jul. 08, 2024. [Online]. Available: https://medium.com/analytics-vidhya/self-organizing-map-customer-segmentation-in-banking-9d7ce96bd3ec.

[21] P. Mohan and K. Patil, "Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application," International Journal of Intelligent Engineering and Systems, vol. 11, no. 4, pp. 167–176, Aug. 2018, doi: 10.22266/ijies2018.0831.17.

[22] Amy, "Gaussian Mixture Model (GMM) for Anomaly Detection," GrabNGoInfo. Accessed: Jul. 08, 2024. [Online]. Available: https://medium.com/grabngoinfo/gaussian-mixture-model-gmm-for-anomaly-detection-e8360e6f4009.

[23] C. O'Sullivan, "Identifying Restaurant Hotspots with a Gaussian Mixture Model," Towards Data Science. Accessed: Jul. 08, 2024. [Online]. Available: https://towardsdatascience.com/identifying-restaurant-hotspots-with-a-gaussian-mixture-model-2a840ab0c782.

[24] G. Jouan, A. Cuzol, V. Monbet, and G. Monnier, "Gaussian mixture models for clustering and calibration of ensemble weather forecasts," Discrete and Continuous Dynamical Systems - S, vol. 16, no. 2, pp. 309–328, 2023, doi: 10.3934/dcdss.2022037.

Impact of Read Theory in Mobile-Assisted Language Learning on Engineering Freshmen's Reading Comprehension Using BI-LSTM

E. Pearlin[1], Dr. S. Mercy Gnana Gandhi[2]
Research Scholar, Department of English, Sathyabama Institute of Science and Technology, Tamil Nadu, India[1]
Professor, Department of English, Sathyabama Institute of Science and Technology, Tamil Nadu, India[2]

Abstract—**The effect of Read Theory in Mobile-Assisted Language Learning (MALL) on reading comprehension is critical, especially for engineering freshmen who require excellent language abilities to navigate their complicated academic courses. Read Theory is a customized reading platform that offers adaptive reading activities based on the user's ability level, which is especially useful in MALL settings where accessibility and flexibility are essential. However, traditional methods of MALL have frequently faced with constraints, such as the inability to completely adapt to students' different and dynamic learning demands. This deficiency usually results in poor improvement in reading skills because the conventional paradigms do not capture the intricate and diverse learning processes that are necessary for the effective learning of languages. To fix these issues, a Deep Learning approach that involves the implementation of BI-LSTM networks for enhancing the completion's reading outcomes is offered. BI-LSTM is more suitable for this task because it has forward and backward reading capabilities to better understand and predict the dynamics of language acquisition. The research established improvement and an astonishing accuracy of 99.3%. The implementation was done using Python. This high value of accuracy disproves the common weakness of the strategy and provides convincing evidence that the proposed approach can significantly enhance MALL projects' outcomes. The specified technique, which improves on the flaws of previous approaches, does not only improve the process of reading but has the potential to revolutionize language acquisition for engineering students, making it more effective and conforming to ability.**

Keywords—Read theory; language learning; Bi-LSTM; mobile-assisted; deep learning

I. Introduction

Reading skills in a person's native language usually develop during early childhood. Acquiring reading skills in a second language can be tough and demands an alternative approach to thinking. This is why reading in English as a Foreign Language (EFL) is still being studied. In most contexts, reading and listening are understood as processes for taking in information while speaking and writing are recognized as methods for expressing information. However, newer studies have looked at how reading links to other skills, especially how it relates to writing [1]. Reading is important in people's lives because it helps them learn and stay informed about all the information around them. It becomes a regular part of the free time, school tasks, or job responsibilities. Students in universities need to pay close attention to this because they will be required to read a lot of material for their coursework. Print or digital are the two alternatives available to them. The primary goal of authors is to pique readers' interest and aid in the retention of the content. Typically, the text is enhanced and remembered by adding images or movies. Since practically everyone may now publish and distribute their work, viewers have a wide variety of texts and forms to choose from. [2]. Developing reading skills is very important because reading is a complicated and interactive task that depends on how a person understands a text. It moves back and forth between basic steps, like recognizing words and decoding them, and higher-level thinking. Reading in a foreign language is harder because everyone experiences and understands a text differently based on their own background and thinking skills. Studies show that it's helpful to use a broad approach when working with text [3].

The educational activities implemented within the framework of the initial academic year for the students who have chosen engineering, Read Theory, an online tool aimed at improving the reading comprehension abilities, explore a very effective direction to develop the language competencies with focus on the specific area of interest [4]. Read Theory is computer algorithm-based application that offers reading tests and quizzes at the appropriate level of difficulty for each student making it very effective program for enhancing the students' reading skills. Science students, especially engineering students, may spend most of their time reading technical materials and thus may not be as proficient in reading as they ought to be especially in contexts that demand deep understanding of technical and academic texts as well as industry literature [5]. Incorporation of Read Theory into MALL is proposed because with the help of Mobile-Assisted Language Learning students can practice reading exercises or take reading assessments at any time and at any place they wish by using their mobile devices. MALL is an outreach of the conventional classroom setting and refers to the use of mobile learning as a flexible means of learning a language on one's own time and pace [6].

The relevance of Read Theory in MALL is made prominent by the kind of data-focused learning that it embodies. Since students engage with the application, their reading behaviours, achievements, and challenges are identified and logged. This information can be collected to have an understanding of the dynamics in understanding and areas that need improvement regarding the comprehension of what is taught to ensure that changes that could be made to effectively teach can be made by educators [7]. In addition, extending Read Theory in MALL can promote more learner-centred context that means learners are likely to have more control of how they learn. The learners can be informed promptly of their performance, get many reading materials, and build their skills in a very systematic yet somewhat freeway. It can be inferred from the above findings that for engineering freshmen particularly, who rely on their capacity to rapidly put into effect and decode technical content for class performance, advantages of applying Read Theory in MALL are vast. Students should also learn to improve their general performance in class, comprehension of class content and improve their skills in interpreting texts. This can help them in the future employment in the engineering field where the clarity of understanding as well as the ability of conveying that understanding becomes crucial [8]. Most engineering students go through a tight outline of subjects that are technical that do not let them practice reading a lot or even do so with much effort [9]. Apparently, the given competence of understanding rather complex texts is important for their further academic and work experience. Several programs have been named earlier: Read Theory with the built-in option of adaptive learning that allows the learner to focus on reading exercises with the relevant difficulty level and a great choice of texts. When it is used through mobile environment, it becomes more convenient since the students can be involved with the reading exercises at any spare time as they wait for their classes, during transport among other times hence making reading part and parcel of their everyday lives [10].

Thus, the efficiency of Read Theory in a mobile frame can be defined by the regular interest of students; steadily increased level of comprehension of the texts read; and flexibility to the time limitations of students. Mobile devices can prove beneficial when learning is being done at a variety of locations including the college, homes or while on the move. This goes hand in hand with increased chances of going through the material, which is beneficial in check and balance learning session and to reinforce what has been learnt. In addition, the mentioned environment is highly flexible and it provides students with a number of opportunities regarding feedback; it is possible to provide immediate feedback in the context of a mobile environment so that students could see the results of their efforts and alter their approach in case of it is necessary [11]. The post Sherpa use of Read theory desired outcome is also measurable by comparing pre and post-test where, if the scores were to rise significantly after the inclusion of the mobile assisted use of Read theory, then the desired outcome would have been achieved. Also, questionnaire or survey with the students in terms of their experience with the particular developed platform can afford prose description of its usefulness as well as its effectiveness in enhancing student learning. To the engineering students the deployment of Read Theory in the mobile context means that those overloaded with multiple difficult subjects can use it as the supplement to the standard approaches. It is not only a method of developing the skills in reading comprehension but also a current method of constant learning. Making use of this instrument, students have the chance to foster gradually their independent reading-skills, respectively instrumental reading-skills which are important for the engagement with more advanced academic texts [12]. The key contribution of the work.

- Utilizes BI-LSTM models to increase the accuracy of reading comprehension prediction, delivering enhanced insights into the usefulness of mobile-assisted learning technologies.
- Implements Read Theory inside a mobile-assisted learning framework to assess its influence on engineering freshmen's reading comprehension, adding to the practical use of technology in education.
- Assesses improvements in reading comprehension scores before and after the intervention, providing measurable proof of the efficacy of mobile educational platforms.
- Provides empirical data on the benefits and problems of using Read Theory in a mobile learning environment, contributing to the corpus of knowledge in educational technology.
- Improves the design and effectiveness of mobile-assisted learning tools by combining user input and performance data, resulting in more effective instructional tactics and resources.

The study is structured as Section I provides an introduction of the research, and Section II has Related Work which reviews the existing literature. Problem statement is given in Section III. Section IV details the data collection methods, preprocessing steps, and analytical approaches, including the use of BI-LSTM models and Section V presents the Result and Discussion and ends with Conclusion and future works in Section VI.

II. Related Work

Many adults learning a second language or a foreign language don’t have enough chances to practice. However, learning a new language well is important for many reasons. It helps them fit in quickly in a new country and adjust easily to new jobs or schools. This means they need more help to learn their second language successfully thinking that advancements in learning analytics, self-directed language learning, and mobile language learning will enhance this support. MALLAS, which stands for mobile-assisted language learning through learning analytics for self-regulated learning, is a novel concept that is presented in this study. It is intended to assist individuals who plan educational initiatives in assisting folks acquiring a second language. This is accomplished by utilizing learning data to assist students in better managing their own learning. Here, the MALLAS framework is demonstrated as a tool to aid with the application of mobile-assisted language learning in a particular scenario. Researchers who wish to learn more about and support language learners who utilize mobile devices for self-study may also find it helpful. A possible problem with the MALLAS framework is that it depends a lot on how good the learning analytics tools are. These tools can have problems if the data is not high quality or if it’s not understood correctly. Also, it may not consider the different ways people learn and what motivates them, which could make it less effective for various groups of learners [13].

In this study, Mortazavi et al. [14] looked at various factors that affect how MALL can help improve speaking and listening skills. 100 research papers were chosen from the best journals about how MALL affects language learning in higher education. A selection of eight papers was made according to particular criteria to consolidate findings related to language abilities and technology concepts. So, after carefully looking at the suggested methods and comparing them, explaining the basic beliefs about this issue and provide complete and lasting solutions to deal with it. This analysis showed that people in developing countries use mobile devices a lot for learning. The main area where technology helps is vocabulary, and it is giving good results. According to this survey, university students prefer to use WhatsApp for chatting and writing, and LINE for listening and reading comprehension. Teachers should consider using the TAM to modify their present and future lesson plans. This means they can improve students' learning by using more than just in-person classes [14].

Mobile technologies and the reasons that affect EFL learners' desire to use them in their studies are important. This study looks at how factors like support, ease of use, and usefulness affect Iranian EFL learners' views on MALL. Data were gathered from 223 Iranian students learning English as a foreign language, who took their classes in two main places: public schools and private institutes. In the first part of the study, a meaningful connection between three important factors and what learners think about MALL is found. Also, the support available was linked to how easy learners felt using it. In the second part of the study, the results showed that how useful people think MALL is really affects their views about it. This study also shows that having good support makes it easier for learners to use tools. The results are talked about based on other studies, and ideas for more research are given [15].

In this study, university students learning EFL used mobile devices to help them learn. The research focused on understanding how students feel about using mobile phones for language learning, how they use these tools, and how helpful they are for their studies. The findings demonstrated that students utilize different MALL software for distinct objectives. Five important criteria that influenced students' adoption of MALL software were identified from a study conducted among 581 students from Indian colleges and universities. These factors are individual desires and reasons, how easy the software is to use, technology features, social influence, and how useful the students think the software is. These factors impacted students' readiness to use the software, their motivation, and ultimately their performance. The study showed how prepared and motivated students are affected by the way they use MALL and the results they see from it. The importance of good language learning results, the ideas added to the theory, and practical uses for managers from the study are talked about. Some language abilities, such as speaking and listening, require more practice due to hardware restrictions, especially for mobile activities that may be improved [16].

This study looked at how using mobile apps for language learning can help university students who are learning English improve their skills and support their ability to learn on their own. It also aimed to help them become more skilled with technology. The global epidemic and spread of COVID-19 have accelerated the usage of technology in schools because to the urgent necessity to keep education going. As digital technology has advanced, many smart mobile apps are now used for learning and teaching English. Using these technologies can help language learners practice learning on their own outside of the classroom. This is important for staying motivated and becoming independent learners. Mobile phone apps can be very helpful for teaching because they are easy to use, available to many people, and have various features. There hasn't been much research on using mobile phones to help people learn languages on their own outside of school. In this research, university students studying English utilized their smartphones to facilitate independent learning outside the classroom. This made their learning more lasting and independent. The findings of this study showed that students had a positive attitude towards using mobile devices to help them learn English. The results show that using mobile apps for learning languages can increase students' motivation and make their learning more enjoyable and lasting compared to traditional teaching methods. People in this study said that the biggest advantages of using mobile learning for studying English were easy access to learning materials, the ability to carry their tools anywhere, the freedom to learn on their own, better interaction with others, and feeling more confident in their English skills. The study talks about how to teach better and suggests ways to create a learning environment that fits with the changes brought by technology. Learners plan to use mobile phone apps to learn English in the future. This shows that these apps can be useful for helping them learn on their own and make their learning more effective over time [17].

Mobile-assisted language learning (MALL) has had a significant impact on language learning by encouraging self-directed learning and providing practice opportunities outside of traditional classroom settings MALL has shown have proved particularly effective for vocabulary acquisition and are widely used in developing countries, where mobile technology facilitates language learning [18]. Key factors affecting acceptance of MALL tools are personal motivation, ease of use, and perceived usefulness, which directly affect learning outcomes and motivation Students’ perceptions of MALL are also influenced by the support and usefulness of technology a they get it and it shows [19]. The COVID-19 pandemic reaffirmed the value of MALL, as it increased student motivation and enabled a more flexible and consistent learning experience compared to traditional methods but the potential of MALL was well realized, challenges such as ensuring data quality in academic research and retrieving a variety of materials must meet student preferences.

III. Problem Statement

The study aims to fill the identified knowledge gap in the analysed literature, especially when it comes to adapting learning analytics practices to broaden the topical area of MALL and foster self-regulated learning strategy among adult second language learners. In the prior research, the effects of introducing the use of mobile technologies to support second language acquisition have been researched but there is a lack of research in the best ways of using learning analytics to facilitate support and feedback in a mobile learning environment [20]. Besides, there is lack of research evidence on how frameworks such as MALLAS can be used to address the needs of the learners given the different learning styles, motivation, and quality of data collected.

IV. PROPSOED METHODOLOGY FOR IMPACT OF READ THEORY IN MOBILE-ASSISTED LANGUAGE LEARNING ON ENGINEERING FRESHMEN'S READING COMPREHENSION USING BI-LSTM

Read Theory in MALL provides an interactive environment that improves reading comprehension through tailored practice. Adapting content to learners' competence levels, Read Theory promotes focused skill development and engagement. The study makes use of Bi-LSTM networks to evaluate the impact of Read Theory on engineering freshman’s studying comprehension. Bi-LSTM networks are well-acceptable for capturing contextual relationships in sequential data due to their ability to process information in each forward and backward directions. This bidirectional method complements the model’s functionality to apprehend the nuances of language and context, making it effective for analyzing comprehension. Bi-LSTM on data collected from Read Theory's platform, focusing on studying comprehension tasks and assessments. The model can be evaluated based on its capacity to assume enhancements in comprehension ranges and usual common performance. Leveraging Bi-LSTM networks, the research offer insights into the efficacy of Read Theory in enhancing reading skills and making a contribution to the optimization of MALL programs for students as shown in the Fig. 1.

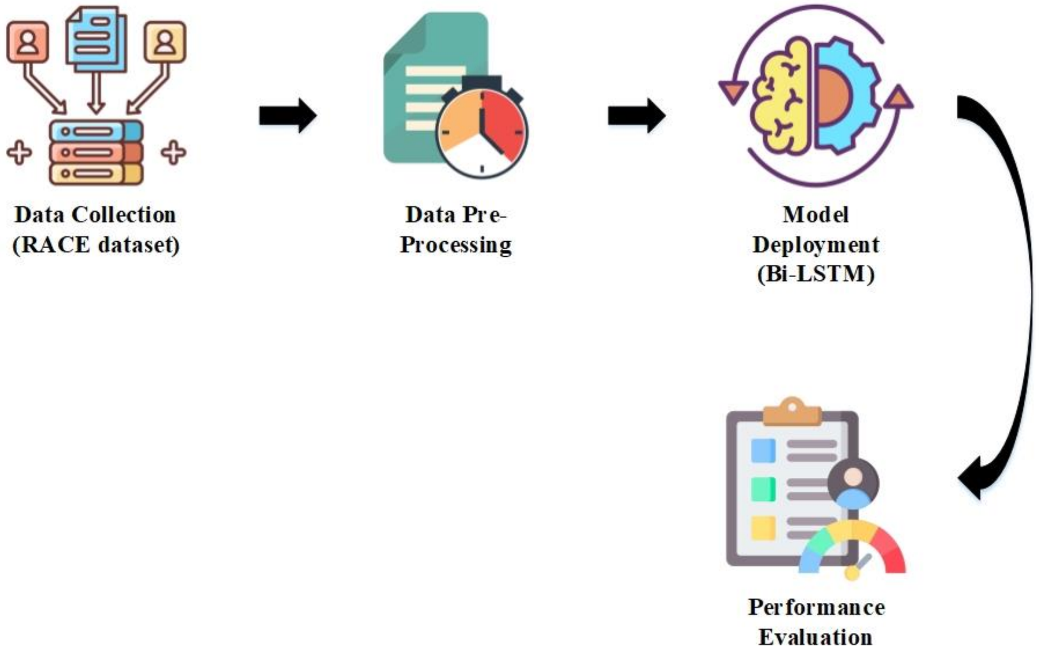


Fig. 1. Block diagram of the proposed study.

The study examines the potential of the mobile learning tool known as Read Theory to contribute to development of freshmen’s engineering reading skills. To examine the impact of Read Theory on students’ ability to understand the difficult texts, the study employs deep learning technique referred to as the Bi-LSTM networks. The study procedure consists of three critical stages: data acquisition which captures students’ details and their reading literacy levels, and data pre-cleaning which sequentially purges the data that it harvested from the internet. Subsequently, an enhanced model in the form of Bi-LSTM is developed and applied in order to estimate students’ reading comprehension according to their progress in Read Theory. Lastly, the ability of the Bi-LSTM model to predict reading comprehension results is assessed. Through such extensive research, the study aims at providing useful information on the effectiveness of Read Theory for enhancing comprehension of freshmen in the engineering field hence help in the development of better language learning techniques.

A. Data Collection

The RACE [21] dataset which is an analyzing dataset comprising of more than 28000 passages and nearly 100000 questions extracted from the English examination conducted in China is used. The forms are designed especially for center and high school students and, therefore, will make the dataset rather useful for measuring reading comprehension. It offers a vast amount of textual information and related questions that can be used as schooling and test sets for the system comprehension models. Using this dataset, analyze the degree to which the Read Theory platform impacts the studying comprehension of engineering newbies within a MALL environment. The collection of passages and questions in the RACE dataset will create a strong basis for the assessment of students’ analyzing skills, and the results will offer an understanding of how the adaptive learning can benefit language education and understanding potential for students in academic context.

B. Data Pre-Processing

Data preprocessing is an initial step in data analysis and ML that emphasizes preparing the raw data for cleaning, transformation, and proper formatting. This process includes numerous key activities: It includes steps like cleaning data by dealing with missing values, errors and inconsistencies; normalize or standardize the information to get certain degree of uniformity and code the categorical data in number form, compatible to the models. It is also equipped with characteristic extraction and selection to determine the maximum eligible variables for the analysis. Data preprocessing increases the purity of data and performance of device by correcting that the record of data is accurate, complete and arranged in the proper format. By performing feature preprocessing, more accurate results are obtained and, therefore, more likely to define the desired version, and is a non-trivial process for any works based on data analysis.

1) Text normalization: Text cleaning is one of the key steps in the preparation of data in the field of NLP that brings raw text data in a specific structure for analysis. Changing every textual content to lower case is also important in order to standardize the text since it is easier to compare lower cases to lower cases from the increased amount of whitespace removed from the text that does not add any value to the meaning of the text. Furthermore, normalization includes enhancement of some occasional contractions to their full bureaucracy, normalization of Unicode character to illustrating equivalent, and erasing normal words like such as those adding semantics. Tokenization process divides the textual content into individual words or tokens which can be lemmatized into their base form so as to reduce variation. Applying these approaches, the textual content is preprocessed: it is normalized and previously analyzed or trained in models, and similarly prepared for further analysis.

2) Tokenization: Tokenization is the procedure of dividing textual content into smaller units, known as tokens, which may be character words, phrases, or symbols. This phase is vital in NLP because it converts raw textual input into a structured format that can then be easily evaluated and processed. Splitting the text into tokens, permits algorithms to handle and understand textual statistics greater efficiently, permitting responsibilities which include parsing, evaluation, and modeling. Tokenization simplifies the textual content, making it feasible to carry out operations like counting word frequencies, analyzing sentence structures, and making use of the system mastering model.

$$Tokens = text.split(delimiter) \tag{1}$$

Where $text$ is the original string of text that will be tokenized and $delimiter$ is the character or collection of characters used to divide the text.

C. Bi-LSTM

Bi-LSTM networks are enhanced LSTM networks that capture time-dependent complexity in sequential data without dissimilarity. Standard LSTM networks, sequential in one direction. In the future, Bi-LSTMs process data in both forward and backward orientations. These two modes can integrate past and future knowledge about a specific event in the sequence to provide a thorough comprehension in the context of the mediation. A Bi-LSTM network consists of two LSTM layers: one that processes the input sequence from beginning to end, and another that processes it from end to beginning. The outputs of these two layers are then combined to provide a more complete representation of the data. For a given sequence $X = (x_1, x_2, x_3 \ldots., x_T)$, where T is the length of the sequence, the forward LSTM creates hidden states $\overrightarrow{h_t}$. The forward LSTM creates hidden states at each time step t, whereas the reverse LSTM generates hidden states at each time step $\overleftarrow{h_t}$. The composite representation for each time step (t) is obtained by concatenating the two hidden states:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{2}$$

In Eq. (2), h_t represents output at time step , which capture the data from both the direction of the sequences. The Input Gate (i_t) controls the quantity of incoming data that is fed into the cell state. The Forget Gate (f_t) regulates the quantity of previous cell state that should be retained. The output gate (o_t) determines the quantity of cell state that will be output.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{3}$$

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{4}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{5}$$

$$C_t = tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{7}$$

$$h_t = o_t * \tanh(C_t) \tag{8}$$

where the hyperbolic tangent function in Eq. (3), Eq. (4), Eq. (5), Eq. (6), Eq. (7) and Eq. (8) is represented by $tanh$, the sigmoid function by σ, and element-wise multiplication by $*$. Because of these equations, LSTMs may better express complex temporal relationships by keeping a consistent gradient across long time periods.

The Fig. 2 illustrates the architecture of the Bi-LSTM model and briefly describes its work. As an advanced type of LSTM network, the Bidirectional LSTM network analyzes the data sequences both in the backward manner and the forward manner, thus, collects contextual information from both the past and the future for each time step. The forward LSTM scans the sequence from start to end while the reverse LSTM scans from the end to the start. The output of these two LSTM layers are then concatenated together to form a new vector representation for each time step. This bidirectional method enhances the possibility of the model to learn and predict based on the whole context of the input. On the whole, the graphic represents the LSTM cells of both, the forward and backward passes, while the arrows demonstrate the information flow across the layers. Bi-LSTMs are useful in tasks that require exploitable amount of context depth such as language modeling, text classification and sequence prediction, where information flow is bidirectional. This structure is especially useful when the correct forecast depends on context data of both past and future time periods. BiLSTMs are often superior to regular LSTMs in hardnesses that require understanding context not only from the previous but the following data input, and they also can discover more sophisticated relations in sequential signs than unidirectional LSTMs. BiLSTMs can be used in any general purpose, for natural language processing and time series analysis and machine translation.

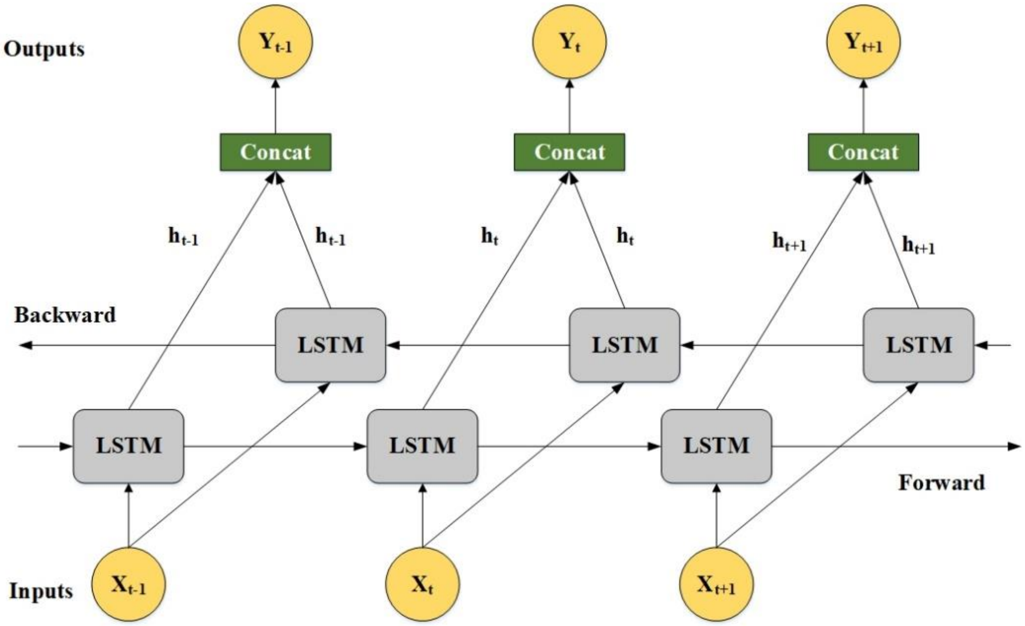


Fig. 2. Architecture of Bi-LSTM.

Bi-LSTM are crucial to investigate and technologically impact the reading theory in MALL for freshmen's reading comprehension. Bi-LSTM interfaces are particularly important in this work, for identifying and applying applications from past and future in contexts, under all comprehending reading tasks. Compared with the single-dimensional model, the Bi-LSTM network can receive textual input in two dimensions that it is more likely to analyse reading and estimating the student comprehension level than the single-dimension model. This bimodal processing assists the communicator to pick the fine details of the language and the context. It also enhances the model's accuracy in forecasting comprehension rates based on engagement with the read theory.

```
Algorithm 1: Reading Comprehension Using BI-LSTM
Step 1   Data Collection
             • Input the RACE Dataset
Step 2   Data Pre-Processing
             • Clean the data
             • Normalize the text data
             • Tokenize the text into smaller units
               $Tokens = text.split(delimiter)$
         Divide the text
             • Lemmatization
Step 3   Model Training
         //Define the architecture of the Bi-LSTM model
               Given sequence $X = (x_1, x_2, \dots, x_T)$
               hidden states $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$
         //Initialize Model Parameters
             //Forward LSTM Layer
             Input Gate $(i_t) = \sigma(W_i.[h_{t-1}, x_t] + b_i)$
             Output Gate $(o_t) = \sigma(W_o.[h_{t-1}, x_t] + b_o)$
             //Calculate Cell State
             Cell State $(C_t) = tanh(W_C.[h_{t-1}, x_t] + b_C)$
             $C_t = f_t * C_{t-1} + i_t * C_t$
             Hidden State $(h_t) = o_t * tanh(C_t)$
         Train the model using the training data
Step 4   Model Evaluation
             • Make Predictions on the test data
             • Calculate model accuracy
Step 5   Model Deployment
             • Evaluate the model
```

In practice, the Bi-LSTM network will be used to assess sequences of student interactions with Read Theory content, such as their replies to comprehension questions and feedback received during the learning process. This investigation seeks to uncover trends and insights into how the Read Theory platform enhances reading comprehension abilities. Using the Bi-LSTM's capacity to integrate information from both sides of the text sequence, the research may gain a better grasp of how various parts of the learning material contribute to comprehension growth. Furthermore, the Bi-LSTM's ability to simulate long-range dependencies within text sequences is critical for capturing the intricate links between many aspects of reading comprehension, such as recognizing context, drawing conclusions, and storing knowledge over time. This capacity will allow the research to more properly and thoroughly assess the success of the Read Theory intervention. Finally, the application of Bi-LSTM networks will give significant insights into the efficacy of mobile-assisted learning tools, help to optimize instructional tactics, and improve the overall impact of MALL platforms on engineering students' academic achievement.

V. Result and Discussion

In the result and discussion section, examining the findings from applying Read Theory inside a MALL framework to engineering freshmen, specializing in studying comprehension upgrades facilitated by using Bi-LSTM networks are done. This section delves into how the gaining knowledge of Read Theory affects comprehension outcomes, inspecting the effectiveness of the Bi-LSTM model in capturing and enhancing reading abilities. The interpretation the statistics to evaluate the realistic implications of these findings, comparing them with existing literature and discussing their relevance for academic techniques in engineering disciplines. Insights gained will highlight the impact of MALL tools on academic performance.

A. *Training and Testing*

The Fig. 3 shows a line graph illustrating the training and testing accuracy of a machine learning model over 100 epochs. The x-axis represents the variety of epochs, at the same time as the y-axis shows the accuracy value, starting from zero. The training accuracy, which typically increase as the model learns from the training data. The testing accuracy, which measures the model's performance on unseen data. Initially, the training accuracy hastily increases, at the same time as the testing accuracy indicates a greater slow upward push. However, after round 60 epochs, the training accuracy plateaus, and the testing accuracy starts to decrease slightly. This indicates that the model might be overfitting the training information, learning its patterns too well and struggling to learn new, unseen examples. Overall, the graph indicates that the model achieves a reasonable level of accuracy on both training and testing data.

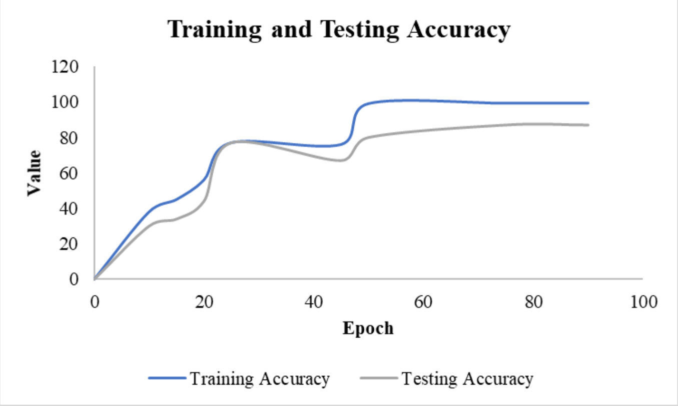


Fig. 3. Training and testing accuracy.

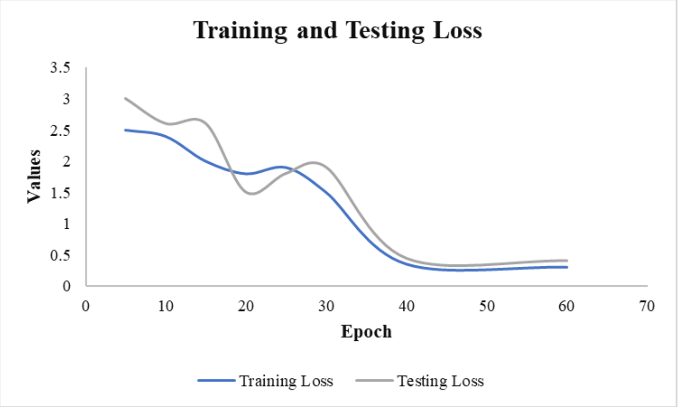


Fig. 4. Training and testing loss.

Fig. 4 shows a line graph illustrating the training and testing loss of a machine learning model over 60 epochs. The x-axis represents the range of epochs, at the same time as the y-axis shows the loss value, starting from zero. The training loss, which commonly decreases because the model learns from the training data. The testing loss, which measures the model's performance on unseen statistics. Initially, both training and testing loss decrease unexpectedly, suggesting that the model is learning efficiently. However, after around 30 epochs, the training loss continues to lower, while the testing loss begins to plateau or even increase slightly. This indicates that the model might be overfitting the training data, learning its patterns too well and struggling to generalize to new, unseen examples. Overall, the graph shows that the model achieves a reasonable level of overall performance.

B. *Performance Metrics*

Performance metrics are quantitative values applied in the assessment of how effectively a model meets intended objectives. These include recall that measures the model's performance in selecting all the actual positives from the entire dataset; accuracy which gives an overall measure of the number of correct predictions; precision that quantifies the number of actual positives per hundred fine predictions to demonstrate the capacity of the version in avoiding false positives; the F1 score that is an integrated measure of recall and precision that provides a balanced measure of performance. Hence, every metric provides information about particular aspects of the model's performance. Thus, for the given classes of precise programs, practitioners could use the above points and make knowledgeable selections about the suitability of the proposed metric and guarantee that the system beneath thought of does fulfill the performance degrees wanted for the preferred solution of the intended problems.

1) Accuracy: Accuracy is one of the evaluation metrics that define the portion of instances that has been classified correctly in a given set. That is because it provides a full picture of a model's performance for categorization tasks. The formula for accuracy is shown in Eq. (9),

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{9}$$

2) Precision: An overall performance indicator called precision assesses how well the model predicts the future with any degree of accuracy. It calculates the percentage of actual positive results among all cases the model has categorized as positive. Because precision is focused on the quality of positive predictions, it is most useful in situations when the cost of false positives is substantial. The formula for precision is shown in Eq. (10),

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{10}$$

3) Recall: The efficiency of pattern recognition for future predictions is evaluated using an overall performance score known as precision. It provides information on actual rate of positives, based on the totality of the cases which the model has classified to have positive results. Since precision measures the accuracy of positive instances, the measure is most handy in applications where false positives are very costly. It is given in Eq. (11).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ negatives} \tag{11}$$

4) F1 score: The F1 score is an all-around performance statistic that gives a fair assessment of a model's accuracy in classification tasks by integrating recall and precision into a single measurement. When working with unbalanced datasets—where accuracy and recall may differ significantly it is helpful. The F1 score is a means to assess a model's performance by taking into account both false positives and false negatives. It is calculated as the harmonic mean of accuracy and recall. The formulation for the F1 score is given in Eq. (12),

$$F1\ score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \tag{12}$$

Table I show that the study obtained great overall performance in comparing the effectiveness of a model. Accuracy at 99.3% suggests that the model is successful in showcasing its general correctness across all predictions. Precision, at 99.1%, measures the share of actual positive predictions amongst all high-quality identifications made by the model, highlighting its reliability in minimizing false positives. Recall, at 98% indicates the model’s ability to find true positives among all cases. Recall, indicates what percentage of positive cases exist in a set of records. It provided an F1 score of 98. 9% is used to balance the precision and recall at once, which gives rather comprehensive information about the model on the whole. This F1 score shows how well the model reduces false positives and how reliable it is, that is how good it is at identifying positive cases. Cumulatively, those measurements show that the version outperforms others in terms of all the measurements which include accuracy, precision, recall, and f1 score making the model extremely useful for the intended purpose. The model’s capability in realizing such excessive confirmed the tool’s potential in providing practical predictions in practical conditions.

TABLE I. PERFORMANCE METRICS

Metrics	Efficiency
Accuracy	99.3%
Precision	99.1%
Recall	98.7%
F1 score	98.9%

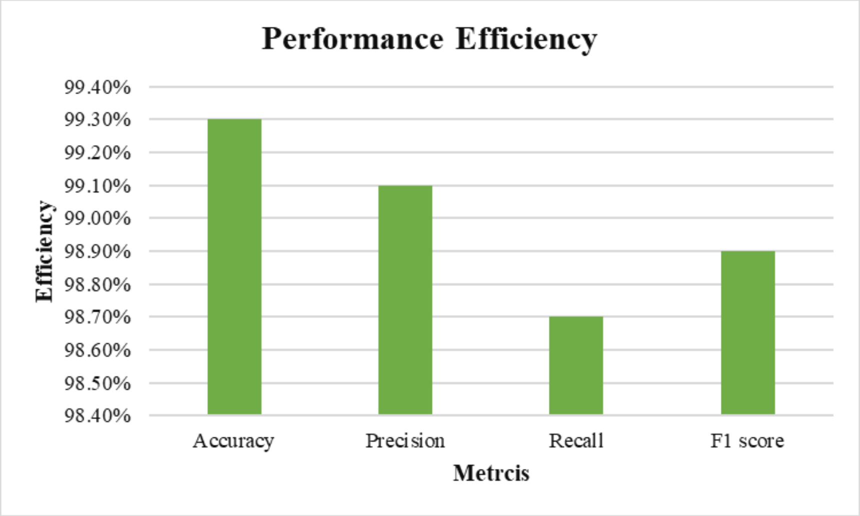


Fig. 5. Performance efficiency of the Bi-LSTM on the proposed study.

Fig. 5 depicts a model's performance parameters, with an emphasis on metrics. The graphic reveals that Accuracy is the highest of the measures, a little over 99.30%, showing that the model's predictions are generally true. Precision follows closely, barely below 99.20%, indicating the model's ability to reliably identify real positive situations while limiting false positives. Recall is significantly lower, slightly around 99.00%, implying that, while the model is reliable and exact, it may miss some actual positive cases, showing a modest trade-off in sensitivity. The F1 Score falls between accuracy and recall, just above 98.90%, indicating a balance between the two measures and reflecting the model's overall strength in classification tasks. The figure’s representation of these metrics emphasizes the model's efficiency in performing classification jobs with high accuracy and precision while keeping an acceptable balance with recall, resulting in a strong F1 score. This figure highlights the model's great performance, particularly in circumstances where accuracy and precision are important, but also illustrates the importance of recall in situations when detecting all positive instances is critical.

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH DIFFERENT METHOD

Method	Accuracy	Precision	Recall	F1 Score
Decision Tree [22]	95.0%	94.5%	96.0%	95.2%
Random Forest [23]	92.3%	91.8%	93.5%	92.6%
KNN [24]	98.0%	97.5%	98.6%	98.0%
SVM [25]	89.5%	88.0%	90.2%	89.1%
Proposed Method	99.3%	99.1%	98.7%	98.9%

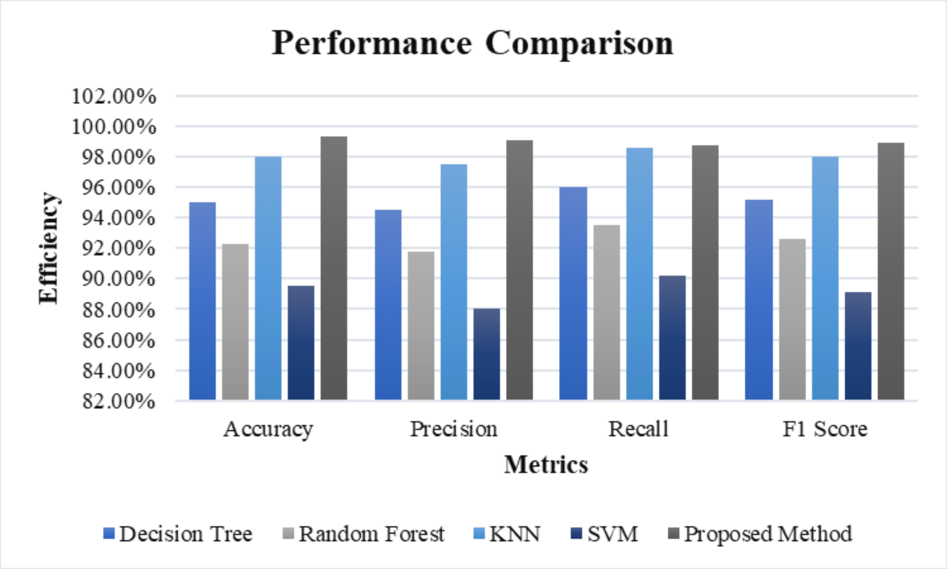


Fig. 6. Performance comparison of the proposed method with different methods.

A comparison of the methodologies across several measures is shown in Table II. The decision tree approach has 95.0% accuracy, 94.5% precision, 96.0% recall, and a 95.2% F1 score. This indicates a strong balance between identifying positive cases and reducing false positives. The Random Forest performs low, with 92.3% accuracy, 91.8% accuracy, 93.5% recall, and 92.6% F1 score, with improved classification but slightly less robust than decision trees K-Nearest Neighbors (KNN) technique achieves 98.0% accuracy, 97.5% precision, 98.6% recall, and 98.0% F1 score, indicating that it performs well in finding positive cases while minimizing error. In contrast, the Support Vector Machine (SVM) method does not perform well, with 89.5% accuracy, 88.0% accuracy, 90.2% recall, and 89.1% F1 score, indicating its low efficiency in comparison. The proposed method outperforms the others by the highest metrics: 99.3% accuracy, 99.1% precision, 98.7% recall, and 98.9% F1 score, demonstrating good performance at all evaluation criteria. This suggests that the suggested strategy gives an accurate and dependable solution to the categorization problem.

Fig. 6 compares the efficiency of five distinct methods such as Decision Tree, Random Forest, KNN, SVM, and the Proposed Method using four major performance metrics. The chart shows that the Proposed Method regularly beats the other approaches, with the highest values in all parameters, including accuracy slightly above 99%, precision just under 100%, recall around 98.7%, and an F1 score close to 99%. KNN also performs well, especially in accuracy and recall, where it outperforms all other approaches save the suggested method. Decision Tree performs well, with accuracy and recall metrics close to 95%, but it falls slightly short in precision and F1 score. Random Forest has balanced performance across all measures but stays lower than the top performers, with values ranging from 92-94%. SVM has the lowest metrics, particularly in accuracy and recall, demonstrating that it fails to compete with the other techniques' efficiency. This comparison demonstrates that the suggested technique outperforms all measures, making it the most trustworthy alternative for classification jobs in this context.

C. Discussion

The limitation of previous research in MALL, on improving reading comprehension, often starts from their lack of ability to evolve to the various learning desires of students. Earlier approaches and methods have failed to deliver in terms of dealing with the dynamic nature of the problem, particularly in the context of specific learning pathways. Most of these methods rarely achieved an optimal blend of accuracy and flexibility which, in turn, affected the performance levels and the progression of students' reading skills. The following shortcomings have been identified with the conventional approach of analysis: The proposed method that uses BI-LSTM networks eliminates these shortcomings. The processing of sequential data in both forward and backward directions, make BI-LSTM capture the requirement of language learning more effectively than the basic models of AI. This allows the model to capture more of the context inherent in language and hence leads to improved prediction and in extension reading comprehension. This high level of accuracy underlines the ability of the strategy to compensate for the problems that have been identified in the previous investigations. The technique enhances the reading comprehension result at the same time establishing a standard of feasibility for the subsequent MALL applications. Such an approach can be successfully spread to other spheres of language learning making the approach more valuable and effective.

VI. Conclusion and Future Work

Combining Read Theory with Mobile Assisted Language Learning or MALL enhances the first-year engineering students' reading comprehension using BI-LSTM networks. The 99. The achievement of 99.3% accuracy, proved that BI-LSTM is effective for dealing with restrictions that traditional MALL methods have. Previous strategies hardly take into account students' evolving and diverse needs for learning and this has led to poor improvement in reading skills. In fact, due to the bidirectional processing of sequential data, the technique providing a more enlightened picture of language acquisition as well as a more personalized learning model to individuals, BI-LSTM is capable to accomplish the idea. It not only set very high standards for MALL systems but also improves the ways in which understanding-oriented results are interpreted, thus showing how state-of-the-art AI techniques have the potential to dramatically transform teaching resources. The outcomes of the study make suggestions for the use and extension of BI-LSTM in educational technologies and highlight that, given that language proficiency is high, the usefulness of this technique could be of great value in engineering courses. The high reliability of this method highlights how the learning process as well as the classroom setting can be enhanced for better performance.

Future studies should examine how well the BI-LSTM model scales and adapts to various academic situations and topic areas. The efficacy and usefulness of BI-LSTM might be further improved by looking at how it integrates with other adaptive learning technologies and pedagogical approaches. Furthermore, carrying out longitudinal research to evaluate the approach's long-term effects on reading comprehension and general academic performance would offer significant insights into its long-term advantages. Including a wider range of student demographics and educational attainment in the research can improve the model's accuracy and scope of application. Additionally, investigating how to include interactive and gamified learning components with BI-LSTM may provide a more stimulating and engaging learning environment. For the strategy to be refined and implemented on a broader scale, cooperation with educational institutions to pilot the model in real-world situations and collect input from educators and students would be essential.

REFERENCES

[1] Y. Arifani, “The effectiveness of Pamanpintermu e-reading program on EFL learners’ reading performances,” in 2018 3rd International Conference on Education, Sports, Arts and Management Engineering (ICESAME 2018), Atlantis Press, 2018, pp. 21–24.

[2] M. Gutiérrez-Colón, A. D. Frumuselu, and H. Curell, “Mobile-assisted language learning to enhance L2 reading comprehension: A selection of implementation studies between 2012–2017,” Interact. Learn. Environ., vol. 31, no. 2, pp. 854–862, 2023.

[3] Y. Xu, J. C. Yau, and S. M. Reich, “Press, swipe and read: Do interactive features facilitate engagement and learning with e-Books?,” J. Comput. Assist. Learn., vol. 37, no. 1, pp. 212–225, 2021.

[4] M. P. Vidal, “Effectiveness of Multimedia and Text-Based Reading Approaches to Grade 10 Students’ Reading Comprehension Skills,” AsiaCALL Online J., vol. 13, no. 4, pp. 55–79, Sep. 2022, doi: 10.54855/acoj.221345.

[5] A. Habók, T. Z. Oo, and A. Magyar, “The effect of reading strategy use on online reading comprehension,” Heliyon, vol. 10, no. 2, p. e24281, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24281.

[6] S. G. T. Ong and G. C. L. Quek, “Enhancing teacher–student interactions and student online engagement in an online learning environment,” Learn. Environ. Res., vol. 26, no. 3, pp. 681–707, Oct. 2023, doi: 10.1007/s10984-022-09447-5.

[7] J. M. Muñoz Rodríguez and A. Sánchez Rojo, “On Blended Learning Flexibility: An Educational Approach,” in Blended Learning: Convergence between Technology and Pedagogy, vol. 126, A. V. Martín-García, Ed., in Lecture Notes in Networks and Systems, vol. 126. , Cham: Springer International Publishing, 2020, pp. 21–44. doi: 10.1007/978-3-030-45781-5_2.

[8] M. L. Bernacki, J. A. Greene, and H. Crompton, “Mobile technology, learning, and achievement: Advances in understanding and measuring the role of mobile technology in education,” Contemp. Educ. Psychol., vol. 60, p. 101827, Jan. 2020, doi: 10.1016/j.cedpsych.2019.101827.

[9] M. Wardak, Mobile assisted language learning (mall): teacher uses of smartphone applications (apps) to support undergraduate students’ english as a foreign language (efl) vocabulary development. Lancaster University (United Kingdom), 2020.

[10] N. Nurjannah, K. Nurhadi, A. R. S. Tambunan, and others, “Assessing Student Empowerment in Mobile-Assisted Extensive Reading in a University Setting.,” Qual. Rep., vol. 28, no. 6, 2023.

[11] M. Valizadeh, “Investigating the impacts of mobile assisted reading on EFL learners’ vocabulary knowledge development,” Sak. Univ. J. Educ., vol. 12, no. 3, pp. 573–590, 2022.

[12] C.-C. Lin, V. Lin, G.-Z. Liu, X. Kou, A. Kulikova, and W. Lin, “Mobile-assisted reading development: a review from the Activity Theory perspective,” Comput. Assist. Lang. Learn., vol. 33, no. 8, pp. 833–864, 2020.

[13] O. Viberg, B. Wasson, and A. Kukulska-Hulme, “Mobile-assisted language learning through learning analytics for self-regulated learning (MALLAS): A conceptual framework,” Australas. J. Educ. Technol., vol. 36, no. 6, pp. 34–52, 2020.

[14] M. Mortazavi, M. K. Nasution, F. Abdolahzadeh, M. Behroozi, and A. Davarpanah, “Sustainable learning environment by mobile-assisted language learning methods on the improvement of productive and receptive foreign language skills: A comparative study for Asian universities,” Sustainability, vol. 13, no. 11, p. 6328, 2021.

[15] S. Ebadi and A. Raygan, “Investigating the facilitating conditions, perceived ease of use and usefulness of mobile-assisted language learning,” Smart Learn. Environ., vol. 10, no. 1, p. 30, 2023.

[16] S. Habib, A. Haider, S. S. M. Suleman, S. Akmal, and M. A. Khan, “Mobile assisted language learning: Evaluation of accessibility, adoption, and perceived outcome among students of higher education,” Electronics, vol. 11, no. 7, p. 1113, 2022.

[17] K.-O. Jeong, “Facilitating sustainable self-directed learning experience with the use of mobile-assisted language learning,” Sustainability, vol. 14, no. 5, p. 2894, 2022.

[18] Z. Chen, W. Chen, J. Jia, and H. An, “The effects of using mobile devices on language learning: a meta-analysis,” Educ. Technol. Res. Dev., vol. 68, no. 4, pp. 1769–1789, Aug. 2020, doi: 10.1007/s11423-020-09801-5.

[19] X. Lei, J. Fathi, S. Noorbakhsh, and M. Rahimi, “The impact of mobile-assisted language learning on English as a foreign language learners’ vocabulary learning attitudes and self-regulatory capacity,” Front. Psychol., vol. 13, p. 872922, 2022.

[20] J. Sabatini, T. O’Reilly, J. Weeks, and Z. Wang, “Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques,” Int. J. Test., vol. 20, no. 1, pp. 1–23, 2020.

[21] “datasets/docs/catalog/race.md at master • tensorflow/datasets,” GitHub. Accessed: Aug. 23, 2024. [Online]. Available: https://github.com/tensorflow/datasets/blob/master/docs/catalog/race.md

[22] J. Sinclair, E. E. Jang, and F. Rudzicz, “Using machine learning to predict children’s reading comprehension from linguistic features extracted from speech and writing.,” J. Educ. Psychol., vol. 113, no. 6, p. 1088, 2021.

[23] S. K. D’Mello, R. Southwell, and J. Gregg, “Machine-learned computational models can enhance the study of text and discourse: A case study using eye tracking to model reading comprehension,” Discourse Process., vol. 57, no. 5–6, pp. 420–440, 2020.

[24] L. S. Riza, Y. Firdaus, R. A. Sukamto, Wahyudin, and K. A. F. Abu Samah, “Automatic generation of short-answer questions in reading comprehension using NLP and KNN,” Multimed. Tools Appl., vol. 82, no. 27, pp. 41913–41940, 2023.

[25] Z. Ye et al., “Towards a Better Understanding of Human Reading Comprehension with Brain Signals,” in Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 380–391. doi: 10.1145/3485447.3511966.

Advancing Natural Language Processing with a Combined Approach: Sentiment Analysis and Transformation Using Graph Convolutional LSTM

Kedala Karunasree[1], Dr. P. Shailaja[2], Dr T Rajesh[3], Dr. U. Sesadri[4],
Dr. Choudaraju Neelima[5], Dr. Divya Nimma[6], Malabika Adak[7]

Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[1]
Department of Computer Science and Engineering, Vaagdevi College of Engineering, Warangal, India[2]
Asst. Professor, Dept. of CSE, G Narayanamma Institute of Technology and Science, India[3]
Associate Professor, Department of CSE, Vardhaman College of Engineering, Hyderabad, India[4]
Associate Professor, Department of Engineering English, College of Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[5]
PhD. in Computational Science, University of Southern Mississippi, Data Analyst In UMMC, USA[6]
Department of Applied Mathematics and Humanities, Yeshwantrao Chavan College of Engineering, Nagpur, India[7]

***Abstract*—Sentiment analysis is a key component of Natural Language Processing (NLP), taking into account the extraction of emotional cues from text. However, traditional strategies often fail to capture diffused feelings embedded in language. To deal with this, we advocate a novel hybrid model that complements sentiment analysis by way of combining Graph Convolutional Networks (GCNs) with Long Short-Term Memory (LSTM) networks. This fusion leverages LSTM's sequential reminiscence abilities and GCN's ability to model contextual relationships, allowing the detection of nuanced feelings regularly overlooked with the aid of conventional techniques. The hybrid technique demonstrates superior generalization overall performance and resilience, making it mainly powerful in complicated sentiment detection responsibilities that require a deeper knowledge of text. These results emphasize the capacity of combining sequential memory architectures with graph-based contextual facts to revolutionize sentiment analysis in NLP. This study not only introduces an innovative approach to sentiment analysis but also underscores the importance of integrating advanced techniques to push the boundaries of NLP research. This cutting-edge hybrid model surpasses the performance of previous techniques like CNN, CNN-LSTM, and RNN-LSTM with an amazing accuracy of 99.33%, creating a new benchmark in sentiment analysis. The results demonstrate how more precise sentiment analysis made possible by fusing sequential memory architectures with graph-based contextual information might revolutionise NLP. The findings provide a new benchmark, advancing the sphere by way of enabling greater specific and nuanced sentiment evaluation for a wide range of programs, inclusive of purchaser remarks analysis, social media monitoring, and emotional intelligence in AI structures.**

Keywords*—Graph Convolutional Networks (GCN); Long Short-Term Memory (LSTM); Natural Language Processing (NLP); sentiment analysis, emotions; text classification; machine learning

I. Introduction

As a core problem of natural language processing and considering the fact that there is a lot of texts, which are published on social networks, blogs, and other platforms, sentiment analysis has become a very important problem. To the corporations and decision-makers, as well as to the researchers who work with huge textual databases, it is crucial to understand these sentiments, emotions, and views [1]. This work introduces a novel approach that incorporates some of the most relevant features of many of the current state-of-art methods and results in the design of a new Hybrid GC-L LSTM model [2]. Due to this aforesaid hybrid model, sentiment analysis will be more accurate and contextual as this incorporates the complex relationships that exist between semantic relations, contextual relations, and temporal relations in the written text. Sentiment analysis of the posts on the social media platforms can therefore help the businesses determine customers' feedback on their products, potential trends in the market and generally, how their brands are being perceived. The sentiment analysis assists in the assessment of the overall perception of customers which in turn the businesses on how to come up with better products and or services [3]. The swift identification of the character of emotions in a particular industry and their disposition can help in the determination of the customer's attitude, the prediction of the market position and thus, the efficiency of the judgments made [4]. Using sentiment analysis in customers' engagements helps in addressing the issues of the client, to the satisfaction of the business. This is rather difficult because many terms have different meanings depending on the context of their use. What makes the analysis challenging is recognizing sarcasm and irony as sentiments can be stated inattentively to the opposite of the conveyed meaning. Sentiment analysis has to tread different languages and hence needs complex models to consider the differences in languages [5].

The ever-proliferation of user-generated content in the World Wide Web has brought about the sentiment analysis into sharp focus, as it is becoming important for organizations to gauge emerging trends, customer satisfaction levels and product performance [6]. Despite the fact that the earlier methods in sentiment analysis have been tried and tested and have achieved some levels of effectiveness, they frequently fail to capture the subtleties of sentiment expression, especially in situations where the meanings of words rely on their interactions with one another. In order to address these problems, this work integrates GCN that are famous for their ability to extract semantic relations between the words with LSTM networks that are extremely efficient in understanding the contextual history of the text [7]. Comparing the identification of sentiment subtleties in a diversification of textual data sources, the mentioned two techniques reached the product of which is known as the hybrid GC-LSTM model [8].

This paper discusses the architecture of the proposed Hybrid GC-LSTM model, its training process, and possible benefits for enhancing sentiment analysis in real-world settings. Text normalization and pre-processing are the preliminary phases of NLP that prepare text data for further analysis coherently and uniformly so as to obtain precise and accurate results of the natural language processing tasks. The aim of the project is not only to propose effective NLP methodologies but also to supply important tips for improving sentiment-oriented tasks in numerous spheres with the help of combining graph convolution with LSTM to stimulate sentiment analysis [9]. The present work aims to discuss the architectural characteristics of the Hybrid GC-LSTM model, its training process, and potential uses as well as to show how this particular model can potentially enhance the state of the art of sentiment analysis in a specific real-life environment. Text normalization and pre-processing are one of the critical stages in natural language processing with an essential impact on the majority of the following steps; determine the quality and comparability of the upcoming results. The algorithm is planned to create NLP techniques and offer valuable guidance for improving sentiment-related tasks of manifold domains by optimistically combining graph convolution and LSTM for accelerating sentiment analysis [10]. The H-GC-LSTM model creates new opportunities in NLP practice and research and allows researchers and practitioners to apply the opportunities of this innovative hybrid model. It also provides capabilities for the analysis of sentiments in textual data with better accuracy and specificity to context [11].

The key contributions of the article are,

- The method under investigation in this work is concerned with strait integration of LSTM and GCN networks within a single framework for sentiment analysis. The integration of GCNs, which applies a contextual approach to learn the interactions between nodes and LSTMs which are particularly good at sequence memory, keeps the benefits of the two architectures in mind.
- The first and major contribution for this research is to enhance the sentiment analysis with the help of GCNs and LSTMs that exhibit the complementing features. Regardless of the fact that LSTMs are more effective in sequential learning, GCNs offer the notable contribution to extract intricate contextual information from a textual input. The combination of both the approaches hence leads to the development of sentiments analysis model that is more complex and immune.
- Extensive testing on standard sentiment analysis datasets demonstrates the model's improved precision in identifying and classifying nuanced sentiment subtleties. The combination of approaches outperforms the state-of-the-art models now in use, particularly in situations when a thorough comprehension of the context is essential.
- Across a range of datasets, the hybrid Graph Convolutional LSTM approach demonstrates improved generalization effectiveness and durability. This highlights the possibilities for practical applications by demonstrating its flexibility to a variety of language situations and datasets.

The rest of the study is structured as follows: Section II comprises relevant material designed to help readers comprehend the proposed paper using existing methodologies, while Section III elaborates on the problem description. Section IV displays the proposed architectures. Section V includes tabular and graphical representations of the results and performance indicators. Finally, in Section VI, the conclusion and future works are discussed.

II. Related Work

The goal of the project is to enhance sentiment categorization by using input from users from social media sites such as Facebook and Twitter [12]. It highlights how sentiment analysis in NLP may be improved by combining word-embedded techniques like Word2Vec and FastText with algorithms for DL like CNN, GRU, LSTM, and Bi-LSTM. This study presents a novel combination model that deliberately blends DL techniques with several word embedding approaches, beating previous research and providing better sentiment classification results in the end.

The study discusses the growing significance of user reviews on forums, social networks, and online stores as well as the requirement for efficient sentiment analysis [13]. The article presents a unique hybrid CNN model that preserves significant data while processing input sequentially by fusing CNN and Bi-LSTM models. The algorithm's exceptional accuracy is demonstrated by experimental findings utilizing benchmark brand and accommodation review information sets, which achieve 93.6% and 92.7%, respectively, for product and hotel reviews, respectively, exceeding modern facilities sentiment analysis approaches.

In Arabic tweets, the paper investigates the application of pre-trained BERT algorithms for Arabic sentiment analysis [14]. The Ara BERT model is refined by the researchers and used in a network design that combines the GRU and BiLSTM models. Their tests show that the optimized AraBERT model combined with the hybrid networks performs exceptionally well, outperforming other popular sentiment analysis techniques like as CNN, LSTM, BiLSTM, and GRU, and achieving up to 94% accuracy.

The study focuses on sentiment analysis of online feedback, which is a challenging undertaking because consumers utilize natural language [15]. The current research emphasizes phrase-level and sentence-level characteristics for sentiment analysis, in contrast to prior approaches that primarily depended on word-level characteristics and other feature weighting techniques. A two-layer CNN and a BGRU are used in the suggested hybrid model, a convolutional CBRNN, to identify abundant phrase-level features and record chronological data by detecting dependencies that span time. The research results show that the CBRNN model is improved to the latest algorithms by 2% to 4%, with an F1 score of 87.62% on the IMDB collection and 77.4% on the Polarity information. However, training this model takes a little longer.

The paper discusses the use of graph neural models in dependency trees as well as aspect-based sentiment analysis [16]. Sentic GCN, a unique strategy that is used to learn the emotive connections of phrases particular to a particular feature, is introduced in this research. Previous efforts have concentrated on learning dependence data gathered from surrounding words to aspect words. The model improves sentence dependence graphs by including SenticNet's emotive information. It does this by taking into account the behavioural data involving opinion words and the aspect as well as the connection involving contextual and aspects words. As experiments on many standards datasets have demonstrated, the proposed Sentic GCN model performs better in sentiment analysis through aspects than the current state of the art.

Sentiment evaluation is a technique inside natural language processing that evaluates and identifies the emotional tone or temper conveyed in textual data. Scrutinizing phrases and phrases categorize them into hopeful, terrible, or neutral sentiments. The significance of sentiment analysis lies in its potential to derive precious insights from huge textual data, empowering organizations to understand consumer sentiments, make informed choices, and enhance their services. For the in-addition development of sentiment evaluation, gaining deep expertise in its algorithms, application, contemporary performance, and challenges is vital. Therefore, on this extensive survey, Jim et al., [17] started exploring the widespread array of software domain names for sentiment evaluation, scrutinizing them within the context of current studies. Then delved into prevalent pre-processing strategies, datasets, and assessment metrics to enhance comprehension. This study additionally explored Machine Learning, Deep Learning, Large Language Models, and Pre-trained models in sentiment analysis, supplying insights into their advantages and drawbacks. Subsequently, this study precisely reviewed the experimental outcomes and boundaries of new state-of-the-art articles. Finally, this study discussed the various challenges encountered in sentiment evaluation and proposed future research guidelines to mitigate these concerns. This great evaluation offers a complete knowledge of sentiment analysis, covering its approaches, utility domains, consequences evaluation, challenges, and studies guidelines.

Natural Language Processing (NLP) is a critical department of artificial intelligence that research how to enable computer systems to recognize, manner, and generate human language. Text category is a fundamental undertaking in NLP, which goals to classify text into exceptional predefined classes. Text type is the most primary and classic task in natural language processing, and maximum of the duties in natural language processing can be regarded as classification obligations. In recent years, deep learning has accomplished terrific fulfilment in many studies fields, and nowadays, it has also become a standard technology inside the discipline of NLP, which is broadly included into textual content class tasks. Unlike numbers and pixels, text processing emphasizes satisfactory-grained processing potential. Traditional textual content classification strategies commonly require preprocessing the input model's text records. Additionally, they also need to obtain good pattern capabilities via guide annotation after which use classical devices to gain knowledge of algorithms for categories. Therefore, this paper analyses the software popularity of deep learning inside the three middle duties of NLP (which include textual content representation, phrase order modelling, and knowledge representation). Xu et al. [18] explore the enhancement and synergy accomplished through natural language processing within the context of textual content category, whilst additionally contemplating the challenges posed through antagonistic strategies in text era, textual content type, and semantic parsing. An empirical look at textual content class obligations demonstrates the effectiveness of interactive integration schooling, particularly in conjunction with TextCNN, highlighting the significance of those improvements in text classification augmentation and enhancement.

The interactive attention graph convolution community (IAGCN), a novel version proposed in the study proposed by Singh et al., [19] will revolutionize element-level sentiment evaluation (SA). IAGCN effectively addresses those key features, in contrast to previous research that left out the means of issue terms and their courting with context. The version combines a modified dynamic weighting layer with bidirectional long brief-time memory (BiLSTM) to appropriately gather context. It makes use of graph convolutional networks (GCNs) to encrypt syntactic records from the syntactic dependency tree. Furthermore, a way for interactive interest is employed to find out the elaborate relationships among context and issue phrases, which ends up inside the reconstruction of these terms' representations. Comparing the proposed IAGCN model to baseline models, incredible gains are made. Across 5 datasets, the model beats preceding methods with a wonderful improvement in F1 scores in those stages from 1.34% to four.04% and an impressive improvement in accuracy that levels from 0.56% to 1.75%. Additionally, the IAGCN model outperforms the global vectors (GloVe)-primarily based approach when the strong pretrained version bidirectional encoder representations from transformers (BERT) is covered inside the venture, resulting in even extra upgrades. The F1 rating notably will increase from 2.59% to 7.55%, and accuracy will increase from 1.47% to 3.95%, making the IAGCN model a standout performer in issue-degree SA.

Thus, the attempt to build fine-grained aspect-based sentiment analysis from existing sentiment analysis algorithms is not always successful because they are known to employ features at the word level and apparently do not address the contextual emotional knowledge. In an effort to eliminate these disadvantages, a new graph GC-LSTM is introduced and it formulates sentence dependency graphs based on certain features and incorporates emotional knowledge. This way, the model may also consider the connections between the opinion words and the aspect, and between the contextual and aspect words. The current models of analysis of the sentiment in Natural Language Processing can fail to capture complex details and the connection between them in a text. This research seeks to do so by presenting a new hybrid model that first consists of Long Short-Term Memory to capture the sequential memory retention and followed by Graph Convolutional Network to capture contextual relationships.

III. Problem Statement

Current sentiment evaluation techniques exhibit several key drawbacks that necessitate the development of advanced models. First, many models like CNN, GRU, and BiLSTM depend heavily on word-stage capabilities, frequently failing to capture the deeper contextual relationships needed to interpret complex texts, together with sarcasm or irony. This loss of contextual understanding limits their ability to as it should be classifying sentiment at the sentencing stage. Second, while LSTM and BiLSTM models cope with sequential memory retention, they're inadequate for extracting deep insights from word-level functions and temporal dependencies [13]. Hybrid models, which include CBRNN and CNN-BiLSTM, provide modest enhancements however nevertheless fall short in dealing with chronological relationships, leading to incomplete sentiment classifications [15]. Additionally, aspect-based sentiment analysis remains inefficient as current algorithms primarily cognizance on neighbouring issue words without fully addressing emotional and contextual dependencies among opinion and thing phrases. Although models like Sentic GCN [16] try and incorporate emotional expertise, they nevertheless warfare with processing complex dependencies, resulting in suboptimal performance. Finally, scalability and performance bottlenecks pose tremendous demanding situations, specifically in area-specific sentiment analysis. Fine-tuned models like AraBERT, although powerful, are computationally highly-priced and gradual to teach, limiting their applicability to big datasets. Consequently, there's a growing need for advanced hybrid fashions which could integrate contextual emotional knowledge, successfully cope with sequential statistics, and enhance aspect-based totally sentiment analysis to reap greater correct, efficient, and scalable sentiment class throughout numerous domains and languages.

IV. Proposed Hybrid Graph Convolutional LSTM Framework

Since the proposed method is the Graph Convolutional LSTM (GCLSTM) model, training of the proposed model is done with care using sentiment analysis datasets. The first process that is called preprocessing of the textual data involves the steps such as tokenization and normalization which is done to make the input text more standardized. Also, a graph representation is built in order to better understand the structure of tokens and their relations, in terms of semantics to the whole text. After that, the resultant model is trained with the semi-supervised learning methods based on the graph and supervised learning. This amalgamation helps the model to incorporate hierarchical features, contextual information and complex text data relationship. Graph based approaches combined with LSTM networks mean that the model is able to capture and understand sequential dependencies as well as the contextual information which improves the prediction of sentiment in general. In order to assess the performances of the model in the context of identifying sentiments from the textual data, benchmarking datasets for the sentiment analysis task are used. As shown in Fig. 1, the model has demonstrated it is efficient in the assessment of the sentiment classification of different datasets. As a result of the specific assessment strategies proposed in the model, the excellent accuracy, stabilization, and versatility of the proposed model surpass prior approaches to sentiment classification. The use of graph-based representation and LSTM networks enables the GCLSTM model to effectively obtain sentiment information hidden in texts. Using both structural and sequential information, the model goes beyond the usual approaches and achieves rather high accuracy in sentiment classification tasks. This proves to be a methodologically innovative that makes a methodological shift on the field of sentiment analysis in Natural Language Processing (NLP), and paves the way for enhanced and efficient analysis of textual sentiment across multifaceted real-world contexts.

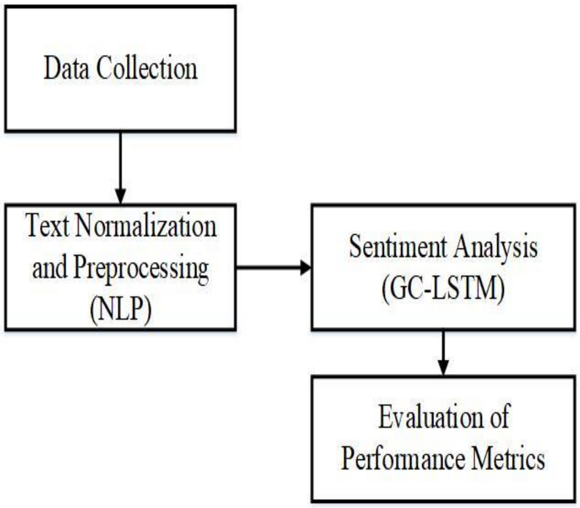

Data Collection
Text Normalization and Preprocessing (NLP)
Sentiment Analysis (GC-LSTM)
Evaluation of Performance Metrics

Fig. 1. Proposed methodology.

A. Data Collection

On Kaggle, a list of the tweets with clearly separating positive and negative attitudes could be seen; clearly, it is better to have a selection of well-curated tweets. Positive tweets refer to preoccupation with ideas, emotions, and opinions that are favourable while negative tweets are those that involve criticism, dissatisfaction or feelings of discomfort. For this purpose, using this dataset, academicians as well as data scientists can build models that could recognize positive as well as negative sentiments present in text data of social media appropriately, so this is going to serve a fruitful tool for sentiment analysis as well as Natural Language processing applications [20].

B. NLP for Text Normalization and Pre-processing

Natural Language Processing involves the important area of text normalization and pre-processing whereby issues associated with linguistic varieties and the like, are dealt with in order to standardize textual data. Lowercasing is a process of taking the whole text and turning all characters to lower case. This assists in discarding case-based dependencies so as to arrive at an equal representation of words in a field. It is especially helpful in such cases as text matching where that changes in upper or lower case can cause disparities. The process of text normalization and pre-processing which is an important process and frequently discussed in the framework of Natural Language Processing (NLP) comprises the following crucial procedures [21]: Some of these responsibilities include converting the text to a list of words where words can be individual words or sub word units, converting the entire text to lowercase and removing special characters, punctuation marks, and other information that is assumed to be noise. In the same way, in order to diminish the words into their base forms, lemmatization or stemming techniques may also be incorporated in this procedure. They are called stop words – these terms are rather ubiquitously used and, thus, less meaningful in specific contexts. This manner, these NLP approaches support a crucial part in ensuring data consistency by providing a commonly accessible pre-processed text corpus that has already gone through denoising, and hence, ease the process of text analysis, sentiment analysis, as well as other applications of NLP [22].

C. Employing Hybrid Graph Convolutional LSTM Model for Sentiment Analysis

In the field of NLP, GC-LSTM for sentiment analysis is one of the modern approaches. The act of determining the extent of positive or negative opinion expressed in textual data is known as sentiment analysis, an activity that may be used in social network monitoring or in understanding users' attitude towards a certain product or service. These two building blocks are powerful and when combined together, this new method we have proposed is not only able to capture the general sentiment poles in text data, that is positive or negative, in most cases, but also give extremely well results for intensity of sentiment poles in the text data as well. GCNs' capability of extracting and expressing semantic relations between words or between subword units is already clear. The model accrues and disseminates knowledge about word associations by constructing a network that consists of nodes – words – and edges – relationships between these nodes. This is especially important when dealing with sentiment analysis since sentiment phrases ordinarily depend on connections with other words and their context. These are the primary GC-LSTM from Eq. (1) to Eq. (5).

$$i_p = \sigma\,(V_{yi} * Y_p + V_{ki} * K_{p-1} + d_i) \tag{1}$$

$$f_p = \sigma\,(V_{yf} * Y_p + V_{kf} * K_{p-1} + d_f) \tag{2}$$

$$C_p = f_p \circ C_{p-1} + i_p \circ \tanh\,(V_{yc} * Y_p + V_{kc} * K_{p-1} + d_c) \tag{3}$$

$$o_p = \sigma\,(V_{yo} * Y_p + V_{ko} * K_{p-1} + d_o) \tag{4}$$

$$K_p = o_p \circ \tanh(C_p) \tag{5}$$

While, the LSTM network is much effective to understand the sequential patterns in the text. It is relevant to word order and it reveals how the initial words in a sequence influence the degree of emotion which is expressed by the final word. In the case of the sentiment analysis, sequential context is important because it gives information on how sentiment flows from one word to the next in a given sentence or a paragraph. This is a combined approach of LSTM and GCN in a way to utilise the benefits of both techniques. This means that when using graph-based paradigms, then students can design effective PDAs that will cater for their needs. It is the synergy that gives a better understanding of the emotional expressions which are present in the textual data. This hybrid model has practical applications in a number of areas of actuality. Where a finer definition of attitude can yield improved customer' satisfaction and, thus, improved products. On a large scale, it may help organizations in responding to the new trends or concerns on the part of communities by monitoring them and understanding the extent of reaction possible. Besides, the hybrid approach can be useful for decision-making in a matter of policy choices and organization of social activity based on more accurate data on public attitudes to a political and/or social concern within the context of opinion mining.

The input L goes through several layers of GC-LSTM as represented by 'GC-LSTM1,' to come out as a 3D tensor, and is then funneled to a regular CNN for sentiment analysis. CNNs produced the following output in Eq. (6),

$$\text{o} = \{o_1, o_2, \ldots, o_c\} \tag{6}$$

The number of action categories is represented by C. To train the GC-LSTM, the categorical cross-entropy loss can be used. Given a sequence L, the expected chance of having the ith class is in Eq. (7),

$$\text{Q}\,(C_i |\, \text{L}) = \frac{e^{o_i}}{\sum_{j=1}^{C} e^{o_j}},\ \text{h=1,…,C} \tag{7}$$

Hybrid Graph Convolution LSTM Model for Sentiment Analysis as NLP progresses shows a great leap in the ability to detect and apply sentiment from texts. Due to the new approach to revealing the relations between semantic relatedness and sequential context, this method encourages the development of sentiment analysis capabilities and will ultimately contribute to the creation of highly accurate and context-sensitive sentiment analysis for a wide spectrum of applications.

V. Results and Discussion

In the proposed methodology, the GCLSTM model is fine-tuned using sentiment analysis datasets so as to improve the process of sentiment analysis following a multiple-step process. First, the text input goes through pre-processing steps which include; tokenization, normalization and constructing a graph that can show the relationships that exist among words and phrases. This pre-processing phase is fundamental and serves the purpose of transforming the textual input into a uniform nature while at the same time capturing closest semantic relationship present within the data. After pre-processing the model is trained by a combination of supervised learning and graph-based semi-supervised learning. This is quite beneficial since it allows the model to gather information from labelled as well as features from unlabelled data due to the graph structure that encodes a lot of contextual information. Thereby, by employing graph-based semi-supervised learning, the model becomes wiser to the correlation between the text input and pyramidal features to make more distant differentiation between the sentiment characteristics. At the time of training, the model keeps on learning to traverse through the structure of the graph, enabling it to adjust its parameters to provide the best sentiment analysis ramifications. As it moves to the subsequent form through each iteration, the model amasses the competence to provide for the entities, such as words, phrases as well as sentiment expressions in the text data. Combined with the structural features' temporal dependencies learnt by LSTM networks and the structural relationships' contextual relationship learnt by GCN networks, GCLSTM obtains good results in sentiment analysis tasks. However, the integration of graph-based techniques improves the model's comprehension of the text's semantic features while considering the sentiment variations that the other strategies might fail to identify. It also ensures that the GCLSTM model receives profound training so that in the future it can handle multiple different datasets and different sentiment analysis environments. When training GCLSTM model it performs the following steps namely tokenization and normalization of the sentences, converting the sentences into a graph-based semantic representation and lastly the hybrid training considering both supervised and semi-supervised learning. This approach of dealing with the data allows the model to capture semantic relations, as well as hierarchy and context of the different features which in turn improves the accuracy and overall robustness of the sentiment analysis model.

A. Model Accuracy

Metrics provides information on how well the model's predictions match the actual values and is commonly used in classification problems. Especially in cases when there is a mismatch in the classes or when there are additional expenses associated with wrong classifications.

The model accuracy achieved by the proposed technique is a very impressive 99.33%. This astounding level of precision suggests that the model performs remarkably well when it comes to outcome prediction. A graphical depiction of this performance and the strategy's effectiveness is shown in Fig. 2.

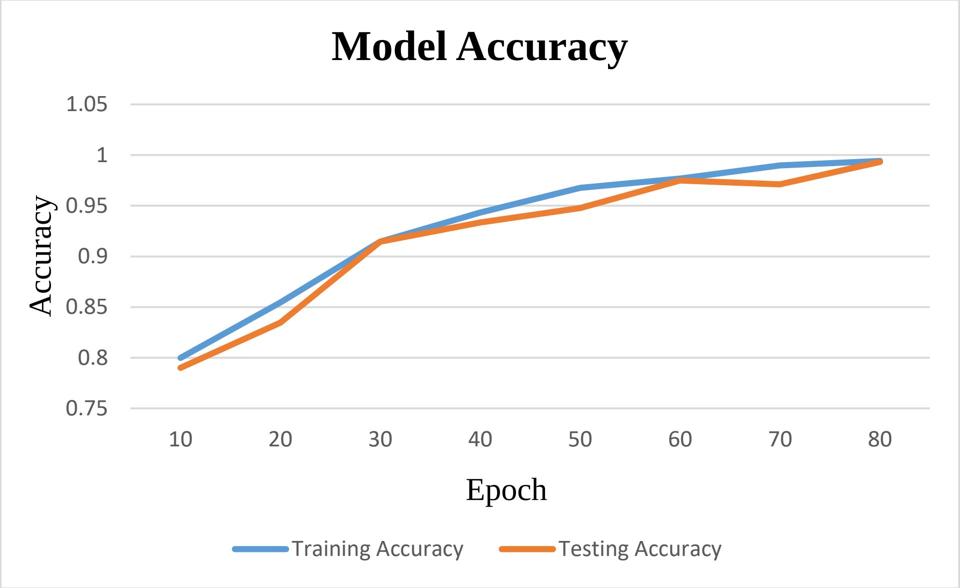


Fig. 2. Model accuracy.

B. Model Loss

The model loss, or the difference between the model's predictions and the actual data, indicates how well the model performed on the training set of data. Lower loss values indicate better performance since they demonstrate how well the model's predictions match the actual data.

This is often computed using a variety of loss coefficients, including as mean squared error and cross-entropy, which help the model learn by adjusting its parameters to reduce the difference between the predicted and actual outcomes.In addition to providing a graphical representation of the model's performance changes during training, Fig. 3 illustrates the model loss of the recommended technique. The variation in the loss over model training epochs, which measures the discrepancy between predicted and actual values, is depicted graphically in this picture.

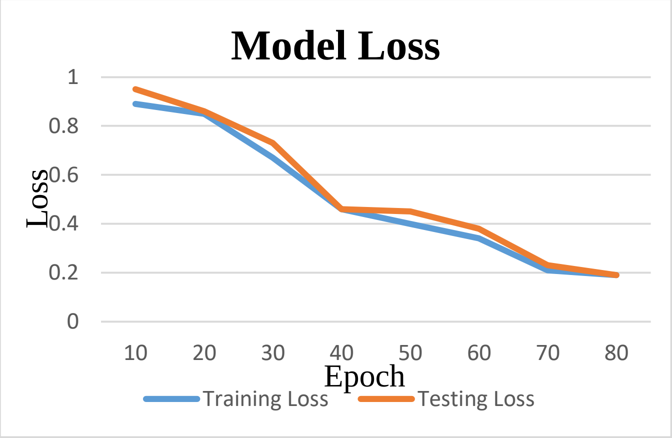


Fig. 3. Model loss.

C. ROC

The Receiver Operating Characteristic (ROC) curve is a graphical representation that is used to evaluate the performance of classification models. It illustrates the trade-off between a model's sensitivity and specificity over several thresholds. When establishing the optimal threshold for classification tasks or assessing the performance of different models, the ROC curve comes in handy. This is especially true when the distribution of classes is not balanced. The ROC value is displayed in Fig. 4.

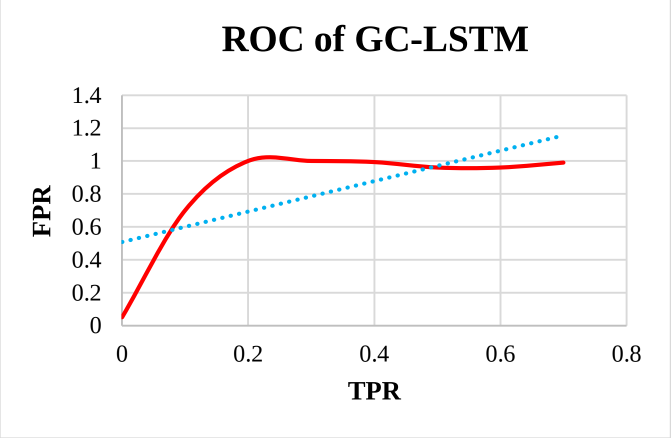


Fig. 4. The Receiver Operating Characteristic (ROC) curve.

The underlying assumption of it is that all interactions are predictable. The precision is given by Eq. (8).

$$Accuracy = \frac{T_{Pos} + T_{Neg}}{T_{Pos} + T_{Neg} + F_{Pos} + F_{Neg}} \tag{8}$$

$$P = \frac{T_{Pos}}{T_{Pos} + F_{Pos}} \tag{9}$$

The appropriate positive for these numbers may be calculated the line is found in Eq. (10).

$$R = \frac{T_{Pos}}{T_{Pos} + F_{Neg}} \tag{10}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{11}$$

TABLE I. COMPARISON OF PERFORMANCE METRICS

Methods	Accuracy	Precision	Recall	F1-Score
CNN	94.56	95.76	95.11	95.88
RNN-LSTM	93.65	92.34	95.78	96.67
CNN-LSTM	97.12	97.87	96.74	93.23
Proposed GC-LSTM	99.33	98.98	98.23	98.11

Table I provides a comparative study of several sentiment analysis techniques, emphasizing measures such as accuracy, precision, recall, and F1-score. Prominently, the suggested GC-LSTM model surpasses other techniques with remarkable accuracy (99.33%) and resilient performance on all measures, demonstrating elevated precision, recall, and F1-Score percentages.

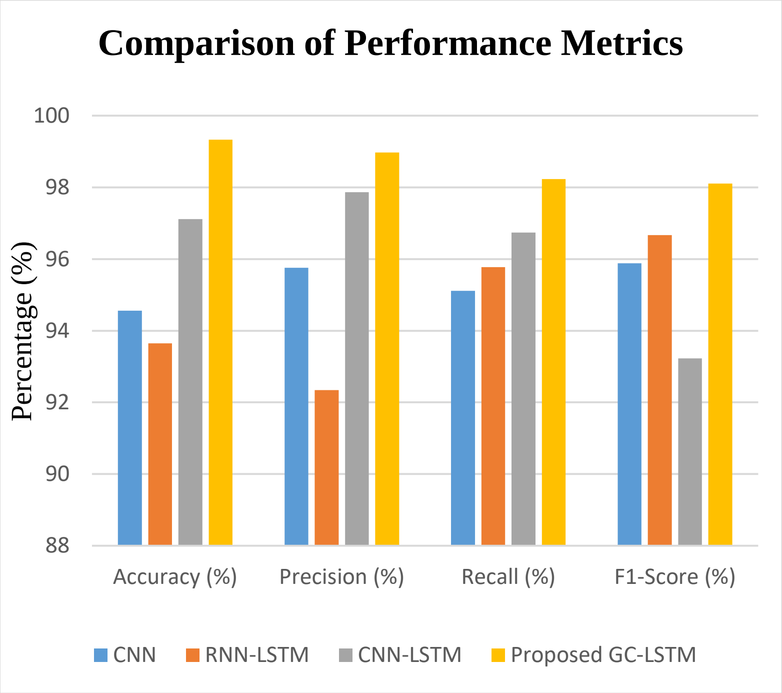


Fig. 5. Performance metrics.

A notable substitute is the CNN-LSTM model, which likewise performs well, especially in terms of accuracy and precision. In the meanwhile, the hybrid models outperform the conventional CNN and RNN-LSTM models, which function well but display somewhat worse outcomes than the former. CNN-LSTM, in particular, strikes an impressive balance between recall and precision. These findings highlight the effectiveness of integrating sequential context analysis (LSTM) with graph-based semantic understanding (GC) in sentiment analysis, which greatly improves the model's capacity to correctly categorize sentiments in text data. In Fig. 5, it is shown.

D. Discussion

The Hybrid GC-LSTM approach gives significant advancements in sentiment analysis; however, several obstacles remain. Its excessive computational complexity, due to the combination of GCNs and LSTMs, makes it resource-extensive, specifically for real-time programs [23]. The graph construction process provides preprocessing overhead, and the model struggles with unstructured text like sarcasm or irony. Additionally, adapting the version for various languages with complex systems may be difficult. It is based on amazing statistics, which might not be available for all domains, and is liable to overfitting on small datasets. Furthermore, its interpretability is limited, complicating transparency in decision-making. The outcomes reveal the advanced overall performance of the proposed GC-LSTM version, reaching an impressive accuracy of 99.33%, drastically outperforming other models like CNN and RNN-LSTM. It also excels in precision (98.98%), consider (98.23%), and F1-rating (98.7%). Compared to the CNN-LSTM, which has robust precision (97.87%) but however slightly decreased taken into account, the GC-LSTM showcases a balanced and robust performance across all metrics. These findings affirm that combining LSTM's sequential context evaluation with GCN's semantic expertise complements the model's ability to correctly classify sentiments in textual content information. Future work for this study includes enhancing the GC-LSTM model to reap actual-time sentiment evaluation and optimizing its computational performance for quicker processing. The model's adaptability throughout exclusive languages and domain names will be increased, addressing demanding situations in multilingual sentiment evaluation. Additionally, quality-tuning for complex textual content kinds together with sarcasm and irony could be explored. Applications in emotion-pushed structures, decision-making strategies, and different rising fields may also be investigated. The purpose is to similarly improve the version's scalability, robustness, and versatility in diverse sentiment analysis environments. This exceptional performance serves as a testament to the synergy between graph-based semantic comprehension and sequential context analysis within the hybrid architecture. By reconciling GCNs and LSTM networks, the proposed GC-LSTM model takes the best out of both methodologies without the missteps of the other [24]. In the matter of the meaning comprehension the graph-based semantic understanding of the relationships between words and phrases is improved, and in the case of the sequential context and temporal connections between the word LSTMs perform well. This reasonable combination of techniques leads to the model that at the same time has market-worthy accuracy, and high enough levels of precision, recall as well as F1-Score in different tasks of sentiment analysis. In addition, the performance of the proposed CNN-LSTM model can also be called good, as it showed a high degree of accuracy and sensitivity in solving the sentiment analysis problems [25]. The better results of the hybrid models over the basic types of the models show that innovation should remain one of the priorities in the development of the sentiment analysis. Through actively incorporating new approaches and incorporating various approaches and techniques, it becomes possible to expand the overall performance of sentiment analysis research by the researchers and practitioner. The findings presented in the table provide compelling evidence of the efficacy of hybrid sentiment analysis models, particularly the GC-LSTM model. With its exceptional accuracy and consistently strong performance metrics, the GC-LSTM model exemplifies the power of combining graph-based semantic comprehension with sequential context analysis. Moving forward, continued exploration and refinement of hybrid techniques promise to further enhance sentiment analysis capabilities, enabling deeper insights into textual sentiment across a wide range of applications and domains [26].

VI. Conclusion and Future Scope

The proposed study combines GCNs and LSTM networks to enhance sentiment evaluation with the aid of capturing both contextual relationships and sequential reminiscence. This hybrid version gives better accuracy, nuanced emotion detection, and advanced generalization, outperforming traditional methods, making it tremendously effective for complicated sentiment analysis responsibilities throughout diverse applications. To highlight, I incorporate LSTM networks with GCNs, which exploits the advantages of both methods to enhance the efficiency and precision of SA applications. This model changes the way sentiment analysis is done as it is better than other approaches to identifying intricate semantic connections between words and phrases. Because of its versatility, the model is suitable for the most NLP tasks, especially for analysing sentiments in social media, product reviews, and customers' feedback. In addition to that, through language expansion it makes the facility more capable of accommodating a wide scope of linguistic settings making it more ‘universal’ in its application. The current focuses would be to extend the current concept with further upgrades which would help to accentuate the fine-grained sentiment analysis, to achieve the ability to work in real-time mode that would provide immediate results, as well as, consider its applications in the emerging areas such as emotion-driven systems and decisions based on sentiment analysis. As the Hybrid GC-LSTM model grows in the future, it will be largely help in analysing text sentiments of the varied languages and conditions and lays down the perfect platform to advance in the more superior sentiment analysis techniques. Furthermore, such integration of this model is expected to deal with aspects pertaining to different domains and different apertures of languages with relative ease in terms of generalization and flexibility for sentiment analysis. Having said that, it is not difficult to note that, because of its flexibility, sentiment analysis can offer a solution adapted to the needs of many industries and types of uses, and remain as efficient as before.

Hence, Hybrid GC-LSTM is a breakthrough in sentiment analysis of NLP due to its uncontrolled accuracy and scalabilities as well as application. As shown in the current paper, it increases numerous academics and organisations' ability and capacity to understand textual sentiment across different languages and situations; it marks the dawn of a new era of advanced sentiment analysis methodologies. The proposed study's barriers consist of capacity overfitting due to the complexity of the hybrid approach, reliance at the first-class of training datasets, and feasible challenges in computational efficiency. Additionally, the version may struggle with nuanced or area-specific sentiments not well-represented in the training data, limiting its widespread applicability.

ACKNOWLEDGMENT

The authors would like to acknowledge that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

REFERENCES

[1] Z. Wang, Y. Zhu, S. He, H. Yan, and Z. Zhu, "Llm for sentiment analysis in e-commerce: A deep dive into customer feedback," Appl. Sci. Eng. J. Adv. Res., vol. 3, no. 4, pp. 8–13, 2024.

[2] "A COMBINED DEEP LEARNING MODEL FOR PERSIAN SENTIMENT ANALYSIS | IIUM Engineering Journal." Accessed: Nov. 03, 2023. [Online]. Available: https://journals.iium.edu.my/ejournal/index.php/iiumej/article/view/1036

[3] J. A. Aguilar-Moreno, P. R. Palos-Sanchez, and R. del Pozo-Barajas, "Sentiment analysis to support business decision-making. A bibliometric study," AIMS Math., vol. 9, no. 2, pp. 4337–4375, 2024.

[4] "A novel weight-oriented graph convolutional network for aspect-based sentiment analysis | The Journal of Supercomputing." Accessed: Nov. 03, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11227-022-04689-9

[5] "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis | Journal of Big Data." Accessed: Nov. 03, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s40537-022-00680-6

[6] "Applied Sciences | Free Full-Text | Enhanced Arabic Sentiment Analysis Using a Novel Stacking Ensemble of Hybrid and Deep Learning Models." Accessed: Nov. 03, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/12/18/8967

[7] "Applied Sciences | Free Full-Text | Graph Convolutional Networks with POS Gate for Aspect-Based Sentiment Analysis." Accessed: Nov. 03, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/12/19/10134

[8] "Applied Sciences | Free Full-Text | Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism." Accessed: Nov. 03, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5841

[9] M. Zhang and T. Qian, "Convolution over Hierarchical Syntactic and Lexical Graphs for Aspect Level Sentiment Analysis," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 3540–3549. doi: 10.18653/v1/2020.emnlp-main.286.

[10] Q. Lu, X. Sun, R. Sutcliffe, Y. Xing, and H. Zhang, "Sentiment interaction and multi-graph perception with graph convolutional networks for aspect-based sentiment analysis," Knowl.-Based Syst., vol. 256, p. 109840, Nov. 2022, doi: 10.1016/j.knosys.2022.109840.

[11] D. Zhao, J. Wang, H. Lin, Z. Yang, and Y. Zhang, "Extracting drug–drug interactions with hybrid bidirectional gated recurrent unit and graph convolutional network," J. Biomed. Inform., vol. 99, p. 103295, Nov. 2019, doi: 10.1016/j.jbi.2019.103295.

[12] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," IEEE Access, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/ACCESS.2020.2982538.

[13] M. Kuppusamy and A. Selvaraj, "A novel hybrid deep learning model for aspect based sentiment analysis," Concurr. Comput. Pract. Exp., vol. 35, no. 4, p. e7538, 2023, doi: 10.1002/cpe.7538.

[14] N. Habbat, H. Anoun, and L. Hassouni, "A Novel Hybrid Network for Arabic Sentiment Analysis using fine-tuned AraBERT model," Int. J. Electr. Eng. Inform., vol. 13, Jan. 2022, doi: 10.15676/ijeei.2021.13.4.3.

[15] S. Soubraylu and R. Rajalakshmi, "Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews," Comput. Intell., vol. 37, no. 2, pp. 735–757, 2021, doi: 10.1111/coin.12400.

[16] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," Knowl.-Based Syst., vol. 235, p. 107643, Jan. 2022, doi: 10.1016/j.knosys.2021.107643.

[17] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. Mridha, "Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review," Nat. Lang. Process. J., p. 100059, 2024.

[18] X. Xu, Z. Xu, Z. Ling, Z. Jin, and S. Du, "Comprehensive implementation of TextCNN for enhanced collaboration between natural language processing and system recommendation," in International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2024), SPIE, 2024, pp. 1527–1532.

[19] N. K. Singh et al., "Deep Learning Model for Interpretability and Explainability of Aspect-Level Sentiment Analysis Based on Social Media," IEEE Trans. Comput. Soc. Syst., 2024.

[20] "Sentiment Analysis Dataset." Accessed: Nov. 02, 2023. [Online]. Available: https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset

[21] M. R. Islam, A. Ahmad, and M. S. Rahman, "Bangla text normalization for text-to-speech synthesizer using machine learning algorithms," J. King Saud Univ.-Comput. Inf. Sci., vol. 36, no. 1, p. 101807, 2024.

[22] E. Frank, J. Oluwaseyi, and G. Olaoye, "Data preprocessing techniques for NLP in BI," 2024.

[23] P. Mei and Y. H. Zhao, "Dynamic network link prediction with node representation learning from graph convolutional networks," Sci. Rep., vol. 14, no. 1, p. 538, 2024.

[24] T. Alsaedi, M. R. R. Rana, A. Nawaz, A. Raza, and A. Alahmadi, "Sentiment Mining in E-Commerce: The Transformer-based Deep Learning Model," Int. J. Electr. Comput. Eng. Syst., vol. 15, no. 8, pp. 641–650, 2024.

[25] X. Shao and C. S. Kim, "Accurate Multi-Site Daily-Ahead Multi-Step PM 2.5 Concentrations Forecasting Using Space-Shared CNN-LSTM.," Comput. Mater. Contin., vol. 70, no. 3, 2022.

[26] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," Knowl.-Based Syst., vol. 235, p. 107643, Jan. 2022, doi: 10.1016/j.knosys.2021.107643.

ABC-Optimized CNN-GRU Algorithm for Improved Cervical Cancer Detection and Classification Using Multimodal Data

Donepudi Rohini[1], Dr. M Kavitha[2]
Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF),
Vaddeswaram, Guntur Dist, Andhra Pradesh, India[1]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF),
Vaddeswaram, Guntur Dist, Andhra Pradesh, India[2]

***Abstract*—Cervical cancer is the second most common malignancy among women, making it a major public health problem worldwide. Early detection of cervical cancer is important because it increases the chances of effective treatment and survival. Regular screening and early management can prevent the growth of cervical cancer, thus reducing mortality. Traditional methods of detection, such as Pap smears, have proven useful, but are time-consuming and rely on behavioral interpretation by cytologists. To overcome these issues the study uses method another for a convolutional neural networks (CNNs) and gated recurrent units (GRUs) to detect and classify cervical cancer in Pap smear images by tuning with Artificial Bee Colony (ABC) Optimizer. This study used several datasets with high-resolution images from the SipakMed collection, with 4049 images and a fetal dataset with patient information for the CNN component of the model, specifically the ResNet-152 system, is extracted spatial attributes from these images. After feature extraction, the GRU component analyzes the sequential data to identify temporal combinations and patterns. This hybrid CNN-GRU algorithm uses the features of two networks: the ability of CNN to learn spatial patterns and the ability of GRU to understand sequential networks and tuning the parameters using ABC. The proposed model outperformed the conventional ML methods with a classification accuracy of 94.89%, and provided a reliable solution for early detection of cervical cancer Using these DL methods role which, not only enables a more accurate diagnosis, but also allows a comprehensive examination of the abnormal cervical cells, making it a positive detections to programs and patient outcomes. This work highlights the promise of cutting-edge AI techniques to improve cervical cancer diagnosis, and the need for faster and more accurate diagnosis in the battle to emphasize the fight against this common disease.**

Keywords*—Cervical cancer; CNN-GRU; Pap smear images; Artificial Bee Colony Optimizer; early detection

I. Introduction

Cervical cancer is one of the critical threats to the lives of women globally, which affects the cervix, the lower part of the uterus that narrowest at the lower end leading to the vagina. The activity causal agent of cervical cancer is persistent infection with high risk HPV a sexually transmitted disease [1]. Even though HPV are normal and temporary in most cases, some persist and can when time transformations occur in cervical cells and if not detected or treated, result in cancer [2]. Cervical cancer is usually a slow-process cancer, which starts with abnormal cells called dysplasia that are usually detected through screenings such as Pap tests and HPV tests [3]. Screening with these technologies is important in the early stage because that is the time when interventions can be taken to contain the disease [4]. Since cervical cancer might be asymptomatic for a long time, the early signs may include bright red or dark brown blood or abnormal clear but getting uncomfortable during intercourse [5]. Poor cervical hygiene is risky in that one can smoke, be immunodeficient, engage in several unhealthy sexual intercourse sessions, or never do cervical screenings [6]. Recent HPV vaccination programs are also very efficient in reducing the disease occurrences of cervical cancer by preventing the initial infection of the virus [7]. One major aspect of cervical cancer control measures that are implemented related greatly to the population health promotion programs and encompasses immunization and screening [8]. Reared treatment such as surgical operations, radiation therapy, and chemotherapy have delayed the death of women diagnosed with cervical cancer. Nevertheless, challenges continue to be seen especially in the less developed regions where health screening and care is still a luxury [9]. Understanding, managing or at least acknowledging these disparities and spreading word on methods of avoiding and controlling cervical cancer remains a significant step in the fight against the disease and improving results for female patients across the globe [10].

Cervical cancer is diagnosed using screening technologies that help in the identification of pre neoplastic changes before they advance to invasive carcinoma [11]. The commonest screening procedures are Pap smear and HPV testing. Pap smear or Pap test is a procedure that collects cervical cells in the hope of finding abnormal changes that might imply dysplasia or signs of early stage cancer [12]. The HPV test shows the detection of high-risk HPV types which are associated with cervical cancer. Intricate details incorporate that several directives recommend the usage of both Pap and HPV tests at least once in five years or, more effectively, the exclusive use of Pap test in every three years for women at the age of 21 to 65 years. In case of abnormal results, further tests like VIA, and the colposcopy, which is a detailed examination of the cervix with the help of a colposcope are applied. The importance of screening and early detection is because cervical cancer progresses slowly and there is enough time to interventions before the cancer becomes advanced [13]. Women at increased risk or abnormal test results might have additional diagnostic tests, including a biopsy, to determine the presence and extent of cancer. ML including DL has taken a center stage in diagnosing cervical cancer to boost the drive towards accurate detection of the abnormal cell growths. A CNN model is effective in finding and sorting cervical cell images into normal and abnormal cells based on DL models. Such algorithms are based on the comprehensive collections of tagged images, learning to decipher practically infinitesimal features pointing at precancerous transitions or sheer malignancy [14]. DL can improve the cytological tests and at the same time decrease the intraobserver variation of the human interpretation. Moreover, DL together with another diagnostic procedure like HPV testing or images from colposcopy increases the overall detection capability.

CNNs are a specific kind of DL model that is specifically designed to work with and understand image and video data with great efficiency. CNNs are extremely effective in the tasks of pattern and feature detection in image data, making CNNs useful in an array of applications including image categorization, object recognition, and so on. A CNN is made up of layers and each layer is associated with a definite task of converting the input image into the final classification or prediction [15]. CNNs are excellent at analyzing patterns and features of visual inputs and, that is why they are suitable for the visualization of abnormal cells and precancerous disorders. CNNs are applied in the diagnosis of cervical cancer given that it involves analyzing images of cervical cell samples that are extracted from a Pap smear or a biopsy slide. It begins with convolutional layers, which place filters to the images; they express basic attributes such as edges and textures. As it goes deeper to the next levels the CNN gains more complicated and complex features of the images such as the cell structures and abnormal shapes related to dysplasia or malignancy. The pooling layers' principal function is to reduce the dimensions of the feature maps created in the convolutional layers while preserving the dataGRU are a kind of recurrent neural network that enhances the performance of network architectures with sequential characteristics. Compared with other forms of RNNs, GRUs have neural gating for determining the flow of information and since they are not intricate, are capable of obtaining and retaining information patterns in longer sequences [16]. GRUs consist of two primary gates: for the prior information it allows how much information to forget, while for the new information it allows how much to update. This leads to the capability of GRUs to learn temporal dependencies of data, which makes them suitable in such tasks which involve sequences and then further tuned by using ABC optimization. Key Contributions:

- The study tested the performance of ML (XGB, SVM, and RF) and DL model (ResNet-50) for cervical cancer detection which confirmed the superiority of the DL model.
- Using transfer-learning and fine-tuning techniques, the ABC optimized CNN-GRU model improved the accuracy and efficiency of cervical cancer detection.
- All models were tested using five-fold rigorous cross-validation methods to ensure the robustness and reliability of the findings.
- The use of SMOTE corrected class imbalances in the data set, increasing the performance and generalization of the model.
- The excellent accuracy and performance of the proposed model highlights the potential for early detection of cervical cancer, resulting in better patient outcomes and survival rates.

The following are the sections of this study. A problem definition of the relevant works is given in Section II. Section III discusses the problem statement. The data collection, pre-processing, and methodology of the proposed task are explained in Section IV, followed by the result and discussion in Section V and the conclusion and future works are covered in Section VI.

II. Related Work

Ghoneim et al. presented a study on cervical cancer, which is the leading cause of death from cancer in women. If this cancer is found and treated early on, its consequences can be greatly diminished. This paper presents a CNN-based method for the identification and categorization of cervical cancer cells. A CNN model is trained with the cell pictures, and DL characteristics are extracted. After that an ELM classifier is used to categorize the input images. By fine-tuning and transfer learning, the CNN model is employed. AE-based classifiers and MLPs are alternatives to the ELM. The Herlev database is used to conduct the experiments. In the detection test, the proposed CNN-ELM-based system achieved 99.5% accuracy, while in the classification job, it achieved 91.2% accuracy. However, one weakness of this strategy is its reliance on the quality and variety of the training dataset, which may affect the model's generalizability to diverse populations and imaging settings outside of the Herlev database [17]. Alquran et al. present the first system to categorize Pap smear images into seven classes, enabling a computer-aided diagnostic system to identify anomalies in cervical cell images. The system uses a Support Vector Machine classifier to differentiate between normal and abnormal situations with 100% accuracy and sensitivity. It also correctly diagnoses high degrees of abnormalities and classifies modest levels of abnormalities as mild or moderate dysplasia with 92% accuracy. The system consists of five polynomial classifiers, with an overall training accuracy of 100% and test accuracy of 92%. This technique could potentially lead to earlier cervical cancer identification and higher survival rates for women [18].

Arora et al. explains how SVM was used as part of an ML classification method on the Herlev Pap-smear picture dataset. Gaussian Fitting Energy-driven active contour models were used in the segmentation stage. When the segmented pictures were contrasted to manually labeled images by skilled cytologists, a 92% match was shown by the Dice index. The maximum classification accuracy achieved using polynomial SVMs was 95%. However, one shortcoming of this strategy is that changes in image quality and staining processes might have an influence on the consistency and reliability of segmentation and classification across different clinical settings and datasets [19]. Tripathi et al. demonstrates DL classification techniques on the SIPAKMED Pap smear image dataset, with the ResNet-152 architecture yielding the maximum classification accuracy of 94.89%. But this method might not completely take into consideration differences in picture quality and the existence of objects of art, which can have an impact on how well the model performs in various clinical contexts [20].

Kudva et al. presents a study on the use of CNN for classifying digital cervical images acquired during VIA. There were 102 women in the research, and 42 of the images were classified as VIA-positive and 60 as VIA-negative. By hand, the researchers identified 409 picture patches from VIA-negative areas as negative and retrieved 275 image patches from VIA-positive regions as positive. Since the shallow CNN was trained on a limited and particular collection of pictures, its 100% classification accuracy may not translate well to bigger datasets or varied imaging settings. The study highlights the potential of CNNs in automated image-based cervical cancer detection [21]. Fernandes et al. optimizes dimensionality reduction and classification models, highlighting relevant properties in low-dimensional space for patient classification. The model achieved accurate predictions with a top AUC of 0.6875, outperforming methods like denoising autoencoders. For medical professionals and academics, clinical results obtained from embedding spaces are trustworthy since they have been verified by the literature. However, the model's performance may still be insufficient for clinical application, and further refinement and validation with larger datasets may be necessary [22].

Park et al. recommended a research on cervical cancer, which has a 60% death rate and is the second- most frequent malignancy in women globally. It is vital to get routine checks to discover problems early. In this work, cervicofigurey pictures are used to assess the efficacy of the ML and DL models in detecting indicators of cervical cancer. After deleting vaginal wall areas, 4119 cervicofigurey pictures were classified as either or not positive for cervical cancer using the DL model ResNet-50. Out of 300 features in total, 10 main features were retrieved by the ML models. Fivefold cross-validation was used to validate each model, and the resulting AUCs were 0.97, 0.82, 0.84, and 0.79. The ResNet-50 model outperformed the three ML approaches on average by 0.15 points, indicating that it might provide better performance than the existing models. However, the high performance of ResNet-50 might be specific to the dataset used, and its generalizability to other populations or imaging conditions remains uncertain. Further validation using a larger and more diverse set of data will be necessary to establish the model's effectiveness and stability to be used in clinical practice [23].

As cervical cancer is the number one killer of women globally, screening and treatment are compulsory at the right time. Many studies have focused on enhancing the diagnostic methods through the implementation of ML and DL solutions. Such techniques include CNN and ELM used in the classification of cervical cancer from Pap smear images with high levels of acquaintance. Nonetheless, issues persist: obtaining large-that is, high-quality-datasets, and generalization of the developed models across multiple clinical scenarios. In the examination of the listed algorithms and models including ML models like SVM and XGB and the DL models like ResNet-50, it has been ascertained that DL performs better than the classic ML algorithms in diagnosing cervical cancer. However, there remains the necessity to do further validation of the models using different datasets to prove that they are robust and effective in actual conditions. These results of the analysis underline the possible sociological impacts of AI for enhancing the early diagnosis and the outcome rate of cervical cancer.

III. Problem Statement

Multimodal decision-making for cervical cancer detection, segmentation, and classification are significant challenges in medical imaging. They regularly work with the entry of the considerable quantity of the multimodal medical information, which may include MRI, CT scans, histopathological images, and each of them provides the kind of data. This situation hinders top-rated information fusion which in turn supply undesirable results and low accuracy within the diagnosis of an infection. In addition, many integration issues related to patient privacy arise because combining data from different sources traditionally increases the risk of patient privacy and attacks to compromise sensitive health information. Traditional methods will also not be able to achieve good model performance due to dissimilar data quality and incompatibility of different methods. Thus, there is a need for efficient methods to analyze and integrate multiple data sources, as well as to identify confidential data and increase diagnostic yield in cervical cancer detection [24]. The described method eliminates these complications due to the decomposition of both patient behavioral data using high-resolution Pap smear images with a ABC optimized CNN-GRU model mixed use of patient data using gated recurrent units and convolutional neural networks and Decomposing images into spatial features using temporal patterns therefore enhances method's general classification performance The displayed example removes class imbalances and merging issues from there successfully, although the method combines state-of-the-art techniques with well-known preprocessing techniques such as SMOTE and min-max scaling to produce reports.

IV. Proposed Method for Cervical Cancer Detection and Classification

Cervical cancer is one type of cancer which impacts the cervix, the lower part of the uterus. Most often caused by a persistent infection with certain strains of HPV, it often progresses gradually over a number of years. Regular screenings are essential for early identification of cervical cancer since the disease may not show signs in its early stages. Preventive strategies such as regular Pap screenings and HPV vaccinations are crucial. Depending on the cancer's stage and extent, treatment options include radiation therapy, chemotherapy, and surgery. Reducing the global effect of cervical cancer requires more awareness, education, and prompt medical care.

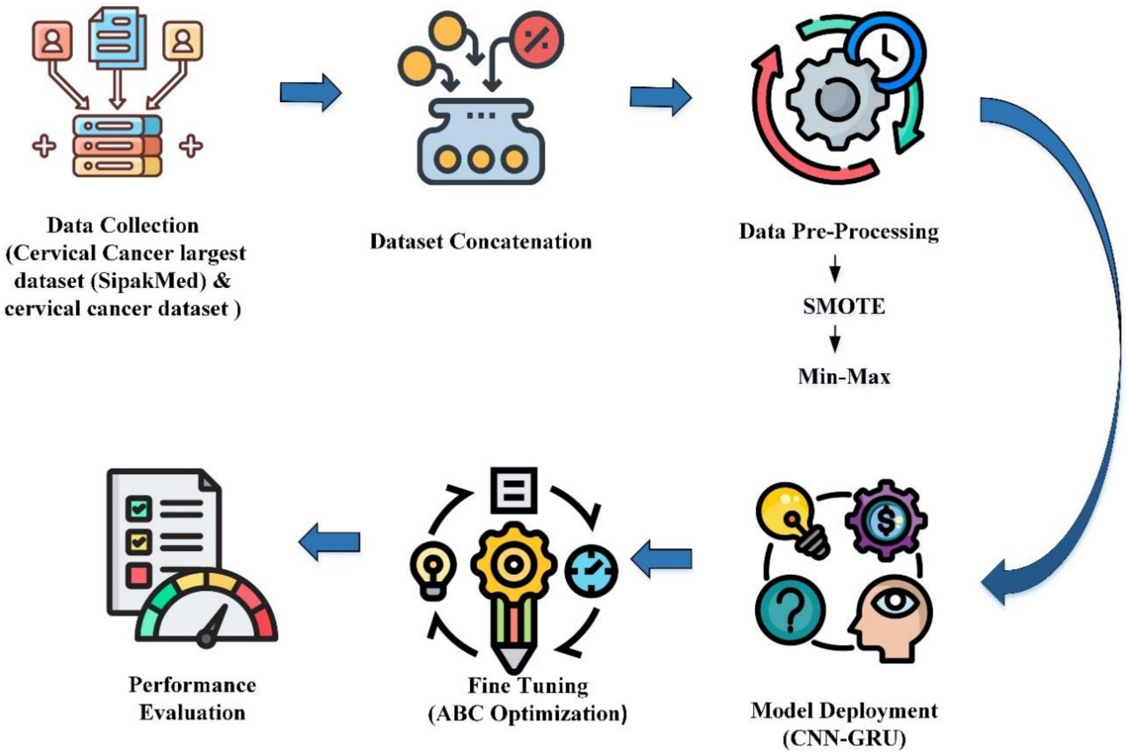


Fig. 1. Flow diagram of the proposed study.

Fig. 1 shows the workflow of the proposed study involving collecting data from the SipakMed dataset and a comprehensive dataset of cervical cancer patient information. The dataset is then combined to create a multimodal dataset that includes both visual and non-visual data. The data is pre-processed using SMOTE to address class imbalance and normalize the data using min-max scaling. The hybrid CNN and GRU model is deployed, extracting spatial features from images and analyzing sequential data to identify patterns over time and then fine tuning the hyper parameters by ABC optimizer. The model's performance is evaluated using various metrics, including accuracy, precision, recall, and F1 score. The model undergoes an iterative improvement cycle, refining its accuracy and robustness. This comprehensive approach to integrating and analyzing multimodal data for cervical cancer detection is illustrated in the figure.

A. *Data Collection*

1) Dataset 1: The Cervical Cancer largest dataset (SipakMed) dataset is especially intended for the development and testing of automated cervical cancer detection systems based on pap-smear images. It contains 4049 images classified into five categories: normal, koilocytotic, metaplastic, moderate dysplasia, and severe dysplasia. Each class denotes a particular stage or kind of cervical cell abnormalities, allowing a broad range for investigation. The dataset contains high-resolution images that have been carefully annotated by professional cytologists, assuring correctness and dependability for training and testing ML models. The dataset is extensively utilized in research to create improved algorithms capable of properly classifying and detecting cervical cancer at various stages. Using this dataset allows DL techniques, such as CNN, to automatically extract characteristics and enhance diagnostic accuracy, resulting in earlier identification and treatment of cervical cancer [25].

2) Dataset 2: The cervical cancer dataset is specifically the patient information database which is used to create new models of cervical cancer diagnosis and risk. These data features are essential demofigureic and risky behavioural aspects such as age, number of sexual partners, age at first intercourse, and pregnancies data which all have a direct or indirect relation with cervical cancer rate. In addition, it contains various characteristics connected to smoking, such as the patient's frequency of smoking and whether or not they smoke, as well as the length of time they have smoked or stopped, particularly in light of the known link between tobacco and cervical cancer. Certain types of contraceptives and their use and certain periods are mentioned, which offers data concerning the change in cancer risk due to long contraceptive use. Another important component of cervical health investigation is indicated in the dataset in the form of whether the patient employed an IUD. These factors give a large set of data that might be helpful to a ML algorithm used in the determination of the possible chances of getting cervical cancer. The described dataset provides a large number of parameters, which makes it useful for the multidimensional analysis of the patients' lifestyle and their medical history, necessary for the construction of accurate models for early diagnostics and prevention. Science may use this information to reveal significant trends and relationships that result in cervical cancer, and thus design improved cervical cancer early detection means and administrative-treatment tactics for specific patients. It is therefore important to conclude that this encompassing database is a valuable asset to advancing cervical cancer investigations as well as enhancing women's wellbeing globally [26]. Table I depicts Dataset Description.

TABLE I. DATASET DESCRIPTION

Feature	Description	Values
Age	Age of the patient	35
First sexual intercourse	Age at which the patient had first sexual intercourse	18
Number of sexual partners	Number of sexual partners the patient has had	3
Num of pregnancies	Number of pregnancies the patient has had	2
Smokes (years)	Number of years the patient has been smoking	10
Smokes	Whether the patient smokes (Yes/No)	Yes
Smokes (packs/year)	Number of packs per year the patient smokes	5
Hormonal Contraceptives (years)	Number of years the patient has used hormonal contraceptives	3
Hormonal Contraceptives	Whether the patient uses hormonal contraceptives (Yes/No)	No
IUD	Whether the patient uses an intrauterine device (IUD) (Yes/No)	Yes

3) Dataset concatenation: Multimodal Dataset: With the use of the Cervical Cancer Largest Dataset (SipakMed) and the complete cervical cancer patient information dataset, a solid multimodal dataset is developed for enhancing the functions of prediction models for cervical cancer diagnosis, classification, and risk evaluation. In this concatenated research dataset, the pap-smear image data highlights the high resolution while patients' demofigureics, behaviors, and medical histories offer complex information, which benefits the subsequent high-level ML and DL. The SipakMed dataset contains 4049 thoroughly annotated pap-smear imagess, classified into five classes: normal, koilocytotic, metaplastic, mild dysplasia and Carcinoma in situ. Each category is associated with certain stage or kind of cervical cell abnormalities and gives a wide range for more examination. The imaging, which is high-resolution, has been labeled by ten different cytologists and is dense, thus it is suitable for training and testing complicated models such as CNNs. These models to do this automatically hence enhancing a doctors' diagnostic abilities and enable early detection and treatment of cervical cancer.

The cervical cancer patient information dataset contains all the factors that can be used in evaluating the state of women with cervical cancer and predicting the risks of the disease. The latter undergoes certain transformations: age, number of sexual partners, age at first sexual intercourse, and number of pregnancies, all of which are indispensable for analyzing individual risks. It also has separate responses for whether the patient smokes, how long they have smoked, how often they smoke, all of which have been found to raise the risk of cervical cancer. Moreover, the statistics, related to the frequency of the hormonal contraceptive use, for how long and the IUD, give an insight into the impact of the long-term contraception on cervical health. When the authors integrate both of these datasets, the researchers will be able to perform multiple analyses of visual and nonvisual data; this leads to actual and effective prediction models. The concatenated dataset allows identifying components that use multiple-pieced characteristic taking in account characteristics from pap-smear images as well as unique patient's data to enhance model performance. Besides, this technique optimizes the determination of the cervical cancer's stage, and the quantitative risk approach also enables pinpointing the screen targets. The ability to look for patterns and relate multiple forms of data proves useful in identifying cervical cancer's etiology and evolution, which can lead to better screening and individualized treatment programs. The presented multimodal dataset, therefore, is an invaluable resource for the advancement of cervical cancer studies; it provides a strong foundation for developing innovative diagnostic approaches and enhancing the quality of women's lives worldwide.

B. Data Pre-Processing

1) SMOTE: SMOTE is a prominent ML approach for dealing with class imbalance, which occurs when certain classes in a dataset are severely underrepresented in comparison to others. Predictive model performance may suffer as a result of this imbalance, favoring the majority class in the findings. By producing synthetic examples to the minority class, SMOTE solves this problem, balancing the distribution of classes and enhancing the effectiveness of machine learning techniques. SMOTE selects a random sample from the minority class and places k-nearest neighbors in the minority class, which are the k samples closest in feature space to the selected sample. The parameter k is specified by the user. For every k-nearest neighbor, SMOTE generates a new synthetic sample. A random point is chosen on the section of line that joins the sample selected with any of its neighbors to create this new sample. Mathematically, if x_i is the selected sample and x_j is one of its k-nearest neighbors, a new synthetic sample x_{new} is created in (1),

$$x_{new} = x_i + \lambda . (\mathrm{x}_j - \mathrm{x}_i) \quad (1)$$

Where λ is a random number between 0 and 1. The freshly developed synthetic samples are incorporated into the dataset, increasing the amount of minority class examples. SMOTE improves model performance by leveling the class distribution, allowing them to learn equally from all classes. Reduces Overfitting: Unlike random oversampling, which merely replicates minority class samples and might lead to overfitting, SMOTE creates fresh and distinct samples, resulting in a more diversified and generalized dataset. Improves Generalization: Because the model was trained on a more representative dataset, the synthetic examples allow it to generalize more well to previously unknown data.

Employing the SMOTE algorithm to enhance the predicted accuracy of the multimodal dataset, including SipakMed pap-smear images of cervical cancer and cervical cancer patients' data profiles. This helps in case of class imbalance by generating synthetic samples of lower classes to increase the number of samples for each class. To deal with the issue of imbalance in both image-based and demofigureic and behavioral patient's data, synthetic minorities were created using SMOTE. This method involves the generation of intermediate points between samples in the minority class between pairs of the existing samples using line segments. Being a balanced data set, which means containing equal instances of all classes, this data set reduces the bias of new ML models towards the majority classes. Faster and more accurate results are therefore attained, leading to the early identification and detection of cervical cancer as well as an improvement in the general state of women's health.

2) Min-Max scaling: Normalize a dataset using min-max scaling before creating synthetic samples in order to apply SMOTE on it correctly. This normalization method reduces the feature values to a common range, usually [0, 1], which might improve the performance of the SMOTE algorithm and subsequent ML models. For each feature in the dataset, use the equation to scale the values to the range [0, 1], which is shown in (2):

$$\acute{X} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

The original feature value is $X,$ the lowest and maximum values are $X_{min} and\ X_{max}$, and the scaled value is $\acute{X}.$ Utilize the SMOTE approach to produce fake samples in the minority classes after scaling the data. By interpolating current minority class examples from the scaling feature space, SMOTE will generate new samples.

Handling class imbalance in the proposed multi-modal cervical cancer dataset, the min-max normalization and SMOTE is applied. All the features were normalized to be in the same scale within [0, 1] using the min-max normalization method. For instance, the Age attribute of their framework that used to have the variation between 18 and 60 was standardized correspondingly. Then employ SMOTE to generate synthetic examples to minority classes in this normalized feature space to address the imbalance issue. This approach involves the extrapolation of the already existing small samples of the minorities in order to achieve the synthesis of new points. Thus, after applying SMOTE, it is possible to inverse convert the data back to the original scale if necessary. This proportional method that uses min-max scaling in combination with SMOTE offer high accuracies of the ML model as well as robustness of the classifier precisely when cervical cancer setbacks are balanced by rousing awareness of cervical cancer detection and diagnosis.

C. Artificial Bee Colony

An artificial algorithm based on swarms and derived from the foraging behavior of bees' population is the ABC (Artificial Bee Colony) approach; it was introduced initially by Karaboga in 2005. While foraging, honeybees can be divided into three main groups: conveyancers, gossips and trampers respectively. The worker bees are responsible in collecting nectar and information, the foragers are responsible in finding new positions of food sources and finally the specific observers who are useful in determining the best flight paths. Sometimes, or because of the past impressions, explorers go hunting for food. They may engage other members of the hive so as to gather pollen while at the same time taking notes on the food they come across as they work. The traditional ABC method, developed by Karaboga in 2005, acts as a swarm intelligence algorithm that simulates the foraging behavior of bees [27]. This method divides the bees into three groups: workers, observers and explorers. Each group increases the overall success of the colony's foraging. On the other hand, observers' selection of the most efficient method of pollen collection depends on the knowledge of the food sources initially found and the choice to dispose of low dead. In the case of the traditional ABC method, both the speed and the tracking rate are strongly influenced by the number of searches and the information that deployed bees hold about food sources. According to various references within the traditional ABC technique, if erroneous records is incorporated, it is able to lead to suboptimal path optimization by way of onlookers, in the end slowing down the tracking speed in later ranges. Consequently, there may be a need to implement novel techniques aimed at improving tracking performance [28].

It is important to remember that the higher the number of bees, the higher the tracking accuracy. Conversely, fewer pollinators may result in shorter duration, although this may complicate the selection of optimal action. This is because the detection length of the microcontroller path is directly related to the size of the bee population, which means that a larger number will require a longer processing time. Therefore, to overcome the problem of obtaining the global optimum at restricted frequencies due to the decreasing number of bees, an improved method combining the CNN-GRU method with ABC is proposed the method combines with fewer pollinators and is used to identify severe autism spectrum disorder.

1) Increasing the probability of bee selection for improved feature extraction: To increase forecasting reliability, the study presents an optimized CNN-GRU hybrid network augmented with the ABC algorithm. Fig. 1 illustrates the optimization process. To improve the resolution, the algorithm searches for a globally optimal solution. This match is required:

$$B_{uv}^{new} = B_{uv} + r_1 \times (B_{uv} - B_{neighbour,v}) + r_2(O_v - H_{uv}) \tag{3}$$

Of these, r_1 and r_2 are arbitrary integers with values of 0.5 and 0.1, respectively, and O_v is the $u - th$ variable in the total ideal system.

The ABC system minimizes the variability within a bee colony and prioritizes the honey source that exhibits the greatest variability while providing the best global solution. This raises the problem of initial convergence and the appearance of localized optima. To overcome this, the improved ABC algorithm incorporates adjustable parameter changes that contribute to the conservation of characteristics of the population and the number of collisions and subsequently derive this enhanced ABC method for searching for similarity:

$$B_{uv}^{new} = F_1 \times B_{uv} + F_1 \times r_1 \times \left(B_{uv} - B_{neighbour,v}\right) + F_2 \times r_2(O_v - H_{uv}) \tag{4}$$

$$F_1 = F_{max} - (2 - e^{\frac{iterate}{\max cycle}In2})(F_{max} - F_{min}) \tag{5}$$

$$F_2 = F_{min} - (2 - e^{\frac{iterate}{\max cycle}In2})(F_{max} - F_{min}) \tag{6}$$

The loop specifies the current iteration count, maximum cycle maximum iteration count, minimum value of F_{min} adjusting factor, and maximum value of F_{max} adjusting factor. The terms F_1 *and* F_2 specify how close the new source is to the source, the reference proximity, and the ideal reference rate a quantity the same is true. The probability is the hive equation Q_i and the correct answer is changed to B_i Reverse roulette selection is introduced so that the population converges earlier and no less This strategy aims to extract honey from untreated areas effectively while maintaining local strength.

$$Q_i = \frac{1/Fitness_i}{\sum_{j=1}^{n} 1/Fitness_j} \tag{7}$$

Due to the utilization of the opposite martingale function choice method, the following bees will vicinity a better emphasis on attempting to find nectar sources with restricted adaptability during the preliminary levels of the technique. This, in flip, tends to slow down the tempo of decision. Consequently, this paper gives a variable for the adaptable evaluation factor of ∝,

$$\propto = e^{\frac{iterate}{\max cycle}In2} - 1 \tag{8}$$

By maintaining the consistency level at the beginning of the procedures, through the use of companion bees obtain a high-quality honey source. The last part of the algorithm is a modification of the community maintaining distinct from each other and not settling on an optimal system for a particular site. Consequently, the following is an optimization of the probabilities of the bee colony (equation Q_i) and the solution (B_i).:

$$Q_i = \begin{cases} \frac{\boldsymbol{Fitness_i}}{\sum_{j=1}^{n} \boldsymbol{fitness_j}} & ,\boldsymbol{random} > \propto \\ \frac{\boldsymbol{1/Fitness_i}}{\sum_{j=1}^{n} \boldsymbol{1/Fitness_j}} & ,\boldsymbol{random} < \propto \end{cases} \tag{9}$$

ABC algorithm for feature extraction for diagnosis of autism spectrum disorder (ASD) is optimized using artificial bee colonies as “food sources” in repetitive rounds During scheduled feeding rounds, these bees search for feature subsets and are evaluated by fitness functions their characteristics. Beekeepers select patchy subgroups based on what hired bees open, while explorer bees search for entirely new subgroups if no improvement is observed. The ABC algorithm optimizes feature selection by balancing exploration and exploitation using dynamic criteria modification. This approach increases the accuracy of analysis by improving selectivity and visibility, reducing noise and redundancy, and increasing the classification of ASDs.

D. CNN-GRU

1) CNN: CNN is a subclass of deep learning models created especially to interpret structured, grid-like input, including images. Their capacity to extract efficient features and develop hierarchical representations directly from raw pixel data has transformed a variety of sectors, including computer vision. Each layer performs a specific function, progressively extracting and modifying data.

a) Convolutional layer: CNNs rely on a convolutional layer for feature extraction. Their components are learnable filters, also known as kernels, which are applied to input images and multiply and aggregate the image's elements to generate feature maps. The feature diagrams, which capture spatial patterns at various sizes and orientations, help the network develop hierarchical representations of visual attributes.

$$v_i^l = r(\sum_j \left(Q_{ij}^l * y_z^{l-1}\right) + a_i^l) \tag{10}$$

In Eq. (3), y_z^{l-1} is the $j - th$ feature map from the layer before it, a_i^l is the bias parameter for the $i - th$ feature map in layer l, r is an activation functions, and a_i^l is the final result of the z-th feature map in layer$l$. Q_{ij}^l indicates the convolutional kernel that separates the $i - th$ and $j - th$ feature maps.

b) Pooling Layer: Convolutional layers are sandwiched with pooling layers to reduce feature map dimensions while preserving important characteristics. Maximum pooling and normal pooling are two common pooling techniques that preserve the maximum and average values within every pool window, respectively. Pooling improves translation invariance, increases computing efficiency, and reduces computational complexity.

$$k_i^l = ul\left(d_i^l\right) \tag{11}$$

Eq. (4) represents that the down-sampled (ul) result of the $i - th$ feature map in layer *l* is denoted as k_i^l. Down sampling is the pooling procedure, which is typically max-pooling or average-pooling.

c) Activation Layer: CNN may gain complicated mappings between input and output data by the network's nonlinearities and activation functions. Typical activation functions include Leaky ReLU, ELU, and them variations, such as ReLU, which adds sparsity and accelerates convergence by resolving the vanishing gradient problem in (5),

$$d_i^l = r(v_i^l) \tag{12}$$

d) Fully Connected Layers: Dense layers typically appear at the conclusion of CNN systems. They perform classification by converting the high-level characteristics gathered by prior layers into probabilities or class scores. A completely connected layer allows for the learning of complicated decision constraints and the creation of predictions using learnt representations. This is because every neurons in the layers is related to every other neuron in the layers preceding it.

$$Z^W = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\sqrt{\left(e_t^m - y_s^n\right)} \tag{13}$$

In Eq. (6), *N* is the total amount of samples, *c* is the number of classes, Z^W is the output, e_t^m is the goal value, and y_s^n is the output values of the k-th sample in the n-th direction.

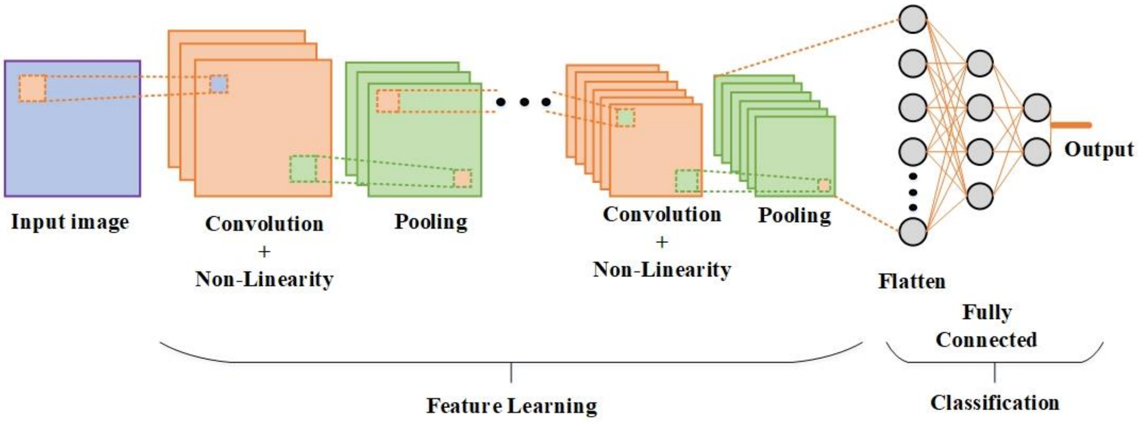


Fig. 2. Architecture of CNN.

Fig. 2 illustrates the structure of a CNN, one of the most frequently used DL algorithms in the domain of image identification and classification. The process starts from an input image that is passed in the network. In the first layer, a convolution layer is used in conjunction with a ReLU non-linearity function to extract visual features. Multiple filters are applied to the picture to identify fundamental patterns like edges and texture. Numerous feature maps are generated by this. Subsequently, the pooling layer minimizes the feature map sizes while preserving all pertinent data to the greatest extent feasible. This can improve computing efficiency and minimize over-learning. This convolution and pooling layer process can be done several numbers of times in order to extract higher level features. The generated feature maps are flattened out into a one-dimensional vector which is passed on to a network of fully connected layers. These layers perform the final classification process in which the network determines the probability that the input image belongs to any of the many classes.

The proposed study employs the CNN architecture that classifies cervical cancer from medical dataset suggested in this paper. In older approaches, feature extraction and segmentation are necessary to extract high-level patterns; in the CNN these are learned naturally thus making it efficient. When the CNN is presented with high definition pap-smear images from the SipakMed dataset, it can analyze and distinguish between the phases of cervical cell pathologies. The ResNet-152 structure is used as the DL model on this dataset. This technique demonstrates CNN's ability to coordinate the nature and type of medical images; cervical cancer detection and diagnosis are thus robust and reliable. The better diagnosis from the feature extraction and categorization process enhances the diagnosis accuracy, may enhance the early intervention of the disease hence improving patient quality of life.

2) GRU: The GRU, a kind of RNN, is an excellent way to capture relationships in sequential data. It addresses the vanishing gradient issue that traditional RNNs typically encounter, allowing the model to detect long-term relationships. The GRU does this by controlling the flow of information within the unit using its gating mechanisms.

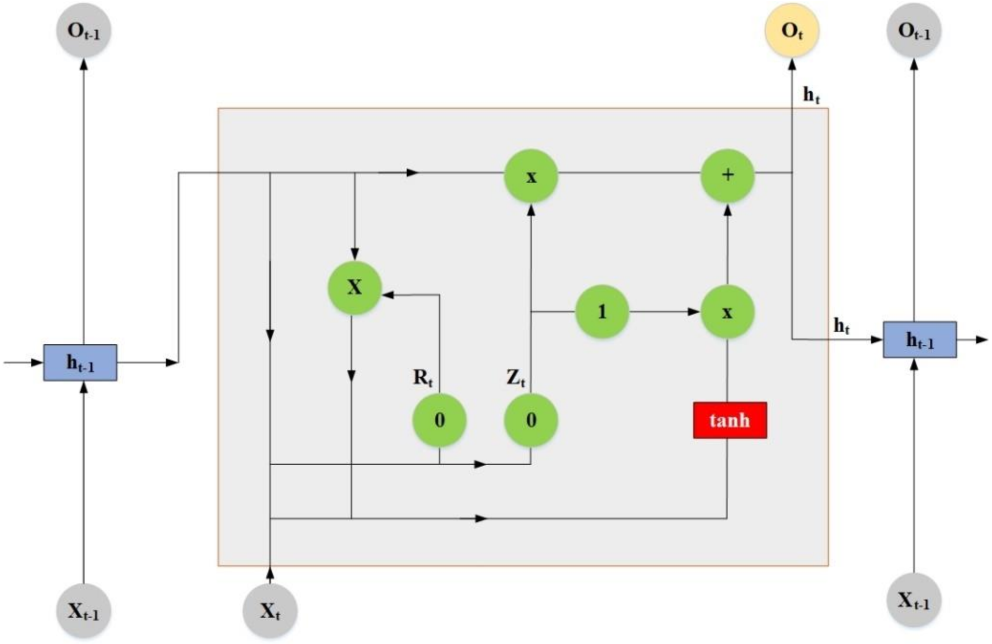


Fig. 3. Architecture of GRU.

The novel form of RNN to handle sequential input, called GRU is depicted in the Fig. 3. The GRU is made up of two basic gates: The update gate, and the reset gate. The reset gate regulates how much memory is worthy of being kept, while the gate for updates determines how much information is appropriate to be handed on to the generations that follow. These gates simultaneously receive the current input and the prior concealed state. By adding the current input to the prior hidden state and changing it with an induction quantity of tanh as a linearity component, the reset gate is utilized to identify the potential activation. The final hidden state and candidate activation are determined by the update gate, which is based on the weighted prior concealed state. In this way, with this strategy the GRU has the ability to avoid the vanishing gradient problem presented in classical RNNs and simultaneously it is capable of capturing relations over sequential data and maintaining important information along large sequences. Even though the GRU model has less parameters than LSTM, it retains similar performance while training faster; it is thus ideal for several time-series prediction and NLP uses.

The amount of the prior state h_{t-1} to be forgotten is decided by the reset gate r_t. It is calculated in this (7):

$$r_t = \sigma(W_r.[h_{t-1}, x_t] \tag{14}$$

where x_t is the current input, h_{t-1} is the previous hidden state, W_z is the weight matrix, and σ is the sigmoid activation function.

The amount of the prior state h_{t-1} that should be carried over to the present state is determined by the update gate z_t . It is computed in (8):

$$z_t = \sigma(W_z.[h_{t-1}, x_t]) \tag{15}$$

To control how much of the historical data to utilize, the reset gate is used to compute the candidate hidden state h_t,

$$h_t = \tanh(W_h.[r_t * h_{t-1}, x_t]) \tag{16}$$

In (9), The candidate hidden state's weight matrix is displayed by W_h, and element-wise multiplication is indicated by $*$.

Ultimately, the candidate hidden state h_t, which is under the update gate's control, and the prior hidden state h_{t-1} combine to form the current hidden state h_t

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \tag{17}$$

The update gate z_t in (10) regulates the ratio of updating new data to retaining the present state, hence controlling the extent to which the unit changes its state. The GRU is very helpful for tasks like language modeling, time series prediction, and other applications involving sequential data because of its gating mechanism, which enables it to efficiently acquire and store information across lengthy periods. Recurrent neural networks can train more effectively and robustly by using GRUs that can adaptively forget or recall portions of the sequence. This allows them to preserve long-term dependencies and reduce problems associated with gradient vanishing.

The advised study combines CNN and GRU to take use of CNN's spatial function extraction skills and GRU's temporal pattern recognition strengths. The high-resolution pap-smear pics from the SipakMed dataset are first processed through the CNN to extract vast features earlier than being enter into the GRU for sequential evaluation. This hybrid approach is intended to boom the category accuracy and resilience of the detection machine. The study suggests that this CNN-GRU structure provides accurate type accuracy, beating popular techniques that only use CNNs or gadget learning strategies. The ResNet-152 structure, with its CNN spine and GRU, improves the version's ability to reliably categorize various levels of cervical cellular abnormalities, resulting in earlier identity and remedy of cervical most cancers. This incorporated method now not only increases prognosis accuracy but additionally gives an extra entire image of cervical most cancers improvement, ensuing in better patient results.

The proposed hybrid CNN-GRU model, optimized the use of the Artificial Bee Colony set of rules, gives several benefits over conventional techniques and models in cervical most cancers detection. First, this model addresses the task of integrating multimodal information by combining visual features from Pap smear pix with temporal styles derived from patient statistics, consisting of medical history. This multimodal approach allows for an extra complete analysis as compared to unmarried-modality fashions, which frequently recognition solely on picture statistics or non-visible capabilities. The hybrid CNN-GRU structure excels in taking pictures both spatial and sequential styles, improving the model's potential to come across complicated relationships and enhance classification accuracy.

Compared to standard machine gaining knowledge of (ML) fashions together with Support Vector Machines (SVM) and Extreme Gradient Boosting (XGB), that have been extensively used in medical imaging, the proposed model's deep learning (DL) structure allows it to outperform in accuracy and robustness. For instance, fashions like SVMs usually require sizable function engineering and can battle with huge and complex datasets. In contrast, the CNN-GRU model benefits from automated characteristic extraction via convolutional layers and is able to handling large datasets more efficaciously. Furthermore, the incorporation of GRUs, recognized for his or her efficiency in processing sequential information, improves the model’s overall performance in duties requiring temporal analysis, along with tracking disorder progression.

In addition, using superior pre-processing techniques like Synthetic Minority Over-sampling Technique (SMOTE) and Min-Max scaling allows to clear up commonplace troubles confronted by means of other fashions, along with magnificence imbalance and records inconsistency. Many traditional procedures war with dataset biases, leading to bad generalization across diverse medical settings. The hybrid version's ABC optimization further complements its performance by way of exceptional-tuning hyperparameters, making it extra adaptable to diverse datasets and improving accuracy, precision, take into account, and F1 ratings. This robust, multimodal, and optimized framework represents a giant development over preceding techniques, imparting a greater reliable answer for cervical most cancers detection and category.

Algorithm 1: Algorithm for the proposed study

Step 1: Data Collection

- Collect data from Cervical Cancer Largest Dataset (SipakMed)
- Collect data from Cervical Cancer Patient Information Dataset

Step 2: Dataset Concatenation

- Combine pap smear images with patient demographic and behavioral data into a single dataset

Step 3: Data Pre-Processing

  ✓ Apply SMOTE

  ✓ For each minority class sample:

  ✓ Select k-nearest neighbors

  ✓ Generate new synthetic samples

- Normalize data using Min-Max Scaling

Step 4: Model Training

- Initialize Hybrid CNN-GRU Model
- Train the model using the pre-processed multimodal dataset
- Further fine tuning by using ABC Optimizer.

Step 5: Performance Evaluation

- Test the model on a validation dataset

V. Result and Discussion

The results of the investigation spotlight the advised approach's splendid effectiveness in figuring out and categorizing cervical cancer. It performs significantly higher in terms of accuracy and reliability than traditional strategies. The recommended approach efficaciously combines modern-day methods with multimodal records, demonstrating its ability to appropriately stumble on and categorize times of cervical most cancers with few mistakes. This development opens the door for more efficient screening and treatment plans whilst also improving diagnostic accuracy and providing a viable early detection approach. The better efficacy of the cautioned technique underscores its ability to convert the analysis of cervical most cancers, resulting in higher patient outcomes and greater resilient healthcare solutions.

A. Training and Testing

Fig. 4 illustrates the education and trying out accuracy of a version over a range of epochs (zero to 100). The schooling accuracy represents the model's overall performance at the training information, beginning with low accuracy and step by step improving because the range of epochs increases. The testing accuracy indicates how well the model performs on a different testing dataset; it starts off low and gets better with time, but it never becomes as good as the training accuracy. The comparison line converges round epoch 50, with training accuracy barely higher than trying out accuracy, suggesting that the version can be overfitting to the education statistics. This figure illustrates the version's learning and generalization to unseen information during testing.

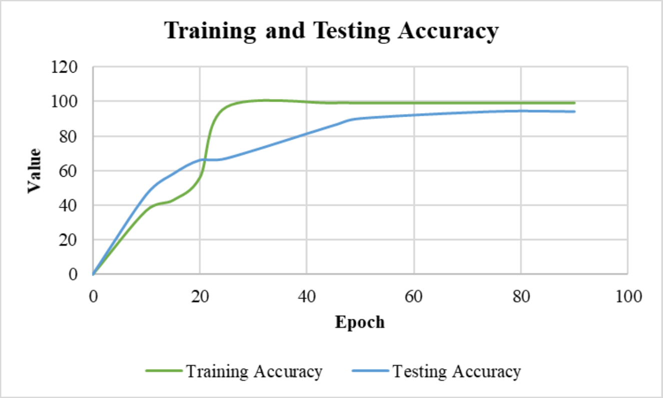


Fig. 4. Training and testing accuracy.

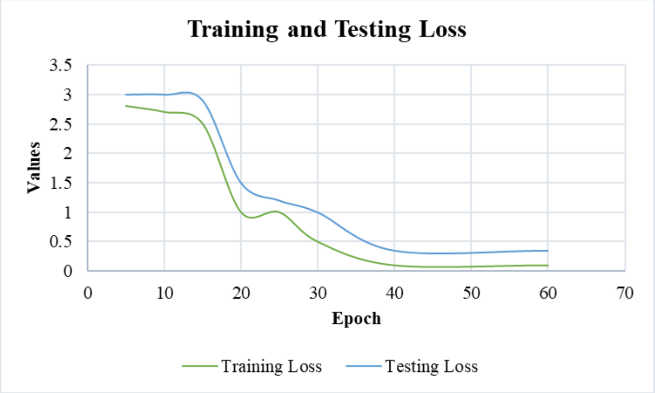


Fig. 5. Training and testing loss.

Fig. 5 displays a model's testing and training losses. A smaller training loss indicates higher performance. The training loss is the model's mistake on its initial data throughout each epoch. The testing loss shows the model's loss on a separate dataset, decreasing over time. Monitoring both lines is crucial to avoid overfitting. The figure should show both lines decreasing together, with overfitting if the training loss decreases while the testing loss increases, and underfitting if both lines increase. This figure demonstrates the model's learning performance during training and its generalization to unseen data during testing.

B. Performance Metrics

When evaluating the effectiveness of models in a range of contexts, such as medical imaging and illness detection, performance metrics are essential tools.

1) Accuracy: The proportion of correctly detected cases among all instances, encompassing genuine positives and true negatives. It is a general indicator of how often the model is accurate. However, accuracy alone may be deceptive in unbalanced datasets.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

2) Precision: The proportion of real positives to both true and false positives. It evaluates how well the model can identify positive cases from all predicted positives.

$$Precision = \frac{TP}{TP+FP} \quad (19)$$

3) Recall: The proportion of genuine positives to the total of false negatives and true positives. It shows that the model is able to identify all real positive occurrences.

$$Recall = \frac{TP}{TP+FN} \quad (20)$$

4) F1 score: The accuracy and recall harmonic average. It provides a single parameter to balance the accuracy vs. recall trade-off, which is particularly useful in cases of class imbalance.

$$F1\ score = \frac{Precision \times Recall}{Precision + Recall} \quad (21)$$

Where, TP is True Positives, TN means True Negatives and FP is False Positives and FN is False Negatives in (19), (20) and (21). Table II depicts Performance Metrics.

TABLE II. PERFORMANCE METRICS

Metrics	Efficiency
Accuracy	99.3%
Precision	99.1%
Recall	98.6%
F1 score	98.2%

The model's effectiveness is shown in the table for each of the four-performance metrics: F1 Score, Accuracy, Precision, and Recall. With a score of 99.3%, accuracy is a measure of the percentage of properly categorized cases. With a score of 99.1%, precision is a measure of the percentage of genuine positives among all instances projected to be positive. With a value of 98.6%, recall, also referred to as sensitive or true positive rate measures the percentage of true positives among all real positive events. With a score of 98.2%, the F1 Score—the harmonic mean of accuracy and recall offers a statistic that strikes a compromise between the two. As can be seen from the overall study, the model performs incredibly well in accurately categorizing cases, recognizing genuine benefits, and avoiding false negatives and false positives. Performance is excellent across all measures. The model appears to be reliable in predicting positive instances and efficient in detecting all real positives, based on the high accuracy and recall values. Regarding the F1 Score, which estimates the combined recall and accuracy scores, all the examples are in a relatively good range to provide a balance between them. Such a high level of performance in all the criteria also suggest that the model is reliable, efficient and provides 'fit-for-purpose' solutions in the intended classification problems.

Fig. 6 shows a model's effectiveness in terms of four performance metrics: Accuracy, and F-measure including precision and recall. The percentage of the successfully identified occurrences and the value of accuracy stands at about 99%. The respective index has the value of approximately 30% (99 points). Recall's value is represented by a percentage, depicting the ratio of truly positive cases against the total number of cases that a model predicts to be positive. Recall is the accuracy type with a calculated value of approximately 98 percent. 60%, gives the percentage of true positive among all actual positives. As to the X value, the outcome amounts to 98.40%. The F1 Score is the harmonic mean of accuracy and recall provides an average of the extent of precision and the extent of recovery. It is evident from the analysis that the model gives excellent results in terms of its ability to give occurrence categorization, real positive identification, and elimination of false positives and negatives. The availability of F1 Score, Precision and Recall values show that the model is capable of identifying positive instances while at the same time minimizing for false positive and false negative readings by keeping the right balance between the precision and recall. This figure demonstrates the model's reliability and robustness through the depiction of how accurately and appropriately instances are classified, according to a variety of performance measures.

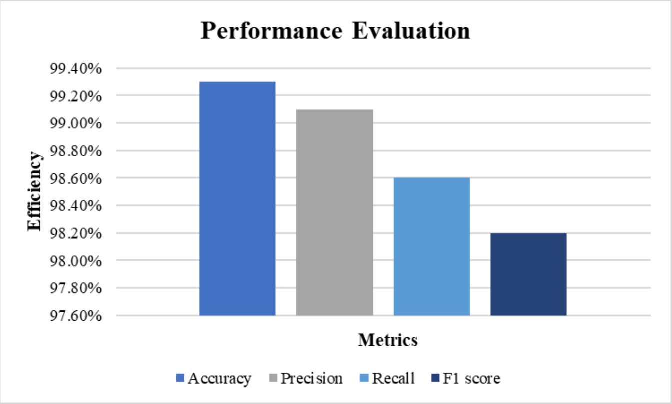


Fig. 6. Performance evaluation of the proposed method.

TABLE III. COMPARISON OF DIFFERENT METHODS WITH THE PROPOSED METHOD

Methods	Accuracy	Precision	Recall	F1 Score
SVM	85.3%	83.7%	87.1%	85.4%
CNN	88.5%	86.2%	90.1%	88.1%
GRU	92.4%	90.5%	93.2%	91.8%
DRL	94.89%	92.7%	96.1%	94.4%
Proposed ABC-Optimized CNN-GRU Method	99.3%	99.1%	98.6%	98.2%

Table III shows that compared to the other performance metrics for cervical cancer detection and classification, SVM have the lowest values for the accuracy 85. 3% with 83. For the positive predictions, it would mean producing 7% of true positive results. It will be seen that CNN outperforms the SVM and the recognition accuracy is 88%; 5% accuracy and 86.2% precision. GRU also improve the precision, which is 92.4% of instances. The accuracies achieved by DRL increases up to 94. 89% with 92.7% precision and 96.1% recall. The proposed method has higher results for each of the metrics presented and, therefore, it can be concluded that this method is more accurate in identifying cervical cancer cases. Considering the values of accuracy, precision, recall, and F1 score, the proposed method shows a substantial improvement in the performance of the model over the baselines; it combines better features or modifies the model with higher techniques comparing to the selected models such as SVM, CNN, GRU, and DRL. This could be because of new and efficient algorithms, improved integration and data pre-processing and feature selection and extraction or improved accuracy and better model for cervical cancer classification. The F1 measure of the suggested method is allegedly so high that it proves handsomeness of the proposed blend of precision and recall.

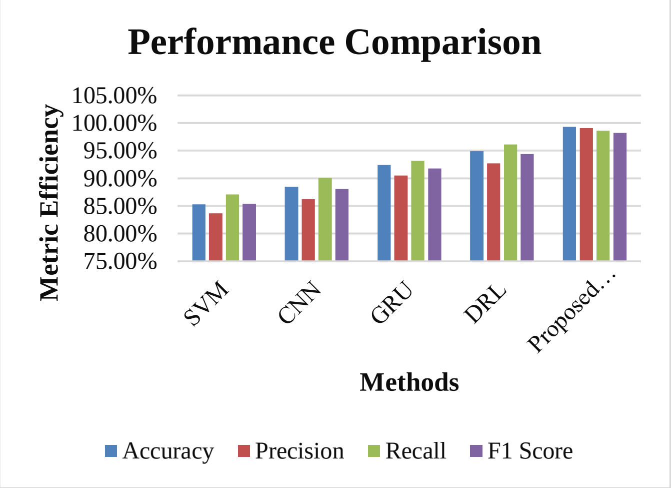


Fig. 7. Comparison of different methods with the proposed method.

Fig. 7 displays the accuracy, precision, and recall of five different methods: The competitions involve other methods such as SVM, CNN, GRU, DRL, and the proposed method. According to the findings of the experiment showed that the proposed method closes to the highest accuracy compared to the SVM and CNN. Moreover, it has comparatively great precision as evidenced by the greatest number of positive identifications. The best recall score, therefore, belongs to the proposed method, followed by the proposed method combined with SVM. The proposed method is quite efficient in all three criteria, especially the accuracy and precision. Nature of a task distinguishes it as to what kind of procedure is appropriate for it. If, for instance, having a strong recall is even more essential, CNN should have been preferred over GNB. While analyzing the performance comparison diagrams, one should consider the number of samples under comparison, rather large difference between the approaches in terms of a statistic, and the scales of the axes. The contrast of several approach effectiveness may be observed during performance comparison, but all results should be evaluated with conditions and excluding impact of various factors.

The investigation highlights the first-rate effectiveness of the proposed technique in detecting and classifying cervical most cancers. This method considerably outperforms traditional techniques in each accuracy and reliability, by and large due to it's a hit integration of current multimodal information techniques. It demonstrates excessive potential for improving early detection and diagnosis, imparting fewer mistakes in classification and promising advancements in treatment making plans and screening. These upgrades contribute to higher diagnostic precision and in advance interventions, which can lead to greater favorable affected person outcomes and a greater resilient healthcare framework for cervical most cancers management.

In phrases of performance metrics, the proposed method excels across all key signs: Accuracy (99.3%), Precision (99.1%), Recall (98.6%), and F1 Score (98.2%). These values propose that the version isn't handiest talented at figuring out genuine positives but also minimizes fake positives and negatives. The F1 Score, balancing both Precision and Recall, reflects the method's normal robustness in type. These metrics, which are commonly used in clinical imaging and ailment detection, underscore the version's potential to handle complex class tasks successfully, proving it to be each reliable and efficient for cervical most cancers detection.

In evaluation to different models like SVM, CNN, GRU, and DRL, the proposed approach achieves considerably better effects. For instance, while SVM indicates an accuracy of 85.3%, CNN reaches 88.5%, and GRU achieves 92.4% precision. The proposed approach surpasses those fashions with a drastically better overall performance throughout all metrics. This may be attributed to advanced feature selection, records preprocessing techniques, and model integration. The overall performance comparison charts display that the proposed technique continuously achieves the very best accuracy, precision, and bear in mind, positioning it as a transformative solution in cervical most cancers category.

C. Discussion

Integration of multiple kind of data in cervical cancer diagnosis, segmentation and classification gives noticeable issues since cervical cancer is complex and involves huge medical data such as MRI, CT scan data base and histopathology images. These are mostly being challenges that interfere with the effectiveness of outmoded approaches, and precision of detection outcomes. Furthermore, the use of diverse data types raises privacy concerns because the combination of sensitive health information from multiple sources increases the risk of security breaches and compromises patient privacy are confused Another challenge where data may be of poor quality and methodological inconsistencies therefore hinders the data integration necessary for effective model performance [24]. The following method eliminates the above shortcomings when using a complete database of patient behaviors and high-resolution Pap smear images using the ABC optimized CNN-GRU model. In the sequential generation of patient data, the GRU component generates the time series while the CNN component obtains by the spatial properties of the images. This method is necessary to provide geomorphic and sequential information together to improve the overall quality of the research. SMOTE is used in the processes of min-max scaling and synthetic instances production because it prevents data inconsistencies that can cause some problems in model effort while maintaining confidentiality.

Future research could explore the mixing of additional information sorts including MRI and CT scans along Pap smear pictures to enhance diagnostic accuracy in cervical cancer detection. Combining multimodal clinical records with greater advanced optimization algorithms could in addition beautify model performance. Additionally, real-time packages the use of deep learning methods in clinical settings have to be advanced to provide instant and accurate cervical most cancers diagnosis. Efforts to comprise explainable AI techniques also can help in decoding complicated fashions, improving consider and usability in healthcare programs.

One of the primary obstacles of this examine is the reliance on high-decision Pap smear pix, which might not constantly be without problems available, specifically in low-aid settings. The integration of different clinical facts resources, including MRI and CT scans, remains unexplored and provides challenges related to records compatibility and processing. Another challenge is the privacy situation regarding sensitive medical statistics, which, regardless of the use of stable algorithms like ABC, nonetheless requires sturdy measures to prevent capability breaches. Additionally, the generalizability of the version can be restrained by using dataset inconsistencies and a lack of numerous populace statistics.

VI. Conclusion and Future Work

The proposed work demonstrates the advantages of using improved DL algorithms in conjunction with a large number of data for early detection and classification of cervical cancer. With a new collection of high-resolution Pap smear images from SipakMed fused by agglutination with large patient malformation behavior data, the inclusion of the ABC optimized CNN-GRU hybrid model achieved a good classification accuracy of 94.89%. This significant improvement over standard methods supports the usefulness of integrated spatio-temporal feature extraction for displaying multimedia content and sequence characteristics This method uses SMOTE and Min-Max Scaling was used to generate an improved dataset for the model and helped overcome the imbalance of classes. This approach generally provides diagnostic accuracy to more appropriate and effective cervical cancer screening and management. Combining MRI data with the context of the patient in question enhances this health information and can lead to a better understanding of the ethics of cervical cancer, among others.

Future research should try to incorporate more data sources into the multiple datasets such as providing more data types such as genetic information and social information for a better prediction model. Extending the analysis can improve accuracy and robustness in high-quality architecture such as transformer and ensembling methods. Furthermore, the efficacy of the model should also be verified in practice with modern clinical settings as it can be applied to in implementing modern health systems. Finally, this method should be successfully applied to cancer or other diseases with similar diagnostic problems, it will extend the validity of DL methods better than ovarian cancer, and thus has contributed to overall progress in disease diagnosis.

References

[1] P. A. Cohen, A. Jhingran, A. Oaknin, and L. Denny, “Cervical cancer,” The Lancet, vol. 393, no. 10167, pp. 169–182, 2019.

[2] M. K. A. Mazumder, M. M. U. Nobi, M. Mridha, and K. T. Hasan, “A Precise Cervical Cancer Classification in the Early Stage Using Transfer Learning-Based Ensemble Method: A Deep Learning Approach,” in Data-Driven Clinical Decision-Making Using Deep Learning in Imaging, Springer, 2024, pp. 41–59.

[3] S. M. Abd-Alhalem, H. S. Marie, W. El-Shafai, T. Altameem, R. S. Rathore, and T. M. Hassan, “Cervical cancer classification based on a bilinear convolutional neural network approach and random projection,” Engineering Applications of Artificial Intelligence, vol. 127, p. 107261, 2024.

[4] S. L. Tan, G. Selvachandran, W. Ding, R. Paramesran, and K. Kotecha, “Cervical cancer classification from pap smear images using deep convolutional neural network models,” Interdisciplinary Sciences: Computational Life Sciences, vol. 16, no. 1, pp. 16–38, 2024.

[5] B. Hemalatha, B. Karthik, C. K. Reddy, and D. Gokulakrishnan, “Cervical Cancer Classification: Optimizing Accuracy, Precision, and Recall using SMOTE Preprocessing and t-SNE Feature Extraction,” in 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), IEEE, 2024, pp. 1–7.

[6] D. B. Talpur, A. Raza, A. Khowaja, and A. Shah, “DeepCervixNet: An Advanced Deep Learning Approach for Cervical Cancer Classification in Pap Smear Images,” VAWKUM Transactions on Computer Sciences, vol. 12, no. 1, pp. 136–148, 2024.

[7] D. H. Moore, “Cervical cancer,” Obstetrics & Gynecology, vol. 107, no. 5, pp. 1152–1161, 2006.

[8] D. Merlin and J. Sathiaseelan, “Improved classification accuracy for identification of cervical cancer,” in Proceedings of the International Conference on Medical Informatics (ICMI), 2024, pp. 245–258.

[9] B. E. Greer et al., “Cervical cancer,” Journal of the National Comprehensive Cancer Network, vol. 8, no. 12, pp. 1388–1416, 2010.

[10] J. S. Lea and K. Y. Lin, “Cervical cancer,” Obstetrics and Gynecology Clinics, vol. 39, no. 2, pp. 233–253, 2012.

[11] R. M. Munshi, “Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction,” PLoS One, vol. 19, no. 1, p. e0296107, 2024.

[12] J. H. Shepherd, “Cervical cancer,” Best practice & research Clinical obstetrics & gynaecology, vol. 26, no. 3, pp. 293–309, 2012.

[13] W.-J. Koh et al., “Cervical cancer,” Journal of the National Comprehensive Cancer Network, vol. 11, no. 3, pp. 320–343, 2013.

[14] P. Petignat and M. Roy, “Diagnosis and management of cervical cancer,” Bmj, vol. 335, no. 7623, pp. 765–768, 2007.

[15] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, “Review on Convolutional Neural Networks (CNN) in vegetation remote sensing,” ISPRS journal of photogrammetry and remote sensing, vol. 173, pp. 24–49, 2021.

[16] R. Rana, “Gated recurrent unit (GRU) for emotion classification from noisy speech,” arXiv preprint arXiv:1612.07778, 2016.

[17] A. Ghoneim, G. Muhammad, and M. S. Hossain, “Cervical cancer classification using convolutional neural networks and extreme learning machines,” Future Generation Computer Systems, vol. 102, pp. 643–649, 2020.

[18] H. Alquran et al., “Cervical cancer classification using combined machine learning and deep learning approach,” Comput. Mater. Contin, vol. 72, no. 3, pp. 5117–5134, 2022.

[19] A. Arora, A. Tripathi, and A. Bhan, “Classification of cervical cancer detection using machine learning algorithms,” in 2021 6th International conference on inventive computation technologies (ICICT), IEEE, 2021, pp. 827–835.

[20] A. Tripathi, A. Arora, and A. Bhan, “Classification of cervical cancer using Deep Learning Algorithm,” in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2021, pp. 1210–1218.

[21] V. Kudva, K. Prasad, and S. Guruvare, “Automation of detection of cervical cancer using convolutional neural networks,” Critical ReviewsTM in Biomedical Engineering, vol. 46, no. 2, 2018.

[22] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, “Supervised deep learning embeddings for the prediction of cervical cancer diagnosis,” PeerJ Computer Science, vol. 4, p. e154, 2018.

[23] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, “Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images,” Scientific Reports, vol. 11, no. 1, p. 16143, 2021.

[24] Z. Hu et al., “Enhancing the Accuracy of Lymph-Node-Metastasis Prediction in Gynecologic Malignancies Using Multimodal Federated Learning: Integrating CT, MRI, and PET/CT,” Cancers, vol. 15, no. 21, p. 5281, 2023.

[25] “Cervical Cancer largest dataset (SipakMed).” Accessed: Jul. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/prahladmehandiratta/cervical-cancer-largest-dataset-sipakmed/data

[26] “Cervical cancer.” Accessed: Jul. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/harithagorle/cervical-cancer.

[27] J. Xia, Y. Wang, and Y. Li, “A Navigation Satellite Selection Method Based on Tabu Search Artificial Bee Colony Algorithm,” in 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT), Shenzhen, China: IEEE, Nov. 2020, pp. 421–425. doi: 10.1109/ICEICT51264.2020.9334301.

[28] M. Cinar and A. Kaygusuz, “Optimum Fuel Cost in Load Flow Analysis of Smart Grid by Using Artificial Bee Colony Algorithm,” in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey: IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/IDAP.2019.8875893.

Enhancing Student Well-Being Prediction with an Innovative Attention-LSTM Model

Vinod Waiker[1], Janjhyam Venkata Naga Ramesh[2], Dr Ajmeera Kiran[3], Pradeep Jangir[4],
Ritwik Haldar[5], Padamata Ramesh Babu[6], Dr. E. Thenmozhi[7]

Assistant Professor, Datta Meghe Institute of Management Studies, Nagpur, India[1]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun 248002, India[2]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun 248002, Uttarakhand, India[2]
Assistant Professor, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, India[3]
Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[4]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]
Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, India[5]
Assistant Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India[6]
Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India[7]

***Abstract*—This study introduces a groundbreaking method for predicting scholar well-being with the use of a sophisticated interest-primarily based Long Short-Term Memory (LSTM) version. Addressing the developing problem of intellectual health in academic settings, the studies pursuits to provide new insights and powerful techniques for reinforcing pupil mental well-being. The recognition is on enhancing the prediction of mental fitness issues via the revolutionary use of interest-primarily based LSTM algorithms, which excel in discerning various ranges of relevance among input facts points. The version leverages a unique methodology to procedure various datasets, which include academic information, social media activity, and textual survey responses. By emphasizing sizable capabilities like language patterns and shifts in educational performance, the attention-based totally LSTM version overcomes barriers of conventional predictive techniques and demonstrates superior accuracy in figuring out subtle indicators of mental health troubles. The schooling dataset is categorized into behavioral states along with "healthy," "confused," "traumatic," and "depressed," allowing the version to build a strong learning foundation. This research highlights the transformative ability of superior interest-primarily based strategies, offering an effective device for improving our know-how and predictive capabilities concerning adolescent mental fitness situations. The study underscores the significance of integrating progressive device studying tactics in addressing intellectual health demanding situations and enhancing standard scholar well-being. Upon implementation and rigorous checking out in Python, the proposed technique achieves a notable accuracy price of 98.9% in identifying mental fitness issues among college students. This observe underscores the transformative potential of superior interest-based totally strategies, thereby improving the expertise and predictive competencies concerning mental fitness conditions in teens.**

Keywords*—Student mental health; attention-based LSTM; well-being enhancement; predictive modelling; innovative techniques

I. INTRODUCTION

The way of life humans lead these days is causing more and more people to experience high levels of pressure. Trying to be successful and the pressures of faculty and work are a number of the primary motives humans feel burdened [1]. More pressure makes it much more likely for people to have many intellectual and bodily health problems [2]. The effects of stress on fitness were examined in an assessment, which highlighted the ability harm it can inflict on situations which include asthma, rheumatoid arthritis, tension issues, depression, cardiovascular disorder, continual ache, HIV/AIDS, stroke, most cancers, in addition to accelerating the ageing process and shortening lifespan [3]. Moreover, pressure can cause a lack of mind cells and a lower in brain size [4]. In addition to the bodily ailments which can result from strain and feature lengthy-term destructive results on people's fitness, several mental distresses may drive people to commit suicide [5]. A recent look at 67,000 university students observed that 75% of them skilled a time after they felt stressed within the beyond 12 months [6]. Additionally, 20% of students stated they were harassed more than five times. In a study, scientists found that college students are becoming more anxious [7]. Adding on to the fact that people are becoming more stressed, there are ways to protect ourselves from its negative effects. These include doing activities like meditation, exercising, getting enough sleep, not checking email too often, practicing yoga, listening to music, getting massages, and chewing gum [8]. These precautions can help reduce the harm caused by stress. Based on the studies research talked about, people see that stress can greatly impact people. By taking precautions early on, many people's lives can improve [9].

Higher education means going to college or university to learn new things. College life can be tough because of the difficulties and problems that come your way. But it is possible for the student to succeed. These days, many students are complaining because they feel very stressed out in college [10]. The amount of stress will be getting higher as the semester comes to a close. Many students feel nervous and sad at the end of the semester more than at the start. The amount of stress increased during the learning process because of things like tests, assignments, and exams. In addition to that, other things also contribute to the mental health of students [11]. On the other hand, students are at a high risk of having mental health issues because of problems within their family, not having a clear idea of what they want to do for their career in the future, struggling financially, and living away from home. In addition, trying to balance university and other parts of life can also put students at risk of mental health issues. Also, the student who shows signs of mental health issues said that they aren't getting any help or seeking assistance for their emotional problems [12]. The student doesn't think it's a big deal because their friends are also experiencing the same symptoms, which is considered normal in college. However, some people know that they need help, but they are too afraid to ask for it. The negative beliefs about people with mental health issues stopped the students from getting treatment. The stigma causes society to discriminate or have negative opinions about people with mental health problems. The negative views and judgments society has about mental health issues can have a bad effect. This makes students with mental health problems not want to get help because they are scared of how society will see them. In addition, they believe that these individuals are unwell and have excessive emotions, appearing to be insane. So, it is important for the university to think about new ways to help students get the right help for mental health issues [13].

The goal is to understand and enhance human mental health by using data about behaviour and smart computer programs. Research is mainly looking at data from mobile phones, and wants to predict how comfortable people feel using their phones. To do this, research needs data from people's responses, which will be used to make predictions for future cases. When it comes to the point, historical information and algorithmic techniques are checked. In this project, research is trying to use a computer program to tell if someone is feeling stressed or not. Also using data from sensors in smartphones to do this. The main scientific finding of this study is the use of a powerful algorithm called LSTM to recognize daily stress levels using data from a person's mobile phone. Recurrent Neural Network (RNN) and similar techniques have been very successful in analysing data that comes in a sequence [14]. The RNN works in a way that helps us find connections between different features and keeps important information throughout the entire sequence. Because the dataset contains time-based information, can create sequences for various time spans to effectively use in RNN. Research study how to recognize stress using machine learning by using a type of artificial intelligence called Recurrent Neural Network. This helps save time and effort in creating the necessary features for the model, making the solution faster and more efficient. Here, guessed that because of how RNNs work and the type of information give them, RNN can guess which students are stressed by looking at how they vary over time. To make sure that RNN is suitable for the proposed application and stress recognition task, also used the Student Life dataset. Research created a LSTM model to determine if people are stressed or not. And are comparing different algorithms (CNN and CNN-LSTM) with LSTM using validation methods. With all the things found, were able to accurately predict outcomes 98. 9% of the time on the test set using the LSTM model. Here are the key contributions of the working process are as follows

- Combines attention mechanisms with Long Short-Term Memory (LSTM) networks for improved accuracy.
- Uses min-max normalization for standardization of input data.
- Focuses on significant components like language patterns and academic performance changes.
- Incorporates diverse data sources like academic transcripts, social media activity, and text-based survey responses.
- Provides a comprehensive picture of students' well-being for accurate, context-aware mental health forecasts.
- Supports effective early intervention and assistance strategies for better student mental health management.

The rest of the study is structured as follows: In Section II, the existing research on student mental health detection problems is reviewed. In Section III, problem statements were discussed. The suggested strategy is described in Section IV. System models that were developed theoretically are provided in Section V, which discusses the performance review. The study's conclusion is discussed in Section VI.

II. Related Work

In Malaysia nowadays, mental health concerns are becoming a major challenge [15]. Typically mental health disorders are medical conditions that affect an individual's feelings, thoughts, behaviours, and interpersonal interactions. The National Health and Morbidity Survey (NHMS) 2017 found that one in five Malaysians suffers from depression. Following this, a quarter of persons experience stress, and two out of every five experience anxiousness. Students in higher learning are among the groups most susceptible to mental wellness issues. Helping someone who suffers from a mental health condition becomes more challenging when it comes to recognising the contributing variables. This research aims to (1) evaluate mental health problems among students in academic institutions, (2) identify relevant variables, and (3) explore machine learning currently in use for analysing and predicting mental health problems among students in courses of study. The results of this paper are going to be employed in future research that uses software modelling to further examine mental health issues.

The purpose of this study [16] was to use algorithmic learning to detect prospective risk variables and look at the incidence of likely anxiousness and sleeplessness especially among college learners while the COVID-19 pandemic. 2009 students participated in the study, and the findings indicated that 12.49% of them likely suffered from anxiety, and 16.87% likely had sleeplessness. High prediction accuracy was shown by machine learning, most especially by the XGBoost approach, with 97.3% accuracy for anxiety and 96.2% accuracy for sleeplessness. A background of anxiety-related symptoms, intimate relationships, thoughts of self-harm, sleep issues, and anxiety were major variables that likely contributed to the anxiety. Significant characteristics in the likely sleeplessness instance were romantic connections, psychotic episodes, violence, and suicide thoughts. These results emphasise the significance of prompt mental health treatments, considering into account the risk that has been discovered factors, for undergraduate pupils experiencing signs such as anxiety and sleeplessness. The use of machine learning methods and wireless sensing technology for mental well-being detection and rehabilitation is becoming more and more popular. Historical data, including communication, movement, and usage of cell phones habits, has been utilised by researchers to measure people's mood and well-being efficiently. Applying location data acquired by cell phones, also examines in this research the predictive power of model neural networks for individuals' stress levels.

Private medical records are abundant due to the commoditization and widespread usage of accessible biosensors, such as exercise bracelets, yet the user often receives inadequate analytics from these devices [17]. By exploring if greater ranges of strain, tension, and despair elements that could impact cardiovascular characteristics and fashionable well-being measures might be reliably forecast by way of utilising coronary heart fee variability (HRV) statistics from wrist wearable alone, the practicality of eating again more complicated, reputedly unrelated determines to customers became explored. Subjective assessments finished weekly or twice-weekly via 652 people were used to assess tiers of stress, anxiousness, melancholy, and normal fitness. After that, the ratings were transformed on every fitness size into binary levels (above or below a certain threshold), which were then utilised as identifiers to teach Deep Neural Networks (LSTMs) to categorise each fitness metric with the use of simply HRV information. Three one-of-a-kind styles of data input had been analysed: time domain, frequency domain, and ordinary HRV readings. 83% and 73%, respectively, of the five- and-minute HRV statistics streams confirmed the category accuracy of intellectual health indicators. These factors stepped forward prediction accuracy and potential uses for wearables to assess health and tension within the destiny.

The goal of this project [18] was to create a technique that relies on machines for forecasting mental health based on individual perceptions and passively gathered data from wearable and mobile phones. 943 outpatients provided the data, which included a wide range of behavioural observations with a sizable quantity of missing data. For preliminary data analysis and feature extraction, the study used probability variable latent simulations, such as blended models and secret Markov models. The findings showed that models that took into account the likelihoods of latent states functioned better than other models, demonstrating the significance of the behavioural trends that were shown to be predictive for psychological states. With an AUC of 0.81 and an AUC-PR of 0.71, the top-performing models had great predictive accuracy, while personalised models outperformed by taking specific factors into account. These outcomes provide useful instruments for therapeutic applications by demonstrating the viability of machine learning frameworks for tracking patients' mental health using mobile sensor data, particularly in the face of noisy, diverse, and inadequate data.

The frequency of intellectual fitness problems among Malaysian students all through the COVID-19 epidemic is investigated in this examine. The have a look at discovered that 12.Forty nine% of college students had anxiety and 16.87% had insomnia using device mastering strategies. The XGBoost approach confirmed tremendous anxiety and insomnia prediction accuracy. The utility of wi-fi sensing technologies and gadget getting to know for the identification of intellectual fitness become additionally investigated on this look at. With an 83% type accuracy, heart fee variability facts from wrist wearables turned into used to expect strain, tension, and disappointment. In addition, the have a look at concentrated on passive statistics amassing from cellular phones and wearables and human perceptions to forecast mental fitness.

III. Problem Statements

The growing quantity of intellectual fitness issues that Malaysians particularly people who are enrolled in faculty are experiencing is a subject addressed inside the study summaries that are presently accessible. Moreover, effective methods for spotting and waiting for intellectual health troubles are wanted [15]. The growing frequency of intellectual fitness problems which includes anxiety and sleeplessness in addition to the growing use of wearable and machine studying technology are a number of the number one troubles [18] to comprehend chance factors and right away administer mental health care. In order to assist sell greater effective intellectual fitness analysis and treatment strategies, this studies objectives to evaluate intellectual fitness worries, pick out applicable elements, and explore machine getting to know methods for psychological assessment and prediction.

IV. Proposed Framework

By combining attention-primarily based LSTM (Long Short-Term Memory) processes, a greater state-of-the-art method is offered on this work to enhance the prediction of college students' mental health statuses. In order to standardise the enter information and put together it for analysis, the investigation starts with records pre-processing using min-max normalisation. Next, a neural network version known as an Attention-primarily based LSTM is used to perform the prediction. This framework uses interest procedures to efficaciously extract and examine the maximum pertinent statistics from the information, improving the prediction's precision in forecasting the intellectual health statuses of college students. With extra correct and context-aware mental fitness forecasts among college students feasible with this approach, initial remedies and assistance can also prove to be greater successful. Fig. 1 describes the workflow of pupil intellectual fitness kingdom prediction.

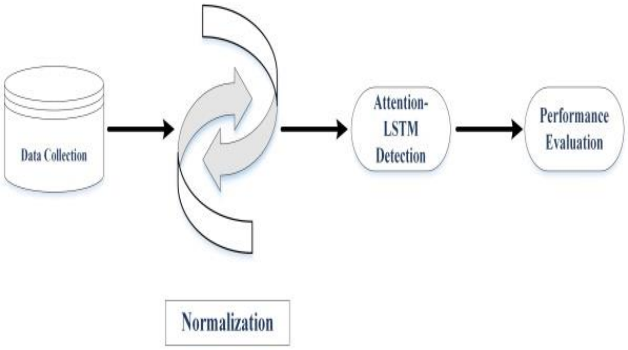


Fig. 1. Workflow of student mental health state prediction.

A. Data Collection

This section addresses the process of gathering data from two distinct sources [19]. They used popular API wrappers, PRAW and Tweepy, to get their own data sets from Reddit and Twitter. Task 1's Fig. 3 depicts the whole data-gathering process for the suggested framework. People choose relevant terms for Twitter that are preceded by the hashtag, which stands for the primary idea of material related to particular subjects. In order to obtain the Reddit data, first narrowed down the subreddits that best suited the search terms. These platforms provide application programming interfaces (APIs) like PRAW and Tweepy, which access the data.

B. Pre-processing with Min-Max Normalization

The transformed records have been subjected to facts standardization so that it will lighten the network's computational load. The role coordinates, x, y, and z are normalized by the use of a Min-Max method to the variety [0, 1]. The capability of the community to converge is advanced with the aid of the usage of Max-Min normalization and gaining knowledge of bounded objectives. The basic data preparation technique of min-max normalization guarantees that the numerical characteristics or parameters are adjusted to a selected variety, often among 0 and 1. In order to improve the suitability for analysis and goal detection algorithms, raw record values ought to be standardised in this system. By aligning the facts into the algorithms' preferred input variety, min-max normalization will increase the efficiency and precision of the techniques. By shifting these outliers in the direction of the top or lower boundaries of the normalized variety, Min-Max normalization may additionally help in highlighting them and lead them to be less complicated to distinguish from regular visitor's styles. The initial statistics set is transformed linearly with the aid of the Min-Max normalization technique [20]. When some characteristic's minimum and maximum values are normalized using the Min-Max formula, the initially set value of the attribute gets replaced with the value within the interval [0,1]. The formula is given in Eq. (1):

$$X'' = \frac{X-Min}{Max-Min} \tag{1}$$

Where Min and Max be the minimum and maximum values of typical X'', accordingly, the initial value of X is changed by Min-Max normalization to the value in the range [0,1].

C. Attention-Based LSTM (AT-LSTM) for Detection

The crucial component for aspect-level categorization of emotions cannot be identified by the conventional LSTM. To tackle this problem, AT-LSTM [21] is suggested creating a system of attention offering can determine the essential portion of a phrase when it reacts with a certain feature.

Considering d is the size of the hidden layers and N is an expression length, let $H\epsilon\mathbb{R}_{d\times N}$ be a structure made up of the hidden vectors $h_1, h_2, \dots, h_N$ that the LSTM generated. Moreover, $e_N\epsilon\mathbb{R}_N$ is a vector of 1s, and v_α denotes the aspect representation. A weighted hidden representational r and a concentrated weight vector (α) will be generated by the mechanism that regulates attention in Eq. (2).

$$M = \tan h\left[\begin{pmatrix} W^h H \\ W^v v_\alpha \otimes e_N \end{pmatrix}\right] \tag{2}$$

Where, $M\epsilon\mathbb{R}_{(d+d_\alpha)\times N}, \alpha\epsilon\mathbb{R}_N, r\epsilon\mathbb{R}_d, W_h\epsilon\mathbb{R}_{d\times d}, W_v\epsilon\mathbb{R}_{d_{\alpha\times d_\alpha}}$ and $w\epsilon\mathbb{R}_{d\times d_\alpha}$ are projection parameters. α is a vector made up of attention weights, and r is a weighted sentence representation with a specified aspect. The operator in 7 (a circle containing a multiplication symbol within, abbreviated as OP) denotes the following: $v_{\alpha\otimes e_N} = v; v; \dots; v$.

The phrase "The operator in 7 (a circle containing a multiplication symbol within, abbreviated as OP)..." refers to a specific mathematical or symbolic operation mentioned earlier in the document or paper, likely in Eq. (7). The description suggests that the "circle containing a multiplication symbol" is a shorthand for an operation involving element-wise multiplication or concatenation of vectors, which is a common operation in attention mechanisms within neural networks.

In this case, it appears that the operator (⨂) is performing a concatenation or element-wise operation on vectors, specifically involving attention weights and sentence representations, to enhance the LSTM model's ability to focus on important features. The specific operator and its notation would be detailed in Eq. (7) of the original document or study. If you are following the cited article or context, ensure you check for the exact meaning of "operator 7" in the associated equations or figures.

This indicates that the operator concatenates v for N times in a row vector called e_N, which has N, 1s in it. The gradually adjusted v_α is repeated as repeatedly as the number of words in the phrase by$W_v v_\alpha \otimes e_N$. The depiction of the last statement is provided by Eq. (3)

$$h'' = \tan h\left(W_p r + W_x h_N\right) \tag{3}$$

Where, W_p, W_x and $h\epsilon\mathbb{R}_d$ are projection parameters that will be discovered during training. Also discover that if includes $W_x h_N$ in the sentence's final form, this practically functions better. The attention mechanism enables the framework, when various elements are taken into account, to identify the most significant portion of a phrase. Considering an input component, h'' is regarded as the symbolic representation of features of a sentence. The Attention-based LSTM (AT-LSTM) design is shown in Fig. 2.

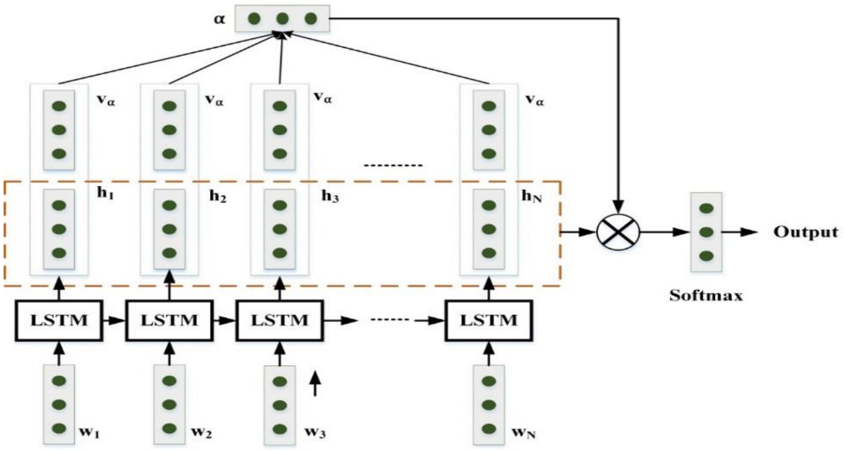


Fig. 2. Design of Attention-based LSTM.

To transform the sentence vector to e, a realvalued vector whose length equals the class number $|C|$, add an exponential layer. After that, e is transformed to a conditional probability distribution using a softmax layer in Eq. (4).

$$y = softmax\ (W_s h'') + b_s \tag{4}$$

where, W_s and b_s are the parameters for softmax layer.

Understanding the steps to solving the problem calls for breaking down the complex system into clear, manageable levels. The objective of the have a look at is to are expecting students' mental health repute using statistics accrued from social media structures and processed through a sophisticated neural community model, the Attention-based LSTM (AT-LSTM). This approach is designed to enhance the prediction of intellectual fitness by focusing at the most applicable statistics points inside massive datasets. By integrating attention mechanisms into LSTM, the version can higher pick out key patterns inside the statistics, leading to greater correct predictions.

The first step involves amassing relevant records, which is critical for any system learning model. In this example, the statistics are sourced from social media platforms like Reddit and Twitter the usage of APIs such as PRAW and Tweepy. The records gathering technique specializes in amassing content material that displays emotional and mental fitness states, as captured in consumer posts and interactions. Once the statistics is collected, it undergoes pre-processing via Min-Max normalization. This normalization guarantees that the information is standardized, making it more green for the model to manner and improving the accuracy of predictions.

The next step involves feeding the processed data into the AT-LSTM model. This model uses an attention mechanism to focus on the most important parts of the data. This has led to more accurate predictions of students' mental health. The attention mechanism within the LSTM framework helps highlight important features within the dataset to ensure that the model does not overlook important emotional signals or patterns based on and predicts—the approach is more structured. It allows readers to easily follow and understand how each step contributes to solving the problem of mental health prediction.

V. Results and Discussion

The application of min-max normalisation in conjunction with attention-based LSTM processes greatly increased the prediction accuracy of students' mental health states. More accurate and context-aware predictions were made possible by the model's improved capacity to extract and analyse pertinent information from the data. This might result in early support and assistance for the well-being of learners in educational settings that are more successful.

A. *Performance Evaluation*

For assessment, the subsequent evaluation standards have been used: don't forget, F1-score, precision and accuracy. These parameters have been used to assess the version. These are depicted below:

The prediction accuracy shown in Eq. (4) that is maximum frequently hired to evaluate category performances is 2nd hand to degree of the classifier's popular usefulness.

$$Accuracy = \frac{Tp' + Tn'}{Tp' + Tn' + Fp' + Fn'} \tag{5}$$

The time period precision is used to describe how well a set of effects accept as true with one another. Precision is normally described because the distinction among a set of outcomes and the set's mathematics suggest. It is shown in Eq. (6).

$$Precision = \frac{Tp'}{Tp' + Fn'} \tag{6}$$

The cause of remember evaluation shown in Eq. (7) is to envision, beneath a positive set of assumptions, how several morals of an independent alterable impact a specific reliant on bendy. This technique is applied inside prearranged bounds which might be dependent on unmarried or additional input facts variables.

$$Recall = \frac{Tp'}{Tp' + Fn'} \tag{7}$$

Outcomes extra than estimate precision had better also be assessed whilst assessing the performance. The F1 rating that is computed for this purpose evaluates the correlation most of the information's expectant statistics and the classifier's predictions. It is shown in Eq. (8).

$$F1\ score = \frac{2Tp\prime}{2Tp\prime + Fp\prime + Fn\prime} \qquad (8)$$

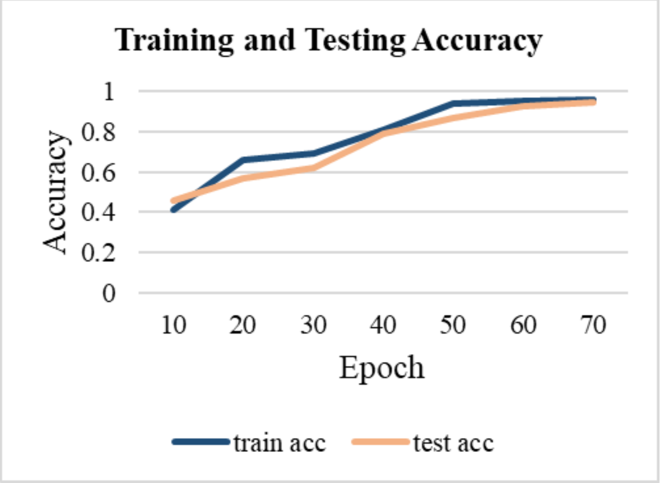


Fig. 3. Training and testing accuracy.

In Fig. 3, the training and testing accuracy graph is depicted. Fig. 4 graphically represents training and testing loss.

Three approaches were assessed in this comparative examination of several techniques for mental health state prediction are listed in Table I and Fig. 5 depicts the comparison of performance metrics. The Heart Rate Variability (HRV), the Hidden Markov Model (HMM), and the suggested Attention-based LSTM (AT-LSTM). The findings showed that the approaches' performance indicators varied significantly from one another.

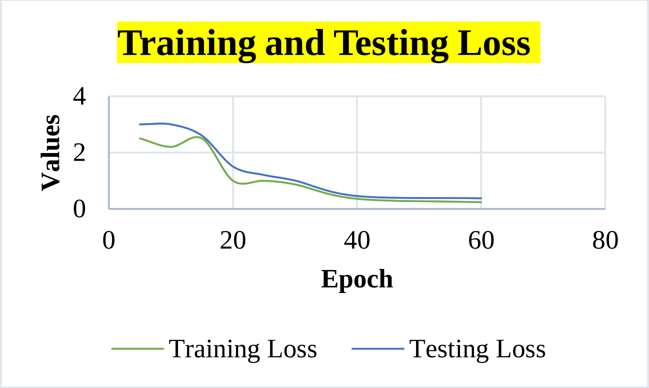


Fig. 4. Training and testing loss.

TABLE I. COMPARISON OF PERFORMANCE METRICS

Method	Accuracy	Precision	Recall	F1-Score
HMM [22]	48	51	57	52
HRV [23]	83	87	82	79
Proposed AT-LSTM	98.9	97	96.5	96

HRV and the counseled AT-LSTM both beat the HMM technique in terms of precision, recall, and F1-rating, while the HMM technique yielded the bottom accuracy (forty eight%) average. The effects showed that HRV had the pleasant accuracy (83%) and the maximum balanced precision, remember, and F1-rating values. The cautioned AT-LSTM, however, became inside the maximum noteworthy performance, reaching a fantastic ninety eight.Nine% accuracy in addition to excessive precision, take into account, and F1-score values, demonstrating its higher predictive abilties for intellectual fitness situation assessment.

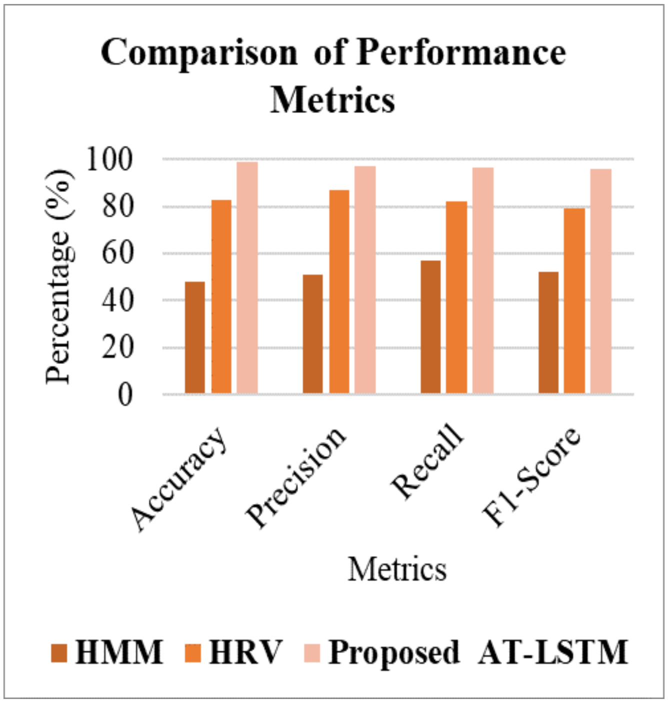


Fig. 5. Comparison of performance metrics.

Table II and Fig. 6 depict the Attention-based LSTM (AT-LSTM) model's performance measures confirmed amazing values for bear in mind, accuracy, precision, and F1 score. It confirmed quite high percentage of correct predictions, with 98.Nine% accuracy. With a precision of 97%, the model validated a low false tremendous price, as measured by the share of proper superb predictions amongst all superb predictions. With a combined precision keep in mind rating of 96. 5%, the F1 score demonstrates a great balance between the 2. Furthermore, they consider, that the percentage of real positives found amongst all actual positives, turned into a robust 96%, indicating that the version should effectively discover and classify fine instances. These first-rate performance metrics demonstrate the AT-LSTM's capacity to accurately and reliably expect and examine mental health situations.

TABLE II. PERFORMANCE METRICS OF AT-LSTM

Performance metrics	AT-LSTM
Accuracy	98.9
Precision	97
F1 score	96.5
Recall	96

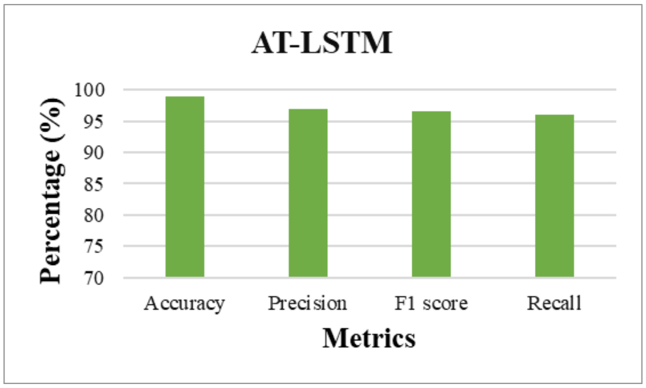


Fig. 6. Performance Evaluation of Hybrid ELM-BiLSTM.

B. Discussion

The proposed take a look at introduces an innovative method using attention-based totally LSTM methods to beautify predictions of scholar mental health. This generation achieves a high diploma of accuracy in figuring out subtle signs and symptoms of intellectual health issues by meticulously studying substantial elements inside the input statistics. This development holds significant potential to enhance early identification and remedy processes, in the end reaping benefits for the overall properly-being of students [16]. By integrating interest-primarily based LSTM procedures, the method addresses the shortcomings of present systems and efficaciously overcomes fundamental barriers in forecasting pupil intellectual health. Unlike preceding strategies, this method goals to beautify the version's capacity to apprehend subtle indicators of intellectual fitness difficulties with the aid of assigning varying stages of relevance to one-of-a-kind enter factors. Compared to current methods, which may struggle with early detection and accuracy, the proposed approach offers a extra precise and context-aware prediction manner. This allows timely assistance and interventions for college kids. However, challenges together with integrating a couple of facts sources for in-intensity evaluation and optimizing the eye technique remain. Future studies could discover similarly upgrades to the attention-based totally LSTM framework, together with refining interest mechanisms to better discover behavioral nuances in students and incorporating extra facts assets for a greater complete evaluation. Additionally, that specialize in developing user-friendly interfaces and deployment techniques could make the proposed approach more widely on hand and relevant in educational institutions.

VI. Conclusion and Future Work

This proposed approach allows greater correct and context-aware forecasts of intellectual fitness amongst students, potentially leading to more powerful early interventions and help strategies in educational environments. Utilizing min-max normalization along side attention-based totally LSTM strategies drastically improves the accuracy of predicting college students' intellectual fitness states. By correctly extracting and reading applicable facts, the version demonstrates more advantageous predictive competencies, supplying the ability for more a hit interventions and support for pupil nicely-being. Performance assessment metrics such as recollect, F1-rating, precision, and accuracy are hired to assess the version's effectiveness. Comparative evaluation with different techniques famous the superiority of the proposed approach in terms of predictive skills for comparing intellectual fitness conditions. In end, this studies highlights the promising capacity of interest-based LSTM fashions in predicting college students' intellectual fitness statuses with high accuracy and precision. By effectively leveraging interest mechanisms and using superior neural community architectures, the proposed methodology offers a valuable tool for educators and mental health professionals to identify and cope with intellectual health concerns amongst students early on. Future studies should discover the application of this technique in real-international educational settings and in addition look into its efficacy in distinctive scholar populations. Additionally, efforts to beautify the interpretability and explainability of the model's predictions ought to contribute to its broader adoption and effect in supporting pupil properly-being.

References

[1] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang, S.-W. Kim, and M. Jhon, ‘Additive Cross-Modal Attention Network (ACMA) for Depression Detection based on Audio and Textual Features’, IEEE Access, 2024.

[2] D. Roy, S. Tripathy, S. K. Kar, N. Sharma, S. K. Verma, and V. Kaushal, ‘Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic’, Asian journal of psychiatry, vol. 51, p. 102083, 2020.

[3] B. M. Mayosi, A. J. Flisher, U. G. Lalloo, F. Sitas, S. M. Tollman, and D. Bradshaw, ‘The burden of non-communicable diseases in South Africa’, The lancet, vol. 374, no. 9693, pp. 934–947, 2009.

[4] M. O. Adebiyi, T. T. Adeliyi, D. Olaniyan, and J. Olaniyan, ‘Advancements in accurate speech emotion recognition through the integration of CNN-AM model’, ᵀᴱᴸKOMNIKA (Telecommunication Computing Electronics and Control), vol. 22, no. 3, pp. 606–618, 2024.

[5] I. M. Tudorancea et al., ‘The Therapeutic Potential of the Endocannabinoid System in Age-Related Diseases’, Biomedicines, vol. 10, no. 10, p. 2492, 2022.

[6] S. A. V. Kolaei, ‘KIANNET: AN ATTENTION-BASED CNN-RNN MODEL FOR VIOLENCE DETECTION’, PhD Thesis, University of Regina, 2024.

[7] T. Thaipisutikul, P. Vitoochuleechoti, P. Thaipisutikul, and S. Tuarob, ‘MONDEP: A Unified SpatioTemporal MONitoring Framework for National DEPression Forecasting’, Heliyon, 2024.

[8] G. Mustafa, Y. Hwang, and S.-J. Cho, ‘Predicting Bus Travel Time in Cheonan City Through Deep Learning Utilizing Digital Tachograph Data’, Electronics, vol. 13, no. 9, p. 1771, 2024.

[9] J. Yoo, ‘Stress release kit for college students through art therapy & activity’, 2018.

[10] H. Yasmin, S. Khalil, and R. Mazhar, ‘COVID 19: Stress management among students and its impact on their effective learning’, International technology and education journal, vol. 4, no. 2, pp. 65–74, 2020.

[11] U. Ravens-Sieberer, A. Kaman, M. Erhart, J. Devine, R. Schlack, and C. Otto, ‘Impact of the COVID-19 pandemic on quality of life and mental health in children and adolescents in Germany’, European child & adolescent psychiatry, vol. 31, no. 6, pp. 879–889, 2022.

[12] J. Maben et al., ‘“You can’t walk through water without getting wet’UK nurses” distress and psychological health needs during the Covid-19 pandemic: A longitudinal interview study’, International journal of nursing studies, vol. 131, p. 104242, 2022.

[13] H. J. Swift, D. Abrams, R. A. Lamont, and L. Drury, ‘The risks of ageism model: How ageism and negative attitudes toward age can be a barrier to active aging’, Social Issues and Policy Review, vol. 11, no. 1, pp. 195–231, 2017.

[14] Y. Acikmese and S. E. Alptekin, ‘Prediction of stress levels with LSTM and passive mobile sensors’, Procedia Computer Science, vol. 159, pp. 658–667, 2019.

[15] N. S. M. Shafiee and S. Mutalib, ‘Prediction of mental health problems among higher education student using machine learning’, International Journal of Education and Management Engineering (IJEME), vol. 10, no. 6, pp. 1–9, 2020.

[16] G. Mikelsons, M. Smith, A. Mehrotra, and M. Musolesi, ‘Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities’, arXiv preprint arXiv:1711.06350, 2017.

[17] L. V. Coutts, D. Plans, A. W. Brown, and J. Collomosse, ‘Deep learning with wearable based heart rate variability for prediction of mental and general health’, Journal of Biomedical Informatics, vol. 112, p. 103610, Dec. 2020, doi: 10.1016/j.jbi.2020.103610.

[18] E. Sükei, A. Norbury, M. M. Perez-Rodriguez, P. M. Olmos, and A. Artés, ‘Predicting emotional states using behavioral markers derived from passively sensed data: data-driven machine learning approach’, JMIR mHealth and uHealth, vol. 9, no. 3, p. e24465, 2021.

[19] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung, ‘A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model’, Computational Intelligence and Neuroscience, vol. 2022, p. e7893775, Mar. 2022, doi: 10.1155/2022/7893775.

[20] W. Cui, Q. Lu, A. M. Qureshi, W. Li, and K. Wu, ‘An adaptive LeNet-5 model for anomaly detection’, Information Security Journal: A Global Perspective, vol. 30, no. 1, pp. 19–29, Jan. 2021, doi: 10.1080/19393555.2020.1797248.

[21] Y. Wang, M. Huang, X. Zhu, and L. Zhao, ‘Attention-based LSTM for aspect-level sentiment classification’, in Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 606–615.

[22] E. Sükei, A. Norbury, M. M. Perez-Rodriguez, P. M. Olmos, and A. Artés, ‘Predicting Emotional States Using Behavioral Markers Derived From Passively Sensed Data: Data-Driven Machine Learning Approach’, JMIR mHealth and uHealth, vol. 9, no. 3, p. e24465, Mar. 2021, doi: 10.2196/24465.

[23] L. V. Coutts, D. Plans, A. W. Brown, and J. Collomosse, ‘Deep learning with wearable based heart rate variability for prediction of mental and general health’, Journal of Biomedical Informatics, vol. 112, p. 103610, 2020.

A Hybrid Intelligent System for IP Traffic Classification

Muhana Magboul Ali Muslam, Senior, IEEE
Department of Information Technology-College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University, P.O. Box 5701, Riyadh 11432, Saudi Arabia

***Abstract*—The classification of IP traffic is important for many reasons, including network management and security, quality of service (QoS) monitoring and provisioning, and high hardware utilisation. Recently, many machine learning-based IP traffic classifiers have been developed. Unfortunately, most of them need to be trained on large datasets and thus require a long training time and significant computational power. In this paper, I investigate this problem and, as a solution, present a hybrid system, which I call the ISITC, that combines the random forest (RF) and XGBoost (XGB) machine learning techniques with the support vector classifier (SVC) as the final estimator, the stacking classifier. This design leads to the development of a model that performs the classification of IP traffic and internet applications efficiently and with high accuracy. I evaluate the performance of the ISITC and various IP traffic classifiers, including neural network (NN), RF, decision tree (DT), and XGB classifiers and SVCs. The experimental results show that the ISITC provides the best IP traffic classification, with an accuracy of 96.7, and outperforms the other IP traffic classifiers: the NN classifier has an accuracy of 59, the RF classifier has an accuracy of 88.5, the DT classifier has an accuracy of 90.5, the XGB classifier has an accuracy of 89.8, and the SVC has an accuracy of 64.8.**

Keywords*—Internet application classification; IP traffic classification; machine learning; machine learning techniques; stacking classifier

I. Introduction

IP traffic classification is crucial to network management and security [1], quality of service (QoS) monitoring and provisioning [2][3], and better hardware utilisation [4]. However, the emergence of encryption and encapsulation [5] is making this a difficult task. Traditional methods such as port- and payload-based identification and deep packet inspection (DPI) are becoming increasingly ineffective due to dynamic port numbers and encryption [6].

Machine learning (ML) techniques, especially decision tree, C4.5, and random forest algorithms, have shown promise in this area [2][6]. These techniques can be used to develop real-time classification systems, with the Bayesian network being particularly effective [7]. However, these machine learning-based IP traffic classifiers need to be trained on a large dataset in order to be able to perform classification with high accuracy. Training on a large dataset is time consuming; it is not always possible to prepare large datasets and use them for the on-flight training of classifiers, and more computational resources are required. Moreover, many machine learning models do not achieve high accuracy when trained on small datasets [8]. Therefore, solutions to this problem are needed. The limitations of small datasets in achieving machine learning models with high accuracy may be due to the need for better methods [8] and the promotion of a data-centric approach to improving model performance [9]. In this research, I attempt to answer the question of how combined machine learning algorithms can be used to improve the accuracy of IP traffic classifiers with small datasets (containing only the most frequent features).

To respond to this challenge, in this paper, I investigate how to develop an effective IP traffic classifier that can be trained on a small dataset. The main objectives of this research are (1) to analyse and evaluate the performance of neural network, random forest, decision tree, XGBoost, and support vector classifiers for IP traffic classification, (2) to develop a hybrid traffic classification system that can be trained on small datasets and used to classify IP traffic with minimum latency and high accuracy, and (3) to compare the performance of individual and hybrid IP traffic classifiers in terms of accuracy. I conclude that an intelligent hybrid system (combining different machine learning methods) can efficiently and effectively classify IP traffic when trained on a small dataset, as combining the strengths of different machine learning models can increase the ability to capture different patterns in datasets that individual classifiers might miss. The proposed system combines the random forest (RF) technique and the XGBoost (XGB) technique with the support vector classifier (SVC) in a way that maximises the possibility of achieving high performance in the IP traffic classifier using a small number of data. The proposed solution is called the Intelligent System for IP Traffic Classification (the ISITC). Efficient IP traffic classifiers such as the ISITC can result in the prioritisation of bandwidth for critical services, the improvement of network performance, a reduction in the need for expensive and computationally intensive manual traffic monitoring tools, and the possibility of faster monitoring, which can lead to anomaly detection and, thus, improve security.

The main contributions in this paper are: (1) the investigation of different machine learning models used for IP traffic classification, (2) the introduction of an IP traffic classifier that can be easily implemented in networks without requiring high-performance computing and (3) depends on a small dataset with few and general features to classify IP traffic, and (4) a comparative analysis of widely used machine learning models in the field of IP traffic classification.

The remainder of the paper is organised as follows: Section II gives an overview of IP traffic classifiers based on individual machine learning models and IP traffic classifiers based on an ensemble. Section III presents the proposed solution, the intelligent system for IP traffic classification (the ISITC).

Section IV presents the results for the IP traffic classifiers, while Section V discusses the performance evaluation of the IP traffic classifiers. Section VI concludes this paper.

II. Review of IP Traffic Classifiers Based on Individual Machine Learning Models and IP Traffic Classifiers Based on an Ensemble

In this section, I examine IP traffic classifiers based on individual machine learning models and IP traffic classifiers based on an ensemble. Many machine learning algorithms that have been used to classify IP traffic have achieved varying degrees of accuracy. For example, in a previous study, the Bayesian network and C4.5 achieved 94% accuracy, but this dropped to 88% for smaller datasets [10]. Furthermore, [2] showed that the size of the dataset significantly influences the classification performance.

The random forest (RF) classifier is often used as an example of a single classifier [1][11][12]. In [1], the authors used random forest (RF), decision tree (DT), support vector machine (SVM), K-nearest neighbour (KNN), and naive Bayes (NB) classifiers to classify IP traffic. The RF classifier achieved the best accuracy, at around 87%. Random forest (RF) and convolutional neural network (CNN) classifier are used to classify the most common applications [11], and the models (RF and CNN) in this work were trained with datasets comprising more than 2 million samples. In [12], the authors focused on using random forest to study application-based traffic classification in an enterprise network. They collected traffic data in an enterprise network using OpenFlow in SDN. Then, the proposed classifiers were used to classify traffic flows in eight applications, namely: YouTube, Vimeo, Facebook, LinkedIn, Skype, BitTorrent, Web browsing (HTTP), and Dropbox. However, the proposed method is limited by the data provided by OpenFlow.

In [13], the authors combined DPI and machine learning to classify network traffic. They first identified the traffic as far as possible using the DPI module and then used the machine learning module to identify the unidentified traffic. Although the classification accuracy of this model was more than 98%, the privacy of the traffic successfully identified via DPI was compromised and there was an additional delay in classifying the unidentified traffic using DPI.

Decision trees (DTs) were used in [14], [15], and [16]. In [14], the authors determined whether the traffic flow was an elephant flow or a mouse flow. In [15], a DT was used to detect traffic among the top 40 applications in the Google Play Store. In [16], the authors used both decision tree and k-NN classifiers. They used two different datasets: one to classify the IP traffic among the top 37 apps in the Google Play Store and the other to classify the IP traffic among 45 apps.

An SVM was used in [17] to classify IP traffic to one of eight applications (PPlive, TVAnts, SopCast, Joost, Edonkey, BitTorrent, Skype, and DNS) based on Netflow records. In [18], a deep neural network (DNN) was used to classify IP traffic to 1 of 200 mobile applications.

Due to the advantages of ensemble methods, which have shown promising results in internet traffic classification, many such methods have been developed [19][20][21][22]. In one study [19], an ensemble of SVMs with different kernels, extra-tree-based feature selection, and majority voting was presented and achieved better results than single-kernel methods.

Moreover, in another study [20], ensemble learning was combined with co-training techniques to address weak adaptability, limited accuracy of data flow, and the need for large labelled training sets. Xu et al. [21] presented an ensemble method using three neural networks as the base model and weight tuning, achieving an accuracy of 96.38% for the payload of the transportation layer of packets. In [22], an ensemble classifier for IP traffic was proposed for imbalanced but not small datasets.

Although there are many methods of IP traffic classification, they need to be trained on large datasets, so there is still a need for efficient machine learning methods that can be trained on small datasets and achieve a good performance. Therefore, this study investigates this problem and proposes a technique that combines different techniques simply and efficiently to provide a solution that takes into account important factors such as the time required to train the model and the size of the training datasets (in terms of samples and features) while maintaining good performance.

III. An Intelligent System for IP Traffic Classification

This section provides an overview of the proposed solution, an intelligent IP network traffic classification system called the ISITC, and how it works.

A. An Overview of the ISITC

The ISITC is a hybrid intelligent system for IP network traffic classification that uses a combination of XGBoost (XGB) and a random forest (RF) with a support vector classifier (SVC) as the final estimator to efficiently classify network traffic into different application classes.

B. How does the ISITC Work?

During the training of the ISITC, the training data (80% of the total dataset) are divided into three folds. In each iteration, the XGB and RF are trained on two foldings and proceed to perform classification based on the remaining folding. The classifications performed by the XGB and RF are used as features to train the SVC. For each training example, a new feature set consisting of the classifications of the XGB and RF classifiers is obtained. Then, the SVC is trained on these meta-features. It is important to note that the target labels remain the same but the input features are now the classifications of the XGB and RF.

When the XGB and RF classifiers are trained on the entire training dataset, both perform classifications on the test dataset (20% of the entire dataset). These classifications performed by the XGB and RF classifiers on the test dataset are used to create the test meta-features that are used by the SVC to perform the final classifications on the test dataset. Fig. 1 shows the ISITC elements and their interactions during the training and testing processes.

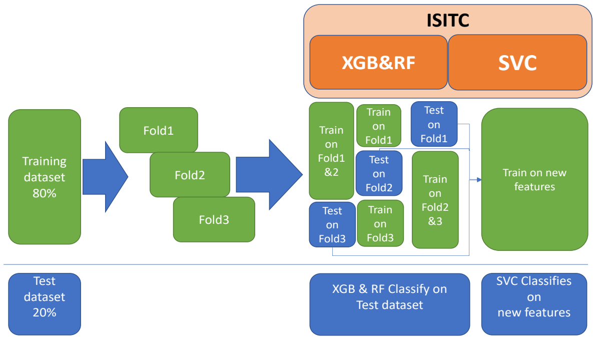


Fig. 1. The architecture of the Intelligent System for IP Traffic Classification (the ISITC).

To ensure that both the hyperparameters of the XGB and RF classifiers and the stacking ensemble are optimised, I use stacking with cross-validation (cv=3) and GridSearchCV in this solution, the ISITC. This method utilises the advantages of grid search to tune the hyperparameters and stacking for the composition of the XGB and RF classifiers. Fig. 2 shows the procedures for loading and splitting the dataset, determining the classifier parameters, the training and testing process, and the stacking process with 3-fold cross-validation for the XGB and random forest classifiers with SVC as the last estimator.

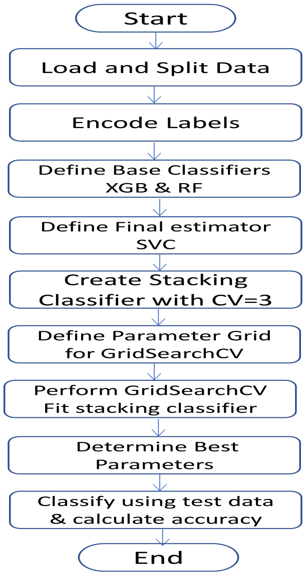

Start
Load and Split Data
Encode Labels
Define Base Classifiers XGB & RF
Define Final estimator SVC
Create Stacking Classifier with CV=3
Define Parameter Grid for GridSearchCV
Perform GridSearchCV Fit stacking classifier
Determine Best Parameters
Classify using test data & calculate accuracy
End

Fig. 2. Training and testing procedures for the Intelligent System for IP Traffic Classification (the ISITC).

The figure above shows the key steps in stacking classifiers with 3-fold cross-validation and optimising their performance using grid search.

IV. Results of the IP Traffic Classifiers

This section presents the dataset and the experimental setup, as well as the results for the IP traffic classifiers and the statistical analysis.

A. Dataset and Experimental Setup

The dataset used in this paper is part of the “IP Network Traffic Flows Labeled with 75 Apps” dataset [23]. The small dataset used comprises only 2172 samples and five applications. Table I shows the characteristics of the datasets and the experimental setup.

TABLE I. Dataset and Experimental Setup Under Which IP Traffic Classifiers are Examined

Parameter	***Value***
Split ratio	80% training, 20% testing
# of samples	2172
# of classes	5
# of instances in class 0	565
# of instances in class 1	439
# of instances in class 2	418
# of instances in class 3	405
# of instances in class 4	345
Data scaling	StandardScaler
Combination Method	Stacking
Parameter tuning	GridSearchCV, 3 folds
Validation	cross-validation using cross_val_score
Statistical tests	t-test using ttest_ind

B. The Results of the Different IP Traffic Classifiers

In this section, the accuracy results for different IP traffic classifiers using different machine learning models are presented. In particular, the accuracy results for the neural network, random forest, decision tree, and XGBoost classifiers, the SVC, and the ISITC are presented. The experimental results show that the NN classifier has an accuracy of 59, the RF classifier has an accuracy of 88.5, the DT classifier has an accuracy of 90.5, the XGB classifier has an accuracy of 89.8, the SVC has an accuracy of 64.8, and the ISITC has an accuracy of 96.7. The accuracy of each classifier is shown in Fig. 3.

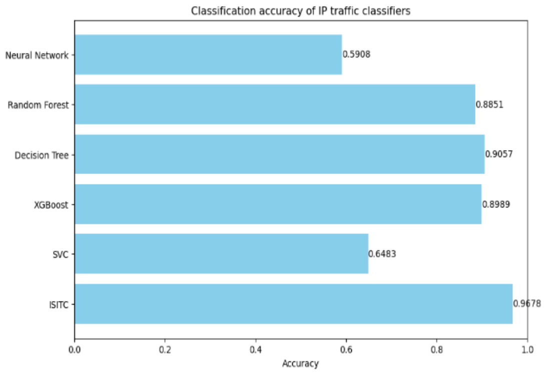


Fig. 3. Classification accuracy of IP traffic classifiers.

To determine the optimal settings for each classifier when using a small dataset, I also evaluated the performance of these classifiers, including neural network (NN), random forest (RF), decision tree (DT), and XGBoost (XGB) classifiers, an SVC, and the ISITC (a stacked model using the RF and XGB, with the SVC as the final estimator). The accuracy results for each IP traffic classifier with different hyperparameters are shown in Fig. 4.

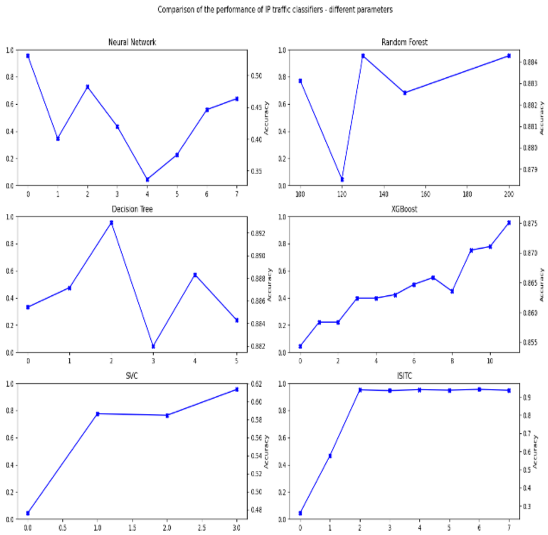


Fig. 4. Comparison of the performance of IP traffic classifiers—different parameters.

I found that the IP traffic classifier using the neural network achieved an accuracy of 0.5195%, with the best configuration comprising hidden layers of size 150 and a learning rate of 0.01. The IP traffic classifier using the random forest model achieved an accuracy of 0.8897% with 100 estimators. The IP traffic classifier using the decision tree model showed an accuracy of 0.9103% at a maximum depth of 30.

The IP traffic classifier using the XGBoost model achieved an accuracy of 0.9609% with 120 estimators and a learning rate of 0.1. The IP traffic classifier using the SVC model showed an accuracy of 0.6483% with an "rbf" kernel. The IP traffic classifier using the ISITC, the proposed solution, achieved an accuracy of 0.9678% with an "rbf" kernel.

To further analyse the performance of the classifiers, a confusion matrix (CM) was created for each IP traffic classifier, providing information on the number of correct and incorrect classifications for each class, as shown in Fig. 5, below.

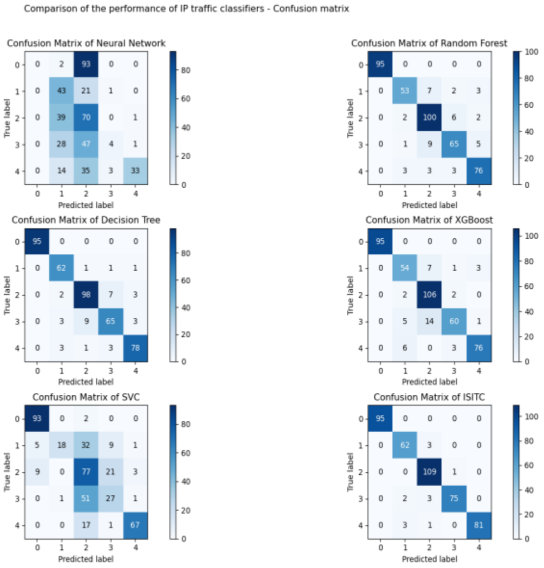


Fig. 5. Confusion matrix for the IP traffic classifiers.

C. Statistical Analysis

To assess the statistical significance of the observed differences in the performance of the IP traffic classifiers, t-tests were performed between the results of the cross-validation of the IP traffic classifiers. The t-test results are shown in Table II.

TABLE II. T-TEST RESULTS BETWEEN IP TRAFFIC CLASSIFIERS

Classifier 1	Classifier 2	t-statistic	p-value
ISITC	Neural Network	7.675364069946557	0.001549315592527276
ISITC	Random Forest	4.29261462099269	0.012719774710311835
ISITC	Decision Tree	5.2256182997013205	0.006402824672843334
ISITC	XGBoost	5.163652433269795	0.006681160371121216
ISITC	SVC	8.813500048755037	0.0009144801911715298
ISITC	Neural Network	7.675364069946557	0.001549315592527276
ISITC	Random Forest	4.29261462099269	0.012719774710311835
ISITC	Decision Tree	5.2256182997013205	0.006402824672843334

V. DISCUSSION ON THE PERFORMANCE OF THE IP TRAFFIC CLASSIFIERS

In this section, the performance evaluation of IP traffic classifiers is discussed. From the accuracy results mentioned above (Fig. 3), the evaluation of the different IP traffic classifiers shows that the ISITC (a stacked model of RF and XGB, with SVC as the final estimator) provides promising results and significantly outperforms all the individual classifiers, including the standalone XGB model, which has the higher accuracy among the individual classifiers.

When evaluating the different IP traffic classifiers with the different hyperparameters (Fig. 4), I found that the performances of the NN, RF, DT and XGBoost classifiers were very sensitive to the learning rate, with the NN accuracy decreasing significantly at higher learning rates. The SVC and ISITC classifiers showed consistent performance at different learning rates and different numbers of estimators, with the highest accuracy achieved with the ISITC, highlighting the advantage of hybrid classifiers with stacking.

The ISITC classifier outperformed all single classifiers and the hybrid model, confirming that the combination of different models can further improve performance in IP traffic classification. The IP traffic classifier with the ISITC, the proposed solution, achieved an accuracy of 0.9678% with an "rbf" kernel, emphasising the advantage of hybrid classifiers with stacking.

The higher performance of the ISITC can be attributed to its ability to capture various patterns in data that individual classifiers may miss. The random forest classifier captures complex relationships, while XGBoost recognises interactions between features perfectly.

Confusion matrices (Fig. 5) enable further analysis of the performance of each classifier by providing insight into the number of correct and incorrect classifications for each class. The NN classifier only performed well in class 2 classification (contributing to its lower overall accuracy), while class 0 is recognised with high accuracy by all other classifiers (RF, DT, XGBoost, and SVC). The ISITC's confusion matrix shows strong performance with a balanced classification, similar to the standalone XGB model, with the exception of class 1 and class 3 classification (but also a slight increase in false positives).

To ensure the scalability of the ISITC, cross-validation (cv=3) is used to evaluate the ISITC on three different datasets derived from the entire dataset during the training and testing phase, and the performance of the ISITC with different parameters and different datasets is shown in Fig. 4.

The results of this study are consistent with those of previous studies that have emphasised the effectiveness of hybrid methods in classification in general. However, the specific combination of random forest and XGBoost classifiers with SVC as the final estimator has not been extensively studied, making this work a novel contribution.

The t-tests performed to compare the ISITC and other IP traffic classifiers showed that the performance improvements observed were always statistically significant. For example, the t-test between the ISITC and the neural network showed a statistically significant difference in performance with a p-value of 0.15. The t-tests between the ISITC and the other classifiers (the random forest, decision tree, XGBoost, and SVC) also showed statistically significant differences in performance with p-values of 1.2, 0.6, 0.6, and 0.09, respectively. This means that the ISITC is different from the other IP traffic classifiers.

VI. Conclusion

In this paper, I propose a hybrid system called the ISITC, which simply and efficiently combines random forest (RF) and XGBoost (XGB) classifier techniques with an SVC as the final estimator to classify IP traffic with a small dataset. The ISITC classifier presented here can efficiently classify IP traffic with a high accuracy of 96.7%, which holds promise for improving network management, security measures, and quality of service (QoS). The ISITC outperforms IP traffic classifiers with NN, RF, DT, or XGB classifiers or SVCs. The t-test values show that there is a statistically significant difference in the accuracy of the ISITC and the other IP traffic classifiers. These results emphasise the potential of advanced hybrid classifiers to significantly improve the accuracy and reliability of IP traffic classification. Furthermore, the ISITC results confirm that the combination of different models can further improve performance in IP traffic classification. In the future, further research should be conducted on combinations of classifiers and their performance should be tested on different datasets. In addition, future work could explore advanced ensemble techniques and further tuning of hyperparameters to improve the performance of hybrid models.

References

[1] Rahul, A. Gupta, A. Raj and M. Arora, "IP Traffic Classification of 4G Network using Machine Learning Techniques," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 127-132, doi: 10.1109/ICCMC51019.2021.9418397.

[2] Jun, Li & Shunyi, Zhang & Yanqing, Lu & Zailong, Zhang, "Internet Traffic Classification Using Machine Learning," in CHINACOM 239 - 243. 10.1109/2007.4469372.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," In IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008.

[4] J. Gómez, V. H. Riaño and G. Ramirez-Gonzalez, "Traffic Classification in IP Networks Through Machine Learning Techniques in Final Systems," In IEEE Access, vol. 11, pp. 44932-44940, 2023, doi: 10.1109/ACCESS.2023.3272894.

[5] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin and J. Aguilar, "Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey," In IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1988-2014, Secondquarter 2019

[6] Donghong Qin, Jiahai Yang, Jiamian Wang and Bin Zhang, "IP traffic classification based on machine learning," in 2011 IEEE 13th International Conference on Communication Technology, Jinan, China, 2011, pp. 882-886, doi: 10.1109/ICCT.2011.6158005.

[7] Kuldeep Singh, S. Agrawal, B.S. Sohi, "A Near Real-time IP Traffic Classification Using Machine Learning," in International Journal of Intelligent Systems and Applications (IJISA), vol.5, no.3, pp.83-93, 2013. DOI:10.5815/ijisa.2013.03.09

[8] Shaoxuan Zhou, "An Analysis of The Small Sample Datasets Based on Machine Learning," in 2022 6th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2022), October 21-23, 2022, Xiamen, China. ACM, New York, NY, USA,

[9] Bhowmik, Pritom and Arabinda Saha Partha, "A Data-Centric Approach to Improve Machine Learning Model's Performance in Production," International Journal of Engineering and Advanced Technology (2021): n. pag.

[10] Singh, Kuldeep and Sunil Agrawal, "Comparative analysis of five machine learning algorithms for IP traffic classification," in 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (2011): 33-38.

[11] V. K. BP, K. SM and P. LV, "Deep machine learning based Usage Pattern and Application classifier in Network Traffic for Anomaly Detection," in 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), Bangalore, India, 2023, pp. 50-54, doi: 10.1109/ICAECIS58353.2023.10169914.

[12] P. Amaral, J. Dinis, P. Pinto, L. Bernardo, J. Tavares, and H. S. Mamede, "Machine learning in software defined networks: Data collection and traffic classification," in Proc. - Int. Conf. Netw. Protoc. ICNP, vol. 2016-Decem, no. NetworkML, pp. 1-5, 2016, doi: 10.1109/ICNP.2016.7785327.

[13] W. A. Aziz, H. K. Qureshi, A. Iqbal, A. Al-Dulaimi and S. Al-Rubaye, "Towards Accurate Categorization of Network IP Traffic Using Deep Packet Inspection and Machine Learning," in GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia, 2023, pp. 01-06, doi: 10.1109/GLOBECOM54140.2023.10437078.

[14] P. Xiao, W. Qu, H. Qi, Y. Xu, and Z. Li, "An efficient elephant flow detection with cost-sensitive in SDN," in Proc. 2015 1st Int. Conf. Ind. Networks Intell. Syst. INISCom 2015, pp. 24-28, 2015, doi: 10.4108/icst.iniscom.2015.258274.

[15] Zafar Ayyub Qazi, Jeongkeun Lee, Tao Jin, Gowtham Bellala, Manfred Arndt, and Guevara Noubir, "Application-awareness in SDN," in SIGCOMM Comput. Commun. Rev. 43, 4 (October 2013), 487-488. https://doi.org/10.1145/2534169.2491700.

[16] M. Uddin and T. Nadeem. Traffic Vision: A Case for Pushing Software Defined Networks to Wireless Edges. Proc. - 2016 IEEE 13th Int. Conf. Mob. Ad Hoc Sens. Syst. MASS 2016, pp. 37-46, 2017, doi: 10.1109/MASS.2016.016.

[17] D. Rossi and S. Valenti, "Fine-grained traffic classification with Netflow data," in IWCMC 2010 - Proc. 6th Int. Wirel. Commun. Mob. Comput. Conf., pp. 479-483, 2010, doi: 10.1145/1815396.1815507.

[18] I. Paper, "INVITED PAPER Special Section on Communication Quality in Wireless Networks Toward In-Network Deep Machine Learning for

Identifying Mobile Applications and Enabling Application Specific Network Slicing," in transcom, No. 7, pp. 1536-1543, 2018, doi: 10.1587/transcom.2017CQI0002.

[19] Manju, N., Harish, B.S. (2020), "Classification of Internet Traffic Data Using Ensemble Method. In: Das, H., Pattnaik, P., Rautaray, S., Li, KC. (eds) Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing," vol 1119. Springer, Singapore. https://doi.org/10.1007/978-981-15-2414-1_39.

[20] Haitao He, Chunhui Che, Feiteng Ma, Jun Zhang, Xiaonan Luo, "Traffic Classification Using En-semble Learning and Co-training," in 8th WSEAS International Conference on Applied Informatics and Communications (AIC'08), Rhodes, Greece, August 20-22, 2008.

[21] L. Xu, X. Zhou, Y. Ren and Y. Qin, "A Traffic Classification Method Based on Packet Transport Layer Payload by Ensemble Learning," in 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 2019, pp. 1-6].

[22] Santiago Egea Gómez, Belén Carro Martínez, Antonio J. Sánchez-Esguevillas, Luis Hernández Callejo, "Ensemble network traffic classification: Algorithm comparison and novel ensemble scheme proposal. Computer Networks," Volume 127, 2017, Pages 68-80, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2017.07.018.

[23] IP Network Traffic Flows Labeled With 75 Apps, Apr. 2019, [online] Available: https://kaggle.com/jsrojas/ip-network-traffic-flows-labeled-with-87-apps. (accessed on 10 July 2024).

Deep Learning-Driven Localization of Coronary Artery Stenosis Using Combined Electrocardiograms (ECGs) and Photoplethysmograph (PPG) Signal Analysis

Mohd Syazwan Md Yid[1], Rosmina Jaafar[2], Noor Hasmiza Harun[3],
Mohd Zubir Suboh[4], Mohd Shawal Faizal Mohamad[5]

Dept. Electrical, Electronic & Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia[1, 2]
Medical Engineering Technology Section, British Malaysian Institute, Universiti Kuala Lumpur, Gombak, Malaysia[1, 3, 4]
Dept. of Medicine, Hospital Canselor Tuanku Muhriz, Cheras, Kuala Lumpur, Malaysia[5]

***Abstract*—The application of artificial intelligence (AI) to electrocardiograms (ECGs) and photoplethysmograph (PPG) for diagnosing significant coronary artery disease (CAD) is not well established. This study aimed to determine whether the combination of ECG and PPG signals could accurately identify the location of blocked coronary arteries in CAD patients. Simultaneous measurement of ECG and PPG signal data were collected from a Malaysian university hospital, including patients with confirmed significant CAD based on invasive coronary angiography. ECG and PPG datasets were concatenated to form a single dataset, thereby enhancing the information available for the training process. Experimental results demonstrate that the Convolutional Neural Networks (CNN) + Long Short-Term Memory (LSTM) + Attention (ATTN) mechanisms model significantly outperforms standalone CNN and CNN + LSTM models, achieving an accuracy of 98.12% and perfect Area Under the Curve (AUC) scores of 1.00 for the detection of blockages in the left anterior descending (LAD) artery, left circumflex (LCX) artery, and right coronary artery (RCA). The integration of LSTM layers captures temporal dependencies in the sequential data, while the attention mechanism selectively highlights the most relevant signal features. This study demonstrates that AI-enhanced models can effectively analyze simultaneous measurement of standard single-lead ECGs and PPG to predict the location of coronary artery blockages and could be a valuable screening tool for detecting coronary artery obstructions, potentially enabling their use in routine health checks and in identifying patients at high risk for future coronary events.**

***Keywords*—Deep learning; CNN; LSTM; ATTN; simultaneous ECG and PPG; coronary artery disease**

I. Introduction

CAD represents a substantial global burden on cardiovascular health [1]. Its impact extends to long-term mortality and morbidity all around the world. Existing studies have demonstrated that ischemic heart disease contributes to approximately 16% of total mortality [2]. Furthermore, epidemiological surveys underscore the escalating prevalence of CAD on a global scale. Nevertheless, the evaluation and diagnosis of CAD persistently hinge upon conventional clinical symptoms, signs, and relevant comorbidities.

It is crucial to determine the location of myocardial infarction and ischemia, as well as identify the specific coronary artery that is blocked and where the occlusion occurs. This information facilitates the diagnosis of ischemia and infarction and guides treatment decisions. For instance, administering nitroglycerin to relieve ischemic chest pain can lead to hemodynamic collapse in patients with right ventricular ischemia/infarction [3]. Therefore, recognizing ECG signs of right ventricular issues is essential. Clinicians, especially interventional cardiologists, benefit significantly from this knowledge, as it directly impacts the selection of coronary catheters.

A variety of non-invasive diagnostic modalities are at the disposal for the assessment of potential coronary artery obstructions in patients with CAD, encompassing stress ECG, and nuclear medicine imaging [4], [5]. Nonetheless, these methodologies are encumbered by several limitations: they are not readily accessible, necessitate the use of specialized apparatus, are laborious, and entail considerable expense. The performance of these tests are moderately suboptimal, ranging approximately between 75-90%, and the issue of radiation exposure cannot be overlooked. Moreover, stress-induced tests that require physical exertion from the patient may not be viable for those in a debilitated state. Hence, there is an imperative need for the development of an easily attainable, economical, and highly precise test for the prediction of ischemic localization.

ECG is recognized as a non-invasive diagnostic instrument that boasts several merits: it is straightforward to operate, consistent in results, broadly accessible, and cost-efficient relative to other diagnostic methods [6]. The ECG is capable of discerning significant CAD by manifesting particular alterations in the ECG patterns, such as deviations in the ST-segment, inversions of the T-wave, and the emergence of Q-waves [7]. Nonetheless, the precision in interpreting ECG data may be compromised by the presence of other medical conditions, including arrhythmias, cardiomyopathies, and bundle branch blocks.

PPG is a non-invasive method utilized to detect variations in blood volume by employing an infrared light sensor positioned on the skin's surface. Beyond its conventional use in heart rate monitoring and pulse oximetry, PPG has garnered attention for its potential in detecting CAD. One of the key methodologies in PPG analysis for CAD detection is Pulse Wave Analysis (PWA). PWA evaluates the PPG waveform to assess arterial stiffness, a hallmark of CAD. The presence of plaque in the arteries can alter the waveform's shape, specifically causing a delay in the pulse wave transit time [8]. In addition, ECG and PPG signals can be combined to enable the assessment of cardiac conditions through heart rate variability (HRV) analysis [9].

Artificial intelligence (AI), particularly through deep learning CNN, has been deployed in diverse disease models [10], [11], [12]. CNNs excel in learning from voluminous datasets, autonomously identifying salient features from data, whether one-dimensional (e.g., signals) or two-dimensional (e.g., images). Recurrent neural network (RNN) architectures such as LSTM is extensively utilized in domains like natural language processing, and analysis of sequential data. As adjunct classifiers to a CNN framework, they fulfill distinct roles in augmenting classification precision. The LSTM architecture, in particular, is proficient in detecting temporal correlations within sequential data [13]. Within deep learning paradigms, attention mechanisms (ATTN) empower models to concentrate on pertinent segments of input data, sidelining the less critical parts. This technique is invaluable in sequential data tasks, where the significance of context and element interrelations fluctuates. Utilizing ATTN, the most informative attributes of an ECG signal can be unearthed across various network layers [14] aiding in functions like classification or regression. In healthcare diagnostics, signal and image data are pivotal. The AI-augmented ECG (AI ECG) algorithm, leveraging deep learning, deciphers significant patterns in ECG data [15]. The effectiveness of deep learning models, particularly a hybrid CNN-LSTM architecture, in enhancing the accuracy of PPG signal analysis for detecting and delineating waveforms was proven by [16]. A deep neural network (DNN) utilizing a multilayer perceptron architecture, enhanced with regularization and dropout techniques, has been employed to enhance the accuracy and reliability of CAD diagnosis and prognosis using clinical data [17]. Prior research has validated its utility in diagnosing heart diseases. Yet, its potential in pinpointing ischemic localizations is a domain yet to be investigated.

The motivation for the study is centered on the need for a more efficient, accessible, and accurate method for diagnosing coronary artery disease (CAD) by identifying the location of arterial blockages. Current diagnostic tools like stress ECG and nuclear medicine imaging, while useful, have limitations including moderate accuracy (75-90%), high costs, and reliance on specialized equipment. Moreover, they may not be viable for patients with debilitating conditions.

ECGs) and PPGs are non-invasive and cost-effective methods, but their full potential in diagnosing CAD is underexplored. By combining ECG and PPG signals and utilizing advanced AI models (CNN + LSTM + Attention mechanisms), this study aims to improve the accuracy of CAD diagnosis, particularly in predicting the location of coronary artery blockages. This approach could lead to a more accessible and precise screening tool for CAD, reducing the need for invasive methods like coronary angiography.

II. Related Work

In a study by Tao et al. [18], an automatic system was developed to detect and localize ischemic heart disease (IHD) using magnetocardiography (MCG) data and machine learning techniques. The authors employed MCG recordings from 227 patients with diagnosed coronary stenosis and 347 healthy controls, using coronary angiography (CAG) as the gold standard for diagnosis. They extracted 164 features from the MCG signals, divided them into time-domain, frequency-domain, and information theory categories, and tested several machine learning classifiers, including k-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), and XGBoost. The XGBoost classifier was used to localize ischemic regions, achieving an accuracy of 0.74 for LAD, 0.68 for LCX, and 0.65 for RCA.

Huang et al. [19] developed and evaluated a deep learning model using CNN to identify significant CAD from standard 12-lead ECG. The study utilized six pre-trained CNN models (VGG16, ResNet50V2, InceptionV3, InceptionResNetV2, Xception, and DenseNet) for feature extraction, ultimately finding that the InceptionV3 model without a dense layer provided the best performance. The model classified patients into four groups: normal (no CAD), and those with obstructions in the LAD, LCX, and RCA. The dataset included ECGs from 2,303 patients with angiography-proven significant CAD and 1,053 control patients without CAD. The AI model demonstrated a macro-average area under the ROC curve (AUC) of 0.869 for CAD detection, with individual AUCs of 0.885 for LAD, 0.776 for RCA, 0.816 for LCX, and 1.0 for non-CAD (normal) cases.

The paper by Roopa and Harish [20] proposes a novel approach using the Information Fuzzy Network (IFN) to analyze ECG signals for identifying and localizing thrombus in culprit arteries. The method involves preprocessing ECG signals using a Savitzky-Golay filter, followed by feature extraction through the Stockwell Transform for point detection, Nearest-Neighbor Interpolation for time interval measurement, and peak amplitude assessment. The classification process differentiates between ischemic and non-ischemic signals, identifies the culprit artery, and pinpoints the thrombus location. The study used ECG datasets from the MIT Physionet databank, including Long-Term ST, Spontaneous Ventricular Tachyarrhythmia, and T-Wave Alternans Challenge databases, providing a comprehensive range of cases. For the LAD artery, ST elevation in lead V3 greater than in V1 suggests an LAD blockage, with further localization determined by elevations in other leads, such as aVF, L2, and L3 indicating a proximal block to the major septal artery. The LCX is identified by ST elevation in L2, L3, and aVF with L2 greater than L3, while the RCA is identified by ST elevation in L2, L3, and aVF with L3 greater than L2 and a higher V1 elevation than V3. The proposed method achieved a classification accuracy of 92.3%, with 87.5% sensitivity and 100% specificity.

The previous studies on CAD detection and localization, such as those by Tao et al., Huang et al., and Roopa and Harish, primarily focused on single modalities (MCG or ECG) and did not explore the combined potential of ECG and PPG signals. Moreover, while machine learning techniques and conventional CNNs were employed, more advanced AI models like LSTM and attention mechanisms have not been fully investigated, limiting the ability to capture complex temporal and spatial features. Additionally, the precision in localizing specific coronary arteries remains moderate, and none of these studies explored the diagnostic potential of PPG signals, leaving a gap in fully leveraging its capabilities for CAD detection. Lastly, the existing research lacks generalizability across diverse and multimodal datasets, potentially limiting the robustness and accuracy of CAD diagnosis. This study aims to address these gaps by integrating ECG and PPG signals, enhanced by advanced AI models (CNN, LSTM, and attention mechanisms), to achieve more accurate and comprehensive CAD detection and localization.

III. Materials and Methods

A. Study Population

This paper analyzes the dataset of a study conducted on patients with angiography-proven significant CAD. All study participants have given their written consent and the study is approved by the Research Ethics Committee of Universiti Kebangsaan Malaysia (UKMPPI/111/8/JEP-2020-806). These patients underwent elective invasive coronary angiography at the Hospital Chanselor Tuanku Mukhriz (HCTM) Malaysia. The criteria for inclusion were severe stenosis (>70%) based on quantitative coronary angiography assessment. All participants enrolled in the study fell within the age range of 20 to 65 years, as the aim was to specifically target individuals without any prior history of CAD. All patients were monitored by their cardiology physicians in outpatient clinics.

B. Data Collection

The algorithm for patients' simultaneous ECG and PPG data recording is shown in Fig. 1. A cohort comprising 60 patients diagnosed with significant CAD via angiography was assembled, and a comprehensive dataset of 7156 simultaneous single-lead ECGs and PPG was amassed for analysis. Subsequently, based on the findings from the patients' angiography reports, they were categorized into three distinct groups: those exhibiting stenosis in the left anterior descending artery (LAD), left circumflex artery (LCX), and right coronary artery (RCA).

Within this cohort, the LAD group consisted of 27 patients, yielding 3884 simultaneous single beat ECG and PPG records, the LCX group comprised 16 patients, corresponding to 1565 simultaneous single beat ECG and PPG records, and the RCA groups encompassed 17 patients, accounting for 1707 simultaneous single beat ECG and PPG records. Fig. 2 shows samples of simultaneous single beat ECG and PPG signals for each class LAD, LCX, and RCA from the dataset.

The data utilized in this study, obtained from the hospital, consisted of simultaneous ECG and PPG time series data collected from patients diagnosed with CAD. This dataset comprised standard single-lead ECG (lead II) signals generated by the MAX86150EVS ECG/PPG module, characterized by a measurement frequency of 400 Hz and a measurement duration of 10 minutes. Prior to commencing the training process, the dataset underwent filtering and segmentation procedures to get the best quality of single beat signals and to augment the number of samples, thereby introducing subsamples for each original sample. As a result, a total of 7165 samples were obtained, each representing a complete cycle of the simultaneous ECG and PPG signal and comprised of 187 data points. These time series data in terms of their shape and size will be utilized for subsequent training of our deep learning models.

C. Dataset Preparation and Preprocessing

Prior to inputting the data into the deep learning model, the ECG and PPG datasets were concatenated to form a single dataset. This integration aims to provide a more comprehensive set of information during the training process, which is hypothesized to enhance the model's performance. Fig. 3 presents examples of concatenated ECG and PPG signals corresponding to each class, specifically LAD, LCX, and RCA. In our experimental setup, the dataset underwent division into two distinct subsets: a training set and a test set. Specifically, 80% of the dataset was allocated to the training and validation process. Validation is done during training utilizing 20% of the training data. The remaining 20% of the original dataset constituted the test set. It is noteworthy to mention that an imbalance in data distribution among the groups was observed.

Previous research has indicated that such imbalances can introduce biases during model training [19]. Consequently, to mitigate this issue, we adopted a down-sampling approach, and randomly removed data so that all the classes have the same number of samples. Fig. 4 shows a bar chart for class distribution before and after the data balancing process.

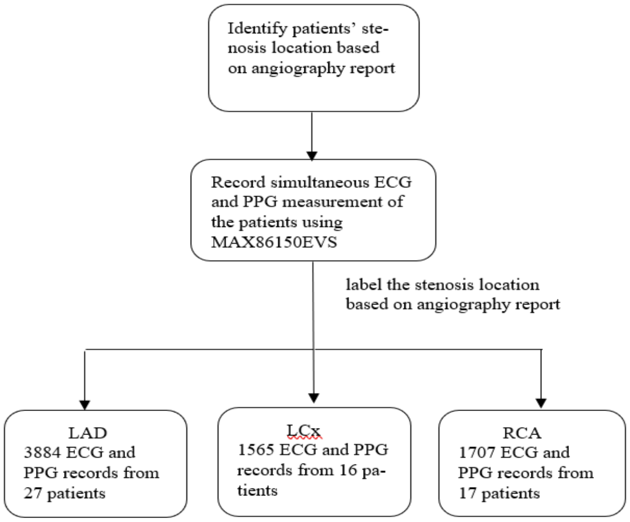

Identify patients' stenosis location based on angiography report
Record simultaneous ECG and PPG measurement of the patients using MAX86150EVS
label the stenosis location based on angiography report
LAD
3884 ECG and PPG records from 27 patients
LCx
1565 ECG and PPG records from 16 patients
RCA
1707 ECG and PPG records from 17 patients

Fig. 1. Data collection protocol.

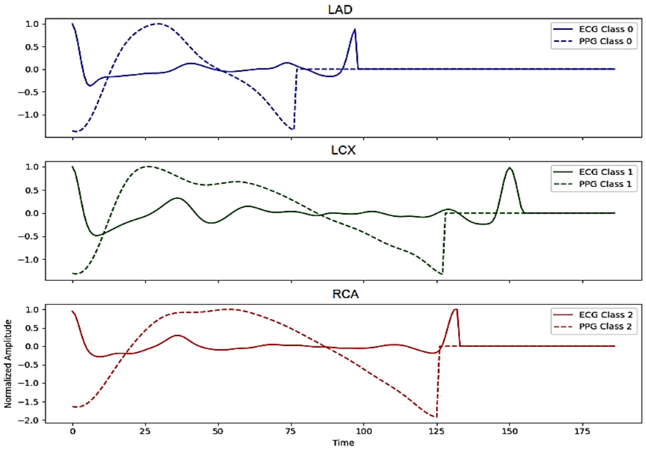


Fig. 2. Normalized simultaneous ECG and PPG measurement samples for patient having blockage at LAD, LCX, RCA.

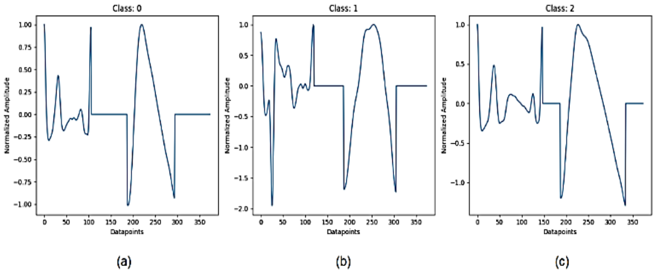


Fig. 3. Concatenated ECG and PPG samples for patient having blockage at (a) LAD, (b) LCX, (c) RCA.

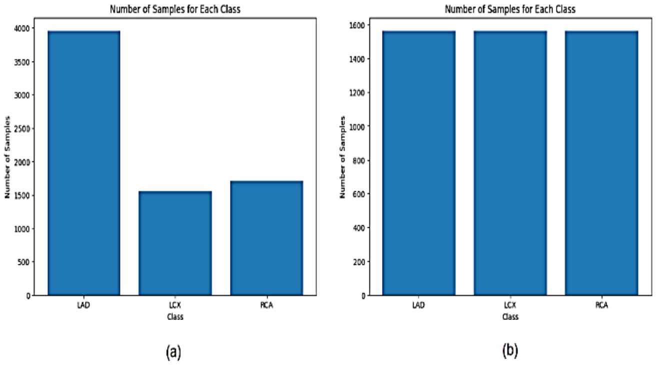


Fig. 4. (a) Imbalanced dataset before data balancing process, (b) Balanced dataset after data balancing process.

D. Model Build-up

In the construction of our deep learning models, we incorporated CNN, LSTM networks, and ATTN along with their respective parameters, resulting in promising outcomes from the integration of all three models. The process of identifying the optimal model involved three key phases.

Initially, we developed a model solely employing 1D CNN. Following the extraction of features by the CNN from the single-lead ECG signals, these features under-went flattening through Max Pooling. Max pooling, a pooling operation technique, extracts the maximum value within each region of the feature map covered by the filter. As a result, the output of the max-pooling layer is a feature map that retains the most prominent features from the preceding feature map. Subsequently, an intermediate dense layer with Rectified Linear Unit (ReLU) activation function was introduced, followed by an additional dense layer with a size of three, representing the three categories of LAD, LCX, and RCA as the output layer. This dense layer employed the Softmax activation function. The performance of this model was then evaluated.

In the second phase, we incorporated an LSTM layer into the existing model and assessed its performance. Finally, in the last phase, we augmented the previous CNN + LSTM model with an ATTN layer and evaluated its performance. The architectural representation of the model is depicted in Fig. 5 after data balancing process.

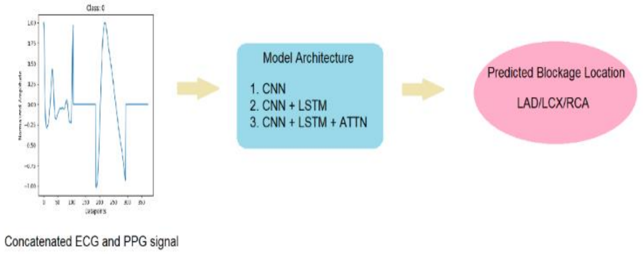


Fig. 5. Architecture for coronary artery blockage localization prediction model.

E. Training Process

The training platform utilized in this study is Google Colaboratory (Colab), which operates within a high-RAM GPU environment. Colab serves as a cloud computing platform supporting Python 3.8 and the TensorFlow package, widely employed for constructing and training deep learning models. As a Google resource, Colab seamlessly integrates with Google Drive, allowing users to access files within Colab by uploading datasets to their personal Google Drives. For model development, we leveraged the Keras application programming interface (API) to construct CNN, LSTM, and ATTN models. Keras not only simplifies the construction of deep learning models but also provides a rich set of APIs and functions, including callbacks, optimizers, metrics, losses, and more, enhancing the versatility and efficiency of model development.

F. Evaluation Metrics

The principal objective of this investigation centered on evaluating the capacity of AI-enhanced ECGs to localize coronary artery blockages utilizing standard single-lead ECG recordings obtained at baseline. The performance of this methodology was evaluated through various assessment metrics, including the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and accuracy. These metrics were computed by averaging the results of five repetitions of training and are reported along with the mean, standard deviation, and 95% confidence interval. The evaluation process also involved the utilization of a confusion matrix, which defined four crucial terms: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These terms were instrumental in computing the aforementioned metrics. Accuracy, represented by Eq. (1), assesses the models' classification proficiency by quantifying the proportion of accurately classified samples out of the total samples. Precision, expressed in Eq. (2), indicates the percentage of correctly predicted positive results among all predicted positive samples. Recall, detailed in Eq. (3), represents the proportion of correctly classified positive samples out of all actual positive samples.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

ROC curve elucidates the fluctuation between the true positive rate (TPR), often termed sensitivity, and the false positive rate (FPR), also known as 1-specificity, over a spectrum of decision thresholds. By modulating the threshold from 0 to 1, a sequence of TPR and FPR coordinates is generated. The ROC curve, plotted with FPR on the X-axis and TPR on the Y-axis, graphically represents the dynamic between specificity and sensitivity in test results. The area under the curve (AUC) signifies the proportion of the area beneath the ROC curve relative to the total possible area. The computation of AUC allows for the ROC curve's quantification, thus enabling the comparative evaluation of model efficacy. The AUC demarcates the model's discriminative capacity into four tiers: (1) AUC < 0.5 (no discrimination), (2) 0.7 ⩽ AUC < 0.8 (acceptable discrimination), (3) 0.8 ⩽ AUC < 0.9 (excellent discrimination), and (4) 0.9 ⩽ AUC ⩽ 1.0 (exceptional discrimination).

ROC-AUC analysis is leveraged in a multitude of fields, including radiology, biology, and, more recently, machine learning and data mining. Within the medical sector, it is prevalently utilized for disease diagnostics, epidemiology, empirical medical research, and radiological methods. The ROC curve's primary advantage is its ability to provide a lucid and direct visual representation of a diagnostic method's clinical precision.

IV. Results and Discussion

To optimize the model architecture, the simultaneous ECG and PPG dataset was utilized to evaluate three distinct model layers. The model demonstrating the highest accuracy was selected as the optimal architecture. The experimental results are presented in Table I. Table I presents the evaluation metrics of three different models utilized in the study: CNN, CNN + LSTM, and CNN + LSTM + ATTN. The performance of these models was assessed based on their accuracy and area under the curve (AUC) values for predicting blockages in three coronary arteries: the left anterior descending (LAD) artery, left circumflex (LCX) artery, and right coronary artery (RCA). Use letters for table footnotes. The corresponding ROC curve and confusion matrix are shown in Fig. 6.

As shown in Table I, The CNN model achieved an accuracy of 94.69%. When the LSTM layer was integrated with the CNN model, the accuracy slightly decreased to 92.47%. However, the introduction of the attention mechanism alongside the CNN and LSTM layers led to a significant improvement, with the CNN + LSTM + ATTN model achieving an accuracy of 98.12%.

TABLE I. Evaluation Metrics of Three Different Models Used in this Study

Model	Accuracy	AUC		
		LAD	***LCX***	***RCA***
CNN	94.69%	0.96	0.97	0.97
CNN + LSTM	92.47%	0.98	0.99	0.98
CNN + LSTM + ATTN	98.12%	1.00	1.00	1.00

For the detection of LAD blockages, the CNN model obtained an AUC of 0.96. The addition of the LSTM layer increased the AUC to 0.98, demonstrating the model's enhanced ability to capture temporal dependencies in the data. The incorporation of the attention mechanism further elevated the performance, with the CNN + LSTM + ATTN model reaching the maximum AUC of 1.00, indicating perfect classification.

In the case of LCX blockages, the CNN model achieved an AUC of 0.97. The CNN + LSTM model improved this metric to 0.99, showing a robust enhancement due to the LSTM layer. The CNN + LSTM + ATTN model again achieved the highest AUC of 1.00, reflecting its superior ability to focus on the most relevant features of the ECG and PPG signals for accurate detection.

For RCA blockages, the CNN model also achieved an AUC of 0.97. The CNN + LSTM model maintained a high AUC of 0.98, confirming the beneficial impact of temporal feature extraction. The CNN + LSTM + ATTN model reached a perfect AUC of 1.00, further validating the effectiveness of the attention mechanism in improving the model's discriminative power.

The comparative analysis reveals that the integration of LSTM and attention mechanisms into the CNN model substantially enhances its performance. The CNN + LSTM + ATTN model consistently outperforms both the standalone CNN and the CNN + LSTM models across all metrics. The inclusion of the LSTM layer helps capture temporal dependencies inherent in the sequential ECG and PPG data, thereby improving the model's ability to differentiate between classes. Furthermore, the attention mechanism selectively emphasizes the most relevant parts of the signals, thereby enhancing the model's focus and leading to more accurate classification.

The significant improvements in both accuracy and AUC, particularly the perfect AUC scores achieved by the CNN + LSTM + ATTN model, underscore its exceptional potential for precise detection of coronary artery blockages. This suggests that the combined approach not only leverages the strengths of each component but also synergistically enhances the overall model performance, making it a promising tool for the diagnosis of coronary artery disease using ECG and PPG signals.

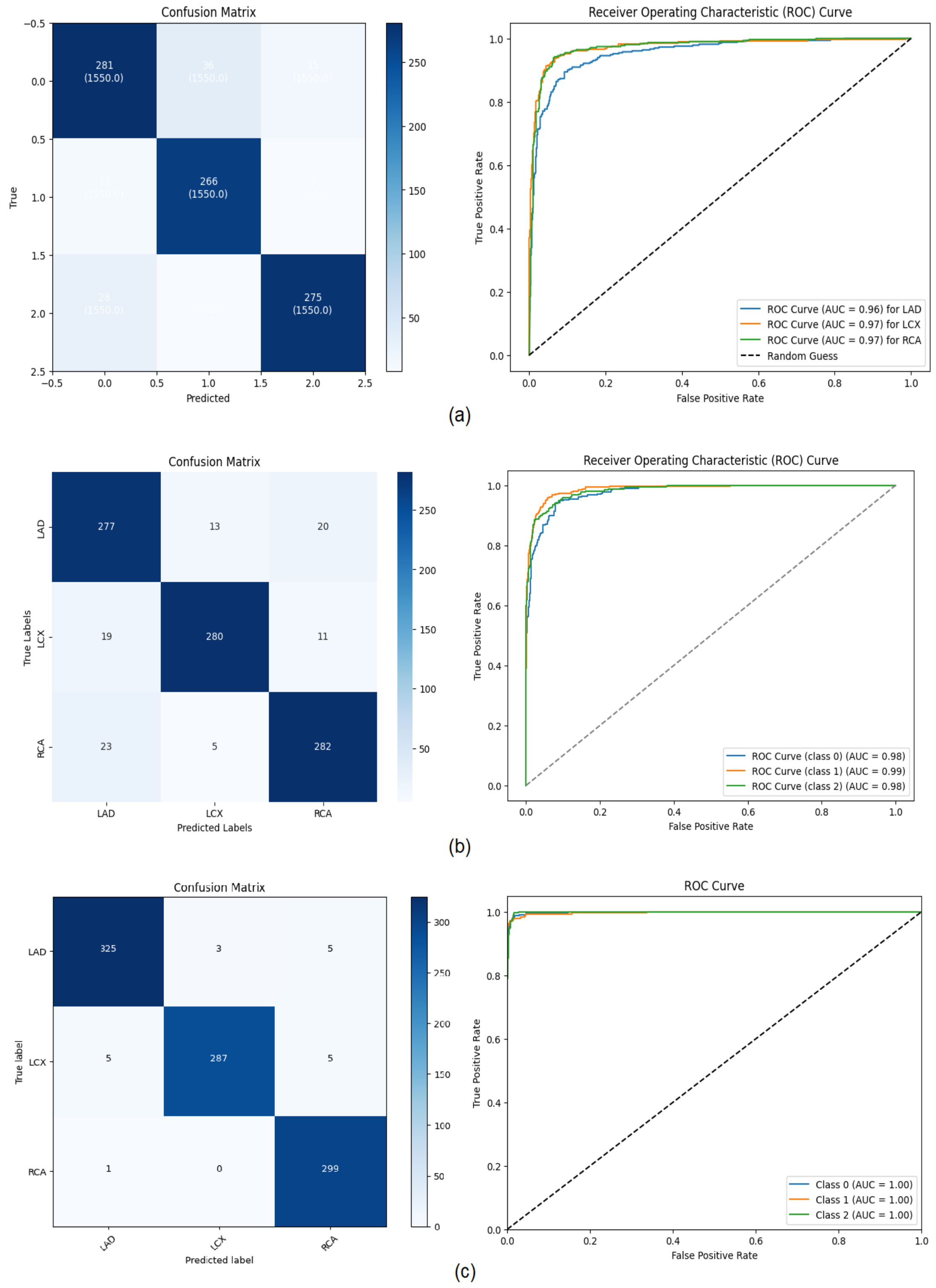


Fig. 6. Confusion matrix and ROC curve of (a) CNN model, (b) CNN + LSTM model, (c) CNN + LSTM + ATTN model.

Table II shows the model performance comparison of the best model obtained in this study with previous works by Tao et al., 2018 [18], Huang et al., 2022 [19], and Roopa and Harish, 2019 [20] in terms of their accuracy and AUC since these studies are closely related to the proposed work.

The results in Table II highlight the superiority of the proposed model, which utilizes combined simultaneous single-lead ECG and PPG signals with a CNN + LSTM + ATTN architecture, in predicting and localizing coronary artery blockages compared to previous studies.

TABLE II. PERFORMANCE COMPARISON WITH PREVIOUS WORKS

Author, Year	**Data**	**AI Model**	**Acc.(%)**	**AUC**		
				LAD	***LCX***	***RCA***
Tao et al., 2018 [18]	MCG	XGBoost	NA	0.74	0.68	0.65
Huang et al., 2022 [19]	12 lead ECG	InceptionV3		0.89	0.82	0.78
Roopa and Harish, 2019 [20]	12 lead ECG	IFN	92.3	NA		
Proposed work	Combined simultaneous single lead ECG and PPG	CNN + LSTM + ATTN	98.12	1.00	1.00	1.00

The proposed model achieved an overall accuracy of 98.12%, surpassing the performance of prior studies. In terms of Area Under the Curve (AUC) metrics for different coronary arteries, the model reached perfect scores (AUC = 1.00) for the left anterior descending (LAD), left circumflex (LCX), and right coronary artery (RCA). These results significantly outperform other models, as shown in the comparison.

Tao et al. [18] use magnetocardiography (MCG) data and an XGBoost model, this study reported lower AUC values of 0.74 (LAD), 0.68 (LCX), and 0.65 (RCA). Huang et al. [19] applied a deep learning model (InceptionV3) on 12-lead ECG data, achieving AUCs of 0.89 (LAD), 0.82 (LCX), and 0.78 (RCA). Roopa and Harish [20] employed an Information Fuzzy Network (IFN) model using 12-lead ECG data with a reported accuracy of 92.3%.

The significant improvement in both accuracy and AUC values of the proposed model can be attributed to the integration of ECG and PPG signals, coupled with the advanced AI architecture combining CNN, LSTM, and attention mechanisms. The LSTM layers effectively capture temporal dependencies in sequential data, while the attention mechanism enhances feature extraction, leading to more precise localization of coronary artery blockages.

These results indicate that the proposed model represents a substantial advancement in the non-invasive diagnosis and localization of coronary artery disease, providing a more accurate and reliable approach compared to existing methods. This makes it a promising tool for future clinical applications in CAD detection.

V. CONCLUSION

In this study, a novel approach for diagnosing the location of coronary artery blockages in CAD patients by utilizing a combination of ECG and PPG signals was presented. The proposed model integrated Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Attention (ATTN) mechanisms to enhance the accuracy and robustness of CAD detection.

The experimental results demonstrated that the CNN + LSTM + ATTN model significantly outperformed the standalone CNN and CNN + LSTM models. Specifically, an accuracy of 98.12% and perfect Area Under the Curve (AUC) scores of 1.00 for detecting blockages in the left anterior descending (LAD) artery, left circumflex (LCX) artery, and right coronary artery (RCA) were achieved by the CNN + LSTM + ATTN model. The superior performance of the attention mechanism in selectively emphasizing the most relevant parts of the ECG and PPG signals, thereby improving the model's discriminative power, was underscored by these results.

The integration of the LSTM layer was found to further contribute to the model's ability to capture temporal dependencies inherent in the sequential ECG and PPG data, enhancing its capacity to differentiate between different classes of coronary artery blockages. The significant improvements in both accuracy and AUC scores highlighted the exceptional potential of the CNN + LSTM + ATTN model for precise detection of CAD.

In conclusion, the combined approach not only leveraged the strengths of each component but also synergistically enhanced the overall model performance, making it a promising tool for the diagnosis of coronary artery disease using ECG and PPG signals. Future work will focus on further validating the model with larger and more diverse datasets, as well as exploring its applicability in real-world clinical settings.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support received from the Ministry of Higher Education Malaysia through the Trans-disciplinary Research Grant Scheme (TRGS/1/2019/UKM/01/4/3).

REFERENCES

[1] A. Joshi and M. Shah, “Coronary Artery Disease Prediction Techniques: A Survey,” in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2021, pp. 593–604. doi: 10.1007/978-981-16-0733-2_42.

[2] L. Zhang et al., “Global, Regional, and National Burdens of Ischemic Heart Disease Attributable to Smoking From 1990 to 2019,” J Am Heart Assoc, vol. 12, no. 3, Feb. 2023, doi: 10.1161/JAHA.122.028193.

[3] G. Femia, J. K. French, C. Juergens, D. Leung, and S. Lo, “Right ventricular myocardial infarction: pathophysiology, clinical implications and management,” Dec. 22, 2021, IMR Press Limited. doi: 10.31083/j.rcm2204131.

[4] M. Matta, S. C. Harb, P. Cremer, R. Hachamovitch, and C. Ayoub, “Stress testing and noninvasive coronary imaging: What’s the best test for my patient?,” 2021, Cleveland Clinic Educational Foundation. doi: 10.3949/ccjm.88a.20068.

[5] U. Sechtem, “Non-invasive testing in patients with suspected coronary artery disease: Some may be more equal than others,” Eur Heart J, vol. 39, no. 35, pp. 3331–3333, Sep. 2018, doi: 10.1093/eurheartj/ehy364.

[6] S. Hadiyoso, F. Fahrozi, Y. S. Hariyani, and M. D. Sulistiyo, “Image Based ECG Signal Classification Using Convolutional Neural Network,” International journal of online and biomedical engineering, vol. 18, no. 4, pp. 64–78, 2022, doi: 10.3991/ijoe.v18i04.27923.

[7] Y. Kaolawanich, R. Thongsongsang, T. Songsangjinda, and T. Boonyasirinant, “Clinical values of resting electrocardiography in patients with known or suspected chronic coronary artery disease: a stress

perfusion cardiac MRI study," BMC Cardiovasc Disord, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12872-021-02440-5.

[8] L. Jeanningros et al., "Pulse Wave Analysis of Photoplethysmography Signals to Enhance Classification of Cardiac Arrhythmias," in Computing in Cardiology, IEEE Computer Society, 2022. doi: 10.22489/CinC.2022.023.

[9] G. N. Georgieva-Tsaneva and E. Gospodinova, "Comparative Heart Rate Variability Analysis of ECG, Holter and PPG Signals." [Online]. Available: www.ijacsa.thesai.org

[10] H. M. Rai, K. Chatterjee, A. Dubey, and P. Srivastava, "Myocardial Infarction Detection Using Deep Learning and Ensemble Technique from ECG Signals," in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2021, pp. 717–730. doi: 10.1007/978-981-16-0733-2_51.

[11] A. A. Ahmed, W. Ali, T. A. A. Abdullah, and S. J. Malebary, "Classifying Cardiac Arrhythmia from ECG Signal Using 1D CNN Deep Learning Model," Mathematics, vol. 11, no. 3, Feb. 2023, doi: 10.3390/math11030562.

[12] B. Tutuko et al., "Empowering AI-Diagnosis: Deep Learning Abilities for Accurate Atrial Fibrillation Classification," International journal of online and biomedical engineering, vol. 19, no. 17, pp. 134–151, 2023, doi: 10.3991/ijoe.v19i17.42499.

[13] A. Makhir, M. H. El Yousfi Alaoui, and L. Belarbi, "Comprehensive Cardiac Ischemia Classification Using Hybrid CNN-Based Models," International journal of online and biomedical engineering, vol. 20, no. 3, pp. 154–165, 2024, doi: 10.3991/ijoe.v20i03.45769.

[14] A. Kuvaev and R. Khudorozhkov, "An Attention-Based CNN for ECG Classification," in Advances in Intelligent Systems and Computing, Springer Verlag, 2020, pp. 671–677. doi: 10.1007/978-3-030-17795-9_49.

[15] B. Zhu and G. He, "Myocardial infarction localization and blocked coronary artery identification using a deep learning method," in Proceedings - 11th International Conference on Prognostics and System Health Management, PHM-Jinan 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 514–519. doi: 10.1109/PHM-Jinan48558.2020.00100.

[16] F. Esgalhado, B. Fernandes, V. Vassilenko, A. Batista, and S. Russo, "The application of deep learning algorithms for PPG signal processing and classification," Computers, vol. 10, no. 12, Dec. 2021, doi: 10.3390/computers10120158.

[17] K. H. Miao and J. H. Miao, "Coronary Heart Disease Diagnosis using Deep Neural Networks," 2018. [Online]. Available: www.ijacsa.thesai.org

[18] R. Tao et al., "Magnetocardiography-Based Ischemic Heart Disease Detection and Localization Using Machine Learning Methods," IEEE Trans Biomed Eng, vol. 66, no. 6, pp. 1658–1667, Jun. 2019, doi: 10.1109/TBME.2018.2877649.

[19] P. S. Huang et al., "An Artificial Intelligence-Enabled ECG Algorithm for the Prediction and Localization of Angiography-Proven Coronary Artery Disease," Biomedicines, vol. 10, no. 2, Feb. 2022, doi: 10.3390/biomedicines10020394.

[20] C. K. Roopa and B. S. Harish, "Automated ecg analysis for localizing thrombus in culprit artery using rule based information fuzzy network," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 16–25, 2020, doi: 10.9781/ijimai.2019.02.001.

Increasing the Performance of Iceberg Query Through Summary Tables

Gohar Rahman[1], Wajid Ali[2], Mehmood Ahmed[3], Hassan Jamil Sayed[4], Mohammad A. Saleh[5*]
Faculty of Computing and Informatics, University Malaysia Sabah (UMS), 88400 Kota Kinabalu, Sabah Malaysia[1, 5]
School of Computer Science and Technology, Jilin University, Changchun, Republic of China, China[2]
Department of Information Technology, The University of Haripur, Haripur, KP, Pakistan[3]
Asia Pacific University of Technology & Innovation (APU) Bukit Jalil, Kuala Lumpur, Malaysia[4]

***Abstract*—One of the key challenging problems in data mining is data retrieval from large data repositories, as the sizes of data are growing very fast, to deal with this situation, there is a need for efficient data mining techniques. For efficient mining tasks number of queries have been emerged. Iceberg query is one of them, in which the output is much smaller like the tip of the iceberg as compared to the large input dataset, these queries take very long processing time and require a huge amount of main memory. However the processing devices have limited memories, so the efficient processing of iceberg queries is a challenging problem for most of the researchers. In this paper we present a novel technique, namely a summary table, to address this problem. Specifically, we adopt the summary table technique to acquire the required results at summary levels. The experimental results demonstrate that the summary table technique is highly effective for large datasets. Compared to bitmap indexing and cubed techniques, the summary table offers faster retrieval capabilities. Furthermore, the proposed technique achieved state-of-the-art performance.**

Keywords*—Threshold (TH); bitmap index; aggregate function; Iceberg Query (IB); anti-monotone; non-anti-monotone aggregation

I. Introduction

Data retrieval and storage play a very important role in databases. The effectiveness of data retrieving techniques depends on specific. Since few years many queries have emerged, one of them is the iceberg query (IBQ) in which the output is significantly small as compared to the input, such query is called IBQ, where the number of above-threshold outcome is usually very small like the tip of an iceberg as compared to large amount of input data [1]. This is a unique class of aggregation queries connecting HAVING () and GROUP BY () clauses, which computes aggregated values below or above a given threshold (TH). This query is first introduced in data mining (DM) [2]. Most of the data DM queries are IB queries. Several applications use aggregate functions such as, Min (), Avg (), Max (), Sum (), and Count () over an attribute or set of attributes to find aggregate values greater than a particular threshold, these aggregate values above the threshold values give more importance. The RDBMS, e.g., MySQL, Postgre SQL, SQL Server, Oracle, DB6, Sybase, and column-oriented databases e.g., Lucid DB, Vertica, and Monet DB all use common aggregation algorithms that first aggregate all rows and then calculate the Having () clause to select the iceberg result [3].

An iceberg query has the following characteristics: (a) Computing aggregate functions on one or more attributes (b) Dealing with large data sets, containing large unique attributes combination (domain size), and (c) Returning results below or above a given TH. These queries face some problems during executions, like 1) It needs to execute within a limited memory, which means memory size is lesser than domain size 2) Computation of aggregation values takes a large amount of time.

The global objective of this work is to reduce the execution time of the iceberg queries within a limited memory. Today’s world is rich in data; every organization and social media generates and stores huge amounts of data which need an efficient way to deal with. For this purpose, IB query is an ideal choice. These queries are used in many applications. Including market basket analysis [4] means finding item pairs (or triplets etc.) that are bought together by many customers in large data warehouses. In other words, market basket analysis means a collection of items purchased by a customer in a single transaction. It is based on two key attributes considered for the threshold value used for finding item pairs; these attributes comprise support and confidence. If support and confidence values are above or below some specified threshold then it identifies products and their content that go well together. Similarly, clustering [5] is a process of partitioning a set of records into groups (clusters), such that all records in a group are related to each other and records that belong to two different groups are different [6]. This helps users to recognize the natural grouping or structure in a data set and this natural grouping is done in clustering based on some specific threshold values in each IB query [7].

A. *Properties of Aggregation Function*

Aggregation function is one the key part of iceberg queries, such as Sum (), Count (), Min (), Max (), and Avg (). Aggregate function is divided into two types (1) Anti-Monotone, and (2) Non-Anti-Monotone aggregation function [8]. An anti-monotone uses apriori [4] property, but non-anti-monotone are not able to use apriori property, examples of anti-monotone are Count (), Sum (), Max (), and Min (), whereas non-anti-monotone are Avg () and Div (). The main benefit of using IB with anti-monotone function is the pruning of computing aggregation functions reduces the time to produce the required query result [9]. On the other hand, non-anti-monotone aggregation IB queries don't take advantage of threshold on Avg () values as anti-monotone aggregation takes on Min (), Sum (), Max (), and Count (). Average IB queries compute average for all unique grouped attributes, and then apply threshold constraint on those average values [10]. To deal with aggregation functions there is a high gape between the researcher's contribution toward these two types with a ratio of 22 to 78 percent [11] as shown in Fig 1.

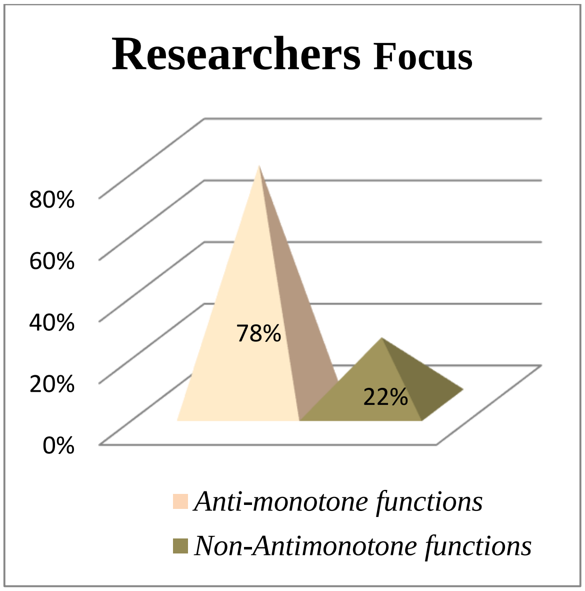


Fig. 1. Aggregation functions.

The rest of the paper is structured as follows, section II presents a review of related research. Section III describes the proposed technique, proposed architecture, and implementation. Section IV describes the results and analysis, and in Section V conclusion and future direction are presented.

II. Review of Related Research

Since a few decades iceberg query has always been an active area of research. Researchers have provided different guidelines and suggestions to improve its performance. We are going to discuss some works in literature based on different specific categories.

A. Bitmap Index Techniques

To accelerate the IB queries, bitmap indices are one of the well-organized and well-known choices in column stores and data warehousing applications. Spiegler et al. [12] first introduced the concept of bitmap index (BI). Basically, a matrix of 0 and 1 bit's makes a bitmap. Its size depends on the number of matchless attributes that exist in vector upon which bitmap is created. Basically, a bitmap index is used to index values of a single column in a table. For illustration Table I indicates a bitmap index with nine rows, and column Y, where column Y is indexed with integer values from 0 to 3 and its cardinality becomes four because it has four different values. Columns X0, X1, X2 and X3 with subscripts signify bitmap index for Y contains four bitmaps. The second bit X1 in Table I is 1 because the second row of Y contains value 1, while corresponding bits of X0, X2 and X3 are all 0 Vuppuand Rao [13] has presented a new evaluation scheme for processing IB queries using bitmap index position. They developed an algorithm based on retrieving index positions of all 1's from each bitmap. Further, these indices positions are processed by using commonality condition which selects whether the pair of directions is iceberg result or not. To retain for future reference, an XOR operation is conducted for bitmaps, which is the iceberg query result. Their experiments show that algorithms which is based on index positions takes less processing time to answer IBQ.

TABLE I. Bitmaps Index for Column Named Y

RID	Y	X0	X1	X2	X3
0	2	0	0	1	0
1	1	0	1	0	0
2	3	0	0	0	1
3	0	1	0	0	0
4	3	0	0	0	1
5	1	0	1	0	0
6	0	1	0	0	0
7	0	1	0	0	0
8	2	0	0	1	0

Padmapriya and Shanmugapriya [1] introduced an index based IBQ assessment method. The key aim of using the index is to convert the bit value into an integer value which speeds up the query evaluation process and takes less memory. This technique performed well on the state-of-the-techniques. O'Neil [14] Model 204 was the first model used bitmap index for wide-spread commercial product making. This was a combination of row identifiers (RID list) and basic bitmap index without compression. In general B+ tree index technique is like the performance of Model 204. Prakash et al. [15] presented a bitmap index as a better choice for querying huge and multidimensional scientific datasets. They have developed a well-organized algorithm based on retrieving index positions of all 1's from each bitmap. These index positions are further processed on common features which decide whether the pair of vectors is IB result or not. Generation of the decision algorithm which involves pre-processing of data sets through bitmap indexing approach is the global objective of [16]. The key benefits of this index strategy are load balancing, identifying frequent patterns of the data sets, kind of data types available in the databases, slowly changing dimension scenarios handling and usage of aggregation in the form of IB querying.

Shankar et al. [17] introduced a cache-based evaluation technique for IBQ by taking threshold value equal to 1 using a compressed bitmap index, and for future situation the required results are saved in cache memory. In future it just picks up the required results from the cache memory as a substitute of executing once again on the database table. Therefore, this approach clearly states that, an execution time of IB query is improved by avoiding duplication of evaluation process several times. In this work testing was conducted by applying an IB query stated on the database table which consists of one million rows with two attributes X and Y, by using COUNT () aggregation function. IB query evaluation method was the first function applied to accept all those tuples as an input and produces the iceberg results with its count value fulfilled by threshold greater than or equal to 1. Then these results are given as an input to the second unit catching IB results in an ascending order. For future position this unit saves the iceberg results in cache memory. The last unit which takes results from cache unit is responsible to answer an IB query for thresholds greater than 1, is just selected from the cache memory and send to output. This cache-based technique enhanced the overall processing time of IBQ for an efficient data retrieval task.

Laxmaiah, Govardhan and Kumar [18] have presented an efficient Database Priority Queue (DPQ) algorithm for processing IBQ using compressed bitmap Index; the impact of this method was to speed up the query evolution method by emptying the compression queue. In this technique, first the iceberg query is responsible to select the similar words with aggregate attributes Y and X from the relation R in which the TH value is taking between 1000 and 9000. By taking a database table which has two attributes Y and X which contain millions of rows and using count () aggregate function. Then the experiment is conducted by applying an IBQ on the first function that generates bitmaps accepts all those rows as an input. The key achievement of this technique is keeping the comparable number of rows in a relational database table and at the same time keeps the results with density queues; consequently, further this experiment is repetitive for different iceberg threshold as well.

Zianget al. [19] suggested a well-organized algorithm for IBQ processing by using compact bitmap indices. The given algorithm does not depend on any testing compression process and demonstrates better performance over presented schemes. Bitmap index has three attractive benefits based on observation such as: first, conducting bitwise operations that reduce computation time. Second, saving disk space by avoiding rows scan on a relation by using attributes group. Third, by leveraging the anti-monotone property of IB therefore this algorithm is not affected by the number of diverse values, and length of attributes in the relation. Rao et al. [20] presented a well-organized technique, known as dynamic pruning technique or vector alignment algorithm to answer IBQ by using compressed bitmap indices, this algorithm guarantees that no empty result is generate by using any bitwise-AND operation. Bitmap indices are presented to get more improved results as compared to tree based index method such as alternatives of R-tree or B-tree [21], explicitly, this work is motivated to compute IBQ using bitmap indices as an index pruning based approach.

Otoo and Shoshani [22] introduced bitmap indexing pattern algorithms for little cardinality attributes to analyze the time and space complexities of Byte Aligned Code (BBC) and Word Aligned Hybrid (WAH) compressed bitmap indices. To demonstrate their success for using high cardinality attributes, for high cardinality attributes, c << N, here c represent compressed bitmap index and N represent the number of words, the WAH algorithm compressed indices uses define 2N words, this 2N words is about half the size of a representative B-tree index. On the other side BBC compressed indices are even smaller but it also represents an in-place algorithm that is linear to the total size of the bitmaps involved to OR many bitmaps in time complexity. The whole size of the bitmaps used is proportional to the number of hits in the worst-case situation. By using compressed bitmap, it shows to search one attribute is optimal and this optimality is established with timing results from a set of real application and random data. In these sets of examinations, WAH compressed bitmap indices were nearly twice as fast as BBC code compressed indices. Both indices could achieve search operations faster than the projection index by using worst cases, on average. By using the WAH compressed indices, time is not more than the projection index. Bitmap indices in study [14] using bitmap vectors for vertical organization of a columns. Every vector characterized the presence of a distinctive value in the column across all rows in the table.

In study [23] an effective bitmap pruning strategy was introduced, which is grounded in order of high cardinality in Priority Queue (PQ) by using compressed bitmap indices for processing an IBQ. By using this method, it allowed the movement of vectors to enter PQs on the high count 1' to get additional benefit for large pruning of bitmaps. The pruned vector essentially improved the response time. Processing huge quantities of data in predetermined time factor is a key challenge faced by data warehousing. By using [24] bitmap indexing which is extra meaningful in quicker data processing generated a strategic decision technique for data warehousing environment. For managing 'Boolean' kind of data, like gender, the bitmap indexing is best suited, such as false and true group of values. Bitmap indexing mainly depends on 1's and 0's kind of data. The data is openly processed by using CPU, which does not support any alteration of the data items into a new format, and greatly decreases the processing time of the records. This work shows the integration of IB querying; within the identified amount of time factors while processing huge amounts of data in data warehousing environments and achieved efficient results. Most of the previous research work mainly centered about identifying "well behaved "constraints with respect to constraint pushing [25, 26], this work proposed a novel pushing technique known Divid-and-Approximate (DnA), which combine two ideas, "Approximate Push" and "Divide-and-Conquer" to generate a strongest constraint for pruning with non-anti-monotone aggregation constraints in IB cubing. The key idea of DnA was to divide a partition of tuples into two subspaces of positive and negative degree values, so that a given constraint could be rewritten using monotone or ant monotone constraints in subspaces. These works mainly focused on (a) SQL like tuple-based aggregates, rather than item-based aggregates (b) General aggregate constraints, rather than only "well behaved", and (c) Constraint independent methods, rather than constraint specific methods. The idea of DnA contributed a new share to constrain data mining techniques.

Laxmaiah et al. [27] presented an efficient Density Priority Queue (DPQ) procedure for an IBQ by using compacted bitmap index based on two stages (1) Using an algorithm for pruning the vectors dynamically by computing newest counts for reinsertion and certifies the proposal using a sample database. By dropping the bitmap vectors dynamically using a high-count attribute to calculate an IB query, and (2) Using a validation of DPQ approach on RDBMS section to show the validity of the proposed DPQ and evaluates an IB query having COUNT () aggregate function. As compared to previous strategies PQ is a more sophisticated technique. Based on large data sets the experimental results indicate significant progress which proves the effect of IBQ computation.

In study [28], the distributed Iceberg Semi-Join operator is proposed, which is used in most of the real-life applications.

This technique is used to get information from two different independent data marts or from a remote digital library and use an efficient technique, which insert the execution of the IBQ and join in the two servers by using Mul-FIS; to prune the non-qualifying groups it uses. This work provides important advantages over its competitors. By using multidimensional databases each dimension is nothing, but one subject oriented table with attributes of related metrics [29], writing queries on multidimensional databases are difficult and include join operations due to which the reply time of query is increased on a massive database. By using queries with aggregation function, and summarization is followed by using 'having' clause. This type of query is very complex and requires extra time. This work focused on two different bitmap indexes search implementations techniques, such as RIDB and Fast Bit. The key improvements in Fast Bit are the Word-Aligned Hybrid (WAH) compression for bitmaps and multilevel bitmap encoding approaches. Fast Bit index is typically greater than RID Bit index, in fewer intervals of time it can answer several queries, as it accesses the required bitmaps in fewer I/O operations [30]. RID Bit normally costs less CPU time in answering queries than that of Fast Bit, though, the CPU time differences are minor matched with I/O time. A brief comparison between "Fast Bit" and "RID Bit" is discussed in [31]. At the end this section Table II categorizes some basic characteristics of bitmap Indexing.

B. *Based on Compound or Hybrid Algorithms*

In study [31], four algorithms namely Partitioned Tree (PT), Breadth-first writing Partitioned Parallel BUC (BPP), Replicated Parallel BUC (RP), and Affinity Skip List (ASL) are introduced, these algorithms are calculated experimentally over a range of parameters to get the necessary condition in which the algorithms could outperform. The Key features of the proposed algorithms are mentioned in Table 2, which described all four algorithms with respect to their writing strategy, Data decomposition, Load Balance, and Relationship of cuboids.

Matias and Segaly [32] presented two effective algorithms based on hash partitioning technique to compute estimated IB queries. Using a hash function to divide a data set into specific values that was independent from a subset resulting with properly smaller independent sub problems that can be handled efficiently with certain performance. In [33] two algorithms which use a concise sample and basic component have been presented. The first algorithm is used to sort the sequence into the necessary number of partitions and the second algorithm is used for computation. Though acting only one pass over the sequence these algorithms are used to compute the approximation query, without accessing a database and without materializing data sets which are stated implicitly, therefore it can be applied online for streaming data. In [34] the author has emphasized two problems; (1) Efficiently classify passing stories from rapid streaming social content and (2) To execute IB queries to form the structural background between stories. To give attention to the solution of the first problem, the social stream is converted into a time gap of tube network, and model passing stories as (k, d) cores in the tube network.

Two polynomial time algorithms were proposed to extract maximal (k, d) cores. The second problem, deterministic context searches and randomized context search is applied to maintain the IBQ efficiently and carefully, which permits performing context search without pair wise relationship.

TABLE II. KEY FEATURES OF FOUR ALGORITHMS

Algorithms	Writing Strategy	Data Decomposition	Load Balance	Relationship of cuboids
RP	Depth-First	Replicated	Weak	Bottom-up
BPP	Breadth-First	Partitioned	Weak	Bottom-up
ASL	Breadth-First	Replicated	Strong	Top-down
PT	Breadth-First	Replicated	Strong	Hybrid

By spreading the probabilistic techniques and suggested hybrid and multi buckets algorithms for processing of IB queries was first considered by study [35]. The sample and multiple hash function are used as an important building chunk of probabilistic events such as scaled-sampling course and count algorithms. It projected the sizes of a query results in order to expect the valid IB results, which decreases memory requirements and raises aggregate query performance. Though, these techniques incorrectly resulted in false negatives and positives. To overcome these bugs, an efficient approach is planned by hybridizing the sampling and coarse count techniques, such as hashing technique that allocated a bitmap of size 'M' in the memory is constructed on linear counting algorithm (LCA). In this method, all entries are initialized with '0's. The linear counting algorithm applies a column interest and then scans the relation. On the other hand, the hash function produces a bitmap address, and the algorithm sets this addressed bit to '1'. This algorithm first counts the number of empty bitmap entries. Then it guesses the column cardinality by distributing the count by the bitmap size 'm' and plugging the given result which increases the overall performance of hybrid technique.

III. PROPOSED TECHNIQUE

In the previous section we discussed in detail IB query processing techniques and algorithms. Researchers have introduced different algorithms and techniques for increasing the performance of iceberg query. Some of the existing techniques focus on the SUM (), MIN (), MAX (), AVERAGE (), and COUNT () aggregate functions, such as, bitmap indexing techniques, cubed techniques, AND operation techniques, POP operation techniques, attribute-based techniques, and hybrid algorithm techniques. All these techniques have some limitations, such as, some techniques have the deficiency to occupy more space in memory, some techniques slow down the system performance, some require complex algorithms which are difficult to maintain, and some take more time to produce the required result in a required time. To overcome the limitations of existing techniques to improve the performance of IBQ, an enhanced technique based on, summary table is proposed to improve the IB query performance. This technique improves the running time of the query for searching a specific record and reduces the elapse time. This section aims to discuss the details of the proposed work using summary table's technique for IB query processing to finds out the required result greater than the given TH.

The proposed technique is applied on sample customer tables of different sizes with the same attributes, such as customer identity (Cus_Id), Expense per day (Cus_Exp-per-day), and job (Job). The values of Cus_Id, Cus_Exp-per-day, and Job attributes classify each group, while Cus_Exp-per-day refers to the field on which the aggregate operator Count () is being computed based on a specific TH value. The focus of this work is Count () aggregate function which is applied on (Cus_Exp-per-day) field, the scope of the proposed work is to find expenses of all those customers whose expense is greater than some specified threshold values. The proposed work is better than the existing work in different perspectives. Different summary tables are created in the proposed work and IBQs are used to extract the required results from these summary tables by ignoring scanning whole data sets one by one. As compared to the proposed technique, the state-of-the-art technique was used to scan all the tables for the required data results which slowed down the query processing time. The advantage of our proposed technique is to improve the running time of the query for searching a specific record and reduces the elapse time.

A. *Proposed Architecure*

The proposed architecture is a robust and efficient summary table creation system based on different threshold values for identifying how summary tables are created from the original tables. Fig. 2 draws our novel architecture, which consists of three phases: an Execute1 phase for processing simple IB queries, that is directly applied on original source table for extracting the required dataset, it scan the whole table for the required data set based on a specific threshold value, the drawback of this technique was the requirement of huge amount of processing time which effect the performance efficiency of this technique, in contrast the Execute phase in the mentioned architecture was used to automatically creating summary tables from original table, and then on those summary tables the required query are executed based on a specific threshold value for the required result instead of scanning the whole table. In the proposed architecture only eight summary tables are mentioned for the sake of simplicity these summary tables are used as a source of IB queries, instead of scanning the whole original data set, and the third phase displays the required output of the processed IB query.

B. *Implemetation*

This section describes different techniques that were planned in the preceding section. The first step toward implementation was by taking different target customers tables ranging from 50K to 500k data set, which store customers' related records. In the second step the existing technique is applied to fetch customer's records by using twenty different threshold values ranging from 1k to 20k on each customer's table. Based on the mentioned threshold values, the times taken by the existing technique are recorded and then the average time is calculated for ten run cycle on a given table. The same experiment is repeated ten times for each data set in total. Similarly, the proposed technique is represented in the same way on the same data accordingly and twenty summary tables are created for each target table, that store pre-calculated aggregate results of the COUNT () aggregate function. The proposed technique then considers the threshold values given in the HAVING clause of the IB query and fetches results from the respective summary tables as per the given threshold values.

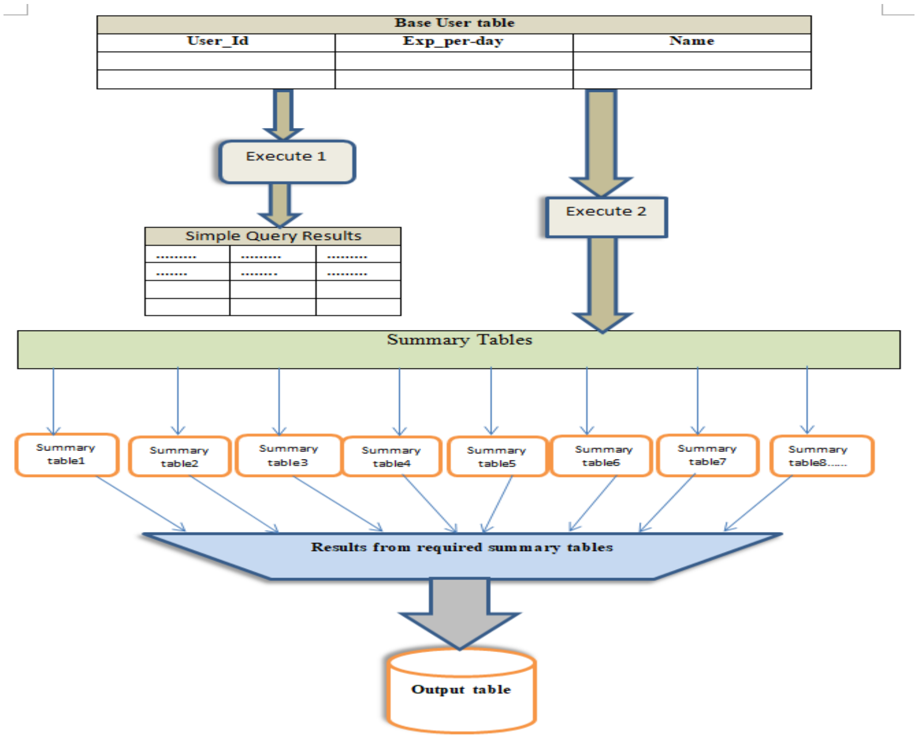


Fig. 2. Proposed architecture.

IV. Result and Analysis

In this section, results obtained during experiments are analyzed. Table III and Fig. 4 represent comparison among existing and proposed techniques based on query performance for different dataset ranging from 50k to 500k. The first column in Table III indicates different data sets, the second and third column correspond to average processing time of both techniques by using ten cycle of run count testing condition on each specific data set accordingly. The “Existing technique time” and “Proposed technique time” represent the average processing time of the existing and proposed IB query by retrieving the required result. The recorded values show high differences in execution times. e.g. when the dataset was 50k, then the corresponding average values of ten count cycle of existing and proposed techniques was about 6.2ms and 1.02ms, similarly for 100k the corresponding values is 6.8ms and 1.1ms, and the same is recorded for the other data sets till to 500k respectively. The above tabulated values are shown in the following Fig. 3 to understand well to the reader, the x-axis represents different datasets and y-axis represents execution time. Then how the execution times of existing and proposed techniques are gradually varying with different datasets.

A. Comparison between Simple and Proposed Techniques

In Table IV, we drawn the comparison between simple and proposed techniques only for one dataset of size 50k, there are nine others different tables is used to store the same comparison used in Table III with different data set (100k, 150k, 200k, 250k, 300k, 350k, 400k, 450k, and 500k) during the whole experiment to record the processing time of both techniques, which is not mentioned in this section due to a large number of comparison. Table IV represents the comparison of “Average of ten run cycle” among simple and proposed query processing. In the Table IV “RECORDS” field represent the number of records in given table, “Cycle of Run Count” represents ten counts of query processing of each proposed and simple technique, “Simple Query Technique” field represent the state-of-the-art technique, “Proposed Query Technique” field represents the proposed technique and “AVERAGE OF ALL” field represent the average time of ten cycle processing of each proposed and simple query technique. To comprehend well to readers, we draw the above tabulated data in graph form which are shown in the following Fig. 4. The x-axis represents the number of processing run count ranging from 1 to 10 cycles, and as well as average of all run count cycles for both simple and proposed techniques, and y-axis represents the execution time in microsecond.

TABLE III. Existing and Proposed Technique Time Interval

Dataset	Existing technique time	Proposed technique time
50K	6.2ms (average time of ten run cycle)	1.02 ms (average time of ten run cycle)
100K	6.8ms	1.1 ms
150K	8.7ms	1.3ms
200K	10.5 ms	1.6 ms
250K	12.5 ms	1.9 ms
300k	14.7 ms	2.1 ms
350K	16.5 ms	2.3 ms
400k	18.3 ms	2.7 ms
450k	20.1 ms	2.9 ms
500k	22.4 ms	3.1 ms

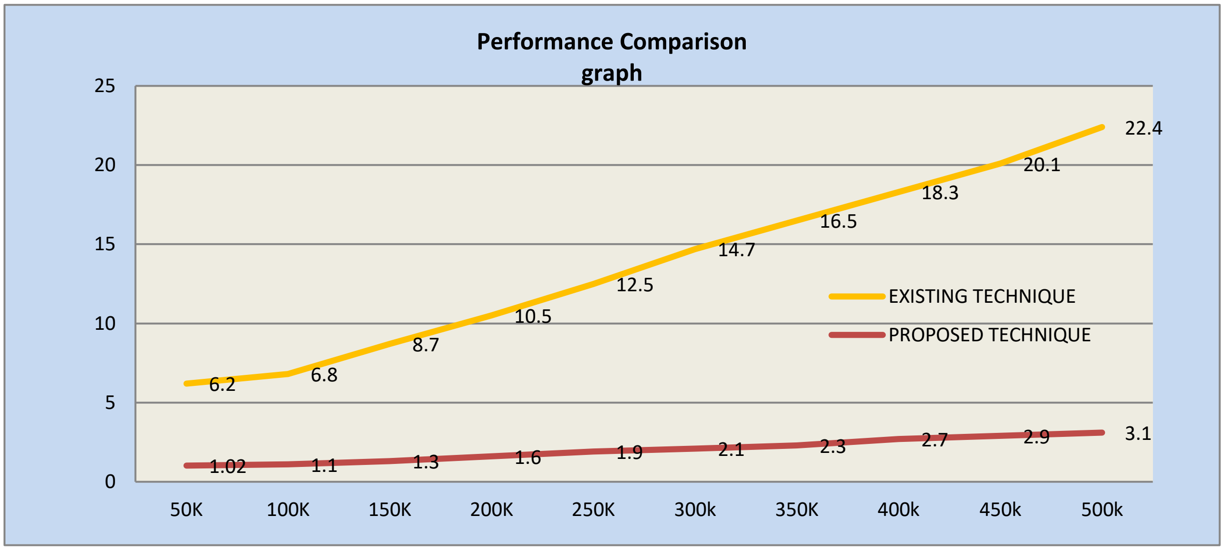


Fig. 3. Performance comparison of existing and proposed techniques.

TABLE IV. COMPARISON OF CYCLE OF RUN COUNT BETWEEN PROPOSED AND EXISTING TECHNIQUE

Table 4	50k Records										
Cycle of Run Count	1	2	3	4	5	6	7	8	9	10	AVERAGE OF ALL
Simple query processing time	6	6.01	6.05	6.07	6.08	6.1	6.3	6.6	6.7	6.9	6.28
Proposed query processing time	1	1.02	1.03	1.05	1.05	1.06	1.08	1.09	1.1	1.4	1.08

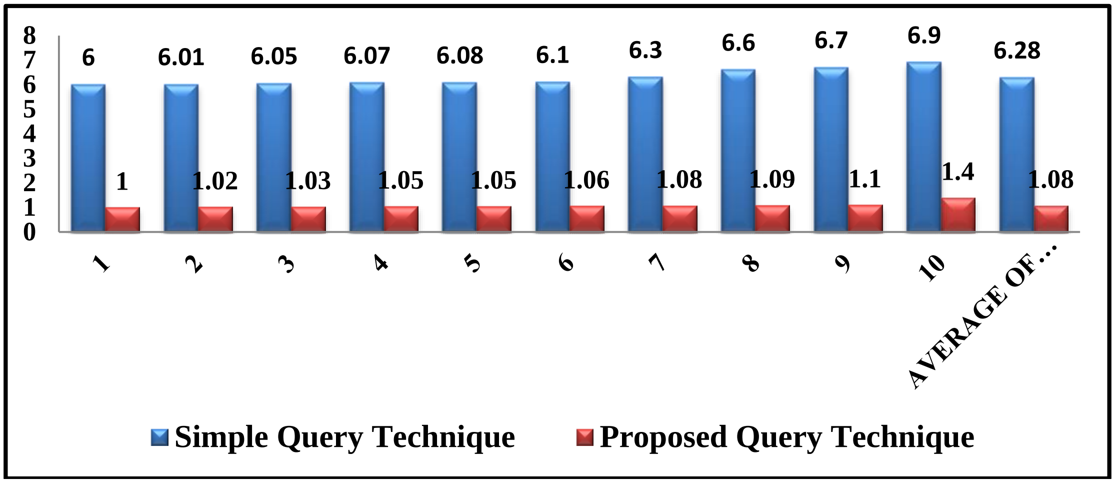


Fig. 4. Comparison of cycle of run Count between existing and proposed technique.

V. CONCLUSION

In this paper, we proposed a summary table technique for processing iceberg queries efficiently. According to IBM 2.5 quintillion bytes of data are generated by different electronic devices on a daily basis. Extraction of useful information from those huge data sets is a well-known challenging problem. From the last few decades most of works have focused to increase the performance of IB queries with respect to time constraint, computing memory, data repositories, computing memories, and data scanning. In this paper our technique is highly simple and different as compared to the previous works based on bitmap indexing and cubed technique. The summary table technique leverages the IB queries at the summary tables; these summary tables are created dynamically from the base table based on different threshold values ranging from 1k to 20k before the execution of queries. At the time of issuing the query, the proposed technique considers the threshold given in the query and fetches the calculated COUNT () aggregation from that specified summary tables as opposed to recalculating the aggregate function from the base table to produce the required results. Our method has improved the main metrics significantly. However, to increase the performance of IB queries in a large dataset is still a big challenge, in the future, this research work focused on the COUNT () aggregate function. In future we would like to extend our work to other aggregate functions such as MAX (), MIN (), SUM (), and AVG (). In this work we only considered the high-level IB queries. Similarly, we would like to continue working on low-level queries as well.

ACKNOWLEDGMENT

The authors would like to thank the University Malaysia Sabah for supporting this research Under the Grant NALS/ SLB2451NALS.

REFERENCES

[1] Ramírez-Gallego, S., García, S., Bergmeir, C., Triguero, I., Mendoza, C., & Herrera, F. (2018). Fast distributed big data preprocessing using Spark. IEEE Access, 6, 21216-21231

[2] Glavic, B., & Alonso, G. (2021). Verifying data-centric programs: Are databases the new bell-bottoms? Communications of the ACM, 64(7), 93-101.

[3] Godfrey, P., & Shipley, R. (2020). Iceberg queries revisited: A multi-dimensional perspective. Proceedings of the VLDB Endowment, 13(11), 2627-2639

[4] Müller, R., & Bach, F. (2023). Optimizing iceberg queries in modern database systems: A comprehensive survey. ACM Computing Surveys, 56(4), 1-36.

[5] Jinuk Bae and Sukho Lee. 2000. Partitioning algorithms for the computation of average iceberg queries. In Data Warehousing and Knowledge Discovery, Springer BerLin Heidelberg, pp. 276-286.

[6] Kevin Beyer and Raghu Ramakrishnan. 1999. Bottom-up computation of sparse and iceberg cube, In ACM SIGMOD Record, vol. 28, no.2, pp. 359-370.

[7] Raghu Ramakrishnan and Gehrke Johannes. 2000. Database management systems, McGraw-Hill New York, vol. 3, pp. 1-1104.

[8] WP Yan and P Larson. 1994. Data reduction through early grouping. In Proceedings of thconference of the Centre for Advanced Studies on Collaborative research, pp .1-74.

[9] Khaled AlSabti. 2006. Efficient Computing of Iceberg Queries Using Quantiling. Journal of King Saud University-Computer and Information Sciences, vol. 18, pp. 53-75.

[10] Ricardo Baeza-Yates.1992. Information retrieval: data structures & algorithms, Prentice Hall.

[11] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases, AI magazine, vol. 17, no.3, pp. 1-37.

[12] Israel Spiegler and Rafi Maayan. 1985. Storage and retrieval considerations of binary databases, Information processing and management: an international journal, vol.21, no. 3, pp. 233-254.

[13] Vuppu Shankar, and CV Guru Rao. 2013. Computing iceberg queries efficiently using bitmap index positions. In Human Computer Interactions (ICHCI), IEEE International Conference on, pp. 1-6.

[14] PE O'Neil. 1989. Model 204 architecture and performance. In High Performance Transaction Systems. Springer Berlin Heidelberg, vol. 359, pp. 39-59.

[15] Prakash, Kale Sarika, and PM Joe Prathap. 2015. Bitmap Indexing a Suitable Approach for Data Warehouse Design. International Journal on Recent and Innovation Trends in Computing and Communication, vol.3, no. 2, pp.680-683.

[16] Uma Pavan Kumar Kethavarapu and B. Lakshma Reddy. 2014. Data Warehousing Security Encapsulation with Bitmap Indexing Mechanisms. International Journal of Emerging Technology in Computer Science & Electronics, vol.1, no. 11, pp.10-13.

[17] Vuppu Shankar and CV Guru Rao. 2014. Cache based evaluation of iceberg queries. In International Conference on Computer and Communications Technologies (ICCCT), IEEE, pp. 1-5.

[18] M. Laxmaiah, K.Sunil Kumar, A. Govardhan, and C.Sunil Kumar. 2013. A Priority Queue Approach to Evaluate Aggregate Queries Efficiently. WAIMS (World Academy of Informatics and Management Sciences), vol. 2, no. 3, pp. 2278-1315.

[19] He B, Hsiao HI, Liu Z, L. Huang Y, and Chen Y. 2012. Efficient Iceberg Query Evaluation Using Compressed Bitmap Index. Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no.9, pp. 1570-1583.

[20] V Chandra Shekhar Rao, and P. Sammulal. 2014. Efficient iceberg query evaluation using set representation. India Conference (Annual IEEE (2014), pp.1-5.

[21] Marcus Jurgerns, Hans-J. Lenz. 2001. Tree based indexes versus bitmap indexes: A performance study. International Journal of Coopertive Information Systems, vol. 10, no. 3, pp.355-376.

[22] Kesheng Wu, Ekow Otoo and Arie Shoshani. 2004. On the Performance of Bitmap Indices For High Cardinality Attributes. VLDB, pp. 24–35.

[23] Vuppu Shankar and CV Guru Rao. 2013. A Density based Priority Queue Strategy to Evaluate Iceberg Queries Efficiently using Compressed Bitmap Indices. International Journal of Computer Applications, vol. 67, no. 21, pp. 39-44.

[24] Uma Pavan Kumar Kethavarapu, Dr.Lakshma Rddy Bhavanam, and Sreedevi.S. Erady. 2015. Improvement of query processing speed in Data warehousing with the usage of components-Bitmap Indexing, Iceberg and Uncertain data. International Conference on Current Trends in Advanced Computing (ICCTAC-), pp. 1-5.

[25] Ke Wang, Yuelong Jiang, Jeffrey Xu Yu, Guozhu Dong, and Jiawei Han.2003. Pushing aggregate constraints by divide-and-approximate. In Data Engineering, Proceedings. 19 th, IEEE, International Conference, pp. 291-302.

[26] Ke Wang, Yuelong Jiang, Jeffrey Xu Yu, Guozhu Dong, and Jiawei Han. 2005. Divide-and-approximate: A novel constraint push strategy for iceberg cube mining. IEEE Transactions on Knowledge and Data Engineering, vol.17, no.3, pp.354-368.

[27] M. Laxmaiah, K. Sunil Kumar, Dr. A. Govardhan, and Dr. C. Sunil Kumar. 2013. An Approach to Evaluate Aggregate Queries efficiently using Priority Queue. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 2, no. 3, pp.341-344.

[28] Every Day Big Data Statistics [online] Available: www.vcloudnews.com/every-day-big-data-statistics.

[29] Ying Mei, Kaifan Ji, and Feng Wang. 2013. A Survey on Bitmap Index Technologies for Large-Scale Data Retrieval. In 2013 6th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), pp. 316-319.

[30] Elizabeth O'Neil, Patrick O'Neil, and Kesheng Wu. 2007. Bitmap index design choices and their performance implications. In Database Engineering and Applications Symposium, 11th International, pp. 72-84.

[31] Raymond T. Ng, Alan Wagner, and Yu Yin. 2001. Iceberg-cube computation with PC clusters. In ACM SIGMOD Record, vol.30, no.2, pp. 25-36.

[32] Yossi Matias and Eran Segaly. 1998. Partitioning based algorithms for approximate and exact Iceberg Queries, pp. 1-30.

[33] Pgukkuo B. Gibbons and Yossi Matias. 1998. New sampling-based summary statistics for improving approximate query answers. In ACM (1998), vol. 27, no. 2, pp. 331-342.

[34] Pei Lee, Lajs V.S. Lakshmanan, and Evangelos Milios. 2014. CAST: A Context-Aware Story-Teller for Streaming Social Content. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 789-798.

[35] Kyu-Yyu Whang, Brad T. Vander-Zanden, and Howard M. Taylor. 1990. A linear-time probabilistic counting algorithm for database applications. ACM Transactions on Database Systems (TODS), vol. 15, no.2, pp. 208-229.

Enhancing Supply Chain Transparency and Efficiency Through Innovative Blockchain Solutions for Optimal Operations Management

Shamrao Parashram Ghodake[1], Vishal M. Tidake[2], Sanjit Singh[3], Elangovan Muniyandy[4], Mohit[5],
Lakshmana Phaneendra Maguluri[6], John T Mesia Dhas[7]

Assistant Professor, Department of MBA, Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[1]
Associate Professor, Department of MBA, Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[2]
Assistant Professor, Department of MBA, Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[3]
Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,
Chennai, India[4]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]
Research Scholar, Institute of Management Studies and Research, Maharshi Dayanand University, Rohtak, India[5]
Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India[6]
Associate Professor, Department of Computer Science and Engineering, School of Computing,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Vel Nagar, Chennai, Tamil Nadu, India[7]

***Abstract*—Blockchain technology holds the potential to revolutionize supply chain management by ensuring transparency, efficiency, and security. This paper presents a detailed examination of blockchain's implementation in supply chain systems, focusing on safeguarding confidential information and preserving supply chain integrity. The method involves extracting 'sales order' data from Walmart's transactional database, which is then encrypted using AES algorithms to protect sensitive details such as client names and geographical information. Utilizing Ethereum's decentralized architecture, smart contracts are employed to manage transactions, encryption, decryption, and access rights. The Ethereum P2P network also aids in data validation and asset preservation, enhancing the system's reliability. Comparative analysis shows that the proposed encryption method, with encryption and decryption times of 2.8 and 3.2 seconds, outperforms traditional methods like RSA and ABE. Implemented in Python, this blockchain-based technique offers a robust, nearly infallible solution that can be applied to various supply chain practices, including Asset Management (AM), Enterprise Asset Management (EAM), and Supply Chain Management (SCM), addressing contemporary challenges and enhancing operational efficiency.**

Keywords*—Blockchain; supply chain management; advanced encryption standard; Ethereum blockchain; data storage

I. Introduction

Fundamentally, a blockchain is a distributed ledger system that statistics transactions on a network of linked nodes in a secure and permanent manner [1]. Blockchain networks function in a decentralized fashion, with every member maintaining a replica of the ledger, in comparison to normal centralized databases, in which facts garage and validation are controlled by using a unmarried authority. The decentralized layout minimizes the possibility of facts modification or illegal get admission to, removes unmarried points of failure, and maintains transparency and robustness. Furthermore, tamper-obvious and immutable, transactions registered on a blockchain are timestamped and cryptographically related [2]. The capability of blockchain technology to allow transactions among individuals without the requirement for middlemen like banks or financial groups is certainly one of its number one characteristic. Blockchain networks offer automatic and accept as true with less transactions between events through the use of smart contracts, which can be agreements that execute themselves with predetermined guidelines and conditions [3]. When certain circumstances are happy, smart contracts run routinely, simplifying operations, reducing prices, and eliminating the want for middlemen. This function of blockchain era has considerable ramifications for sectors like banking, actual property, healthcare, and deliver chain management, where safe and powerful peer-to-peer transactions are vital [4].

Blockchain generation probably clear up lengthy-status problems with transparency, duty, and self-associated with global deliver chains, it has attracted a whole lot of interest in the area of deliver chain management. Blockchain era basically affords a dispersed network of participants with a decentralized, unchangeable ledger to file transactions. This dispensed ledger makes it viable to file all transactions, transportation of commodities, and adjustments in possession in the supply chain ecosystem in a transparent and secure way, that's essential for deliver chain management [5]. Blockchain makes it feasible for all authorized events to access a unmarried source of reality, making an allowance for instant insight into the movement of commodities from the procurement of raw materials to the final customer [6]. The potential of blockchain generation to improve traceability and transparency in supply chain management is one in every of its most important advantages. Businesses may additionally gain a thorough know-how in their deliver chains, which includes the sources of substances required for the method for manufacturing, and the motion of products thru exclusive phases of manufacturing and distribution, through documenting each transaction on a blockchain [7]. Companies that exhibit this diploma of accessibility are better geared up to come across inconsistencies and inefficiencies in addition to react directly to interruptions like recalls of products or first-rate issues. Furthermore, customers can also have by no means-before-seen transparency into the origination and authenticity of merchandise way to blockchain-primarily based deliver chain solutions, for you to boost their self-assurance in corporations [8].

The potential of blockchain era to reduce the opportunity of deception, imitations, and illegal statistics revisions is every other essential benefit for supply chain control. Blockchain statistics are intrinsically tamper-glaring because of their crypto graphical linkage and immutability, which makes it almost tough for malicious actors to change or manipulate transaction statistics covertly. This characteristic of blockchain era is especially useful in sectors like medicines, expensive items, and hospitality, in which product integrity and authenticity are crucial [9]. Additionally, deliver chain techniques is probably automated and streamlined using blockchain technology, increasing productivity and decreasing fees. Automating procedures like processing payments, agreement regulation, and inventory control is possible with clever contracts that are self-executing agreements with predetermined phrases and situations. Blockchain-primarily based clever contracts can decrease transaction prices, remove mistakes, and quicken business enterprise operations by means of getting rid of the want for middlemen and human interaction [10].

There are several blessings to integrating encryption strategies with blockchain generation in supply chain control, from improving facts protection and confidentiality to maintaining the validity and integrity of transaction records [10]. Sensitive statistics, like product specs, fee facts, and patron names, is saved on the blockchain and requires encryption to be secure. Businesses can also enhance the complete safety postures of the deliver chain atmosphere through stopping unwanted get admission to and information breaches through encrypting information before it's far stored at the blockchain [11]. Maintaining confidentiality of records is one of the essential advantages of using encryption strategies in blockchain-based totally deliver chain management [12]. The increasing frequency of statistics breaches and privacy issues in modern digital environment has made it important for corporations running in deliver chains to shield sensitive data. Businesses can use encryption to obscure sensitive data, together with patron names, addresses, and financials, in order that out of doors parties cannot decipher it. This no longer only promotes consider and self-assurance amongst customers, along with purchasers, companions, and regulators, however also aids in complying with information privateness rules like the CCPA and GDPR [13].

Furthermore, by way of protecting records from undesirable changes or tampering efforts, encryption improves the authenticity and integrity of statistics recorded at the blockchain. Each transaction document is given a man or woman fingerprint or virtual signature through cryptographic hashing strategies like SHA-256. These are subsequently encrypted and recorded at the blockchain [14]. These cryptographic signatures assure that any modifications to the records would be fast discovered and characteristic as unchangeable proof of the transaction's legitimacy. Consequently, deliver chain answers primarily based on blockchain and bolstered via encryption methods provide data which are auditable and proof against manipulation, selling accountability and transparency across the deliver chain. Reducing the chance of fraud and counterfeiting is a first-rate gain of incorporating encryption into blockchain-based totally deliver chain management [15]. For each object within the deliver chain, businesses may additionally produce virtual fingerprints which can be verifiable and impervious to tampering via encrypting product identifiers like serial numbers, QR codes, or RFID tags. Stakeholders may also then safely log these encrypted identities at the blockchain, allowing them to affirm the legitimacy and beginning of goods at each step of the supply chain technique. This reduces the danger of fraud and illegal diversion while also assisting groups in monitoring and tracing gadgets with unmatched precision, which in flip enables save you from the boom of counterfeit goods [16].

The key contributions of the proposed system are given as follows:

- The paper sets out the process of implementing blockchain into the supply chain to enhance the processes and offer customers more transparent and coherent data based on Ethereum.
- The sales order data is extracted from Walmart’s transaction history in a methodical manner to construct an extensive data set that is then integrated with blockchain.
- The study uses AES to promote the highest levels of data security; more specifically, the research aims at protecting client names and geographic location details kept in the blockchain network.
- The paper explains how the smart contracts are being used in the proposed SUFS system to manage the transactions in simpler manner, encryption and decryption process and the enforcement of some strict operations and controls.
- The above solution is implemented in Python language with an aim of having a strong, reliable and effective blockchain supply chain management ecosystem that meets supply chain operation in the current world.

The following portions of the chapter are organized as follows. Section II includes an overview of the literature on blockchain technology in supply chain management. The problem statement for the study is presented in Section III. Section IV covers the recommended approach for blockchain technology in supply chain management. Section V compares the method's efficacy to previous techniques, and the performance measures are displayed, along with an explanation of the results. Section VI describes the conclusion.

II. Related Work

Although the food supply chain is one of the areas that blockchain technology has the potential to revolutionize, its deployment is fraught with difficulties. The purpose of this study is to determine and examine the obstacles preventing blockchain from being implemented in food supply chains. After a comprehensive analysis of the literature, 16 main hurdles are found and, with the help of specialists, are further divided into four groups. Then, these obstacles are ranked using the best-worst technique. The results show that organizational and technological limitations are major hindrances to the application of blockchain technology in the food supply chain. To tackle these obstacles, supply chain cooperation must be promoted, blockchain technology must be developed through development and research, and technical proficiency must be increased. However, there are drawbacks, such as the potential for bias in expert judgment and the tendency to overlook specific obstacles throughout the literature review procedure. Despite these drawbacks, the research advances knowledge of how blockchain is being applied in the supply chain and provides policymakers with recommendations on how to remove obstacles and ensure its adoption, especially in emerging economies [17].

The findings advocate that customers apprehend blockchain traceability as especially treasured whilst managing neighbourhood meals assets, leading to extended agree with and pleasant attitudes toward sharing stories [18]. However, on the deliver aspect, regardless of spotting the blessings of blockchain adoption, along with advanced accept as true with, suppliers express reluctance due to inner organizational demanding situations and issues concerning records sharing. While the study provides valuable insights from both consumer and provider perspectives, it's difficult to identify the limitations. These encompass potential biases in player responses and the scope of the examination, which won't absolutely capture all complexities related to blockchain integration in the meals supply chain.

Through a literature assessment and grey Delphi approach, ten CSFs have been identified and analysed, presenting treasured insights for FSC stakeholders [19]. However, it is vital to acknowledge certain barriers on this study. Firstly, the scope of the observation won't encompass all possible CSFs relevant to each FSC context, doubtlessly restricting the generalizability of the findings. Additionally, the gray Delphi method is based on professional critiques, which can also introduce biases or forget about certain perspectives. Furthermore, the gray DEMATEL evaluation affords insights into the significance and causal relationships amongst CSFs however might not capture all nuances of complicated interactions in the FSC ecosystem. However, similar research and practical implementation are had to completely apprehend and address the challenges related to B-IoT adoption in diverse FSC settings [19].

III. Problem Statement

The limitations encompass capability oversights in identifying obstacles, biases in participant responses, and scope constraints that won't absolutely capture the complexities of blockchain integration [17]. To address these obstacles and contribute to ongoing discussions, our look at goals to endorse a novel method for reinforcing traceability and transparency inside the meals deliver chain the usage of blockchain generation. Utilising a mixed-approach method that blends qualitative information and experiments, the aim is to better recognize consumer sentiments on blockchain-enabled traceability and pinpoint manageable answers for providers to recover from adoption hurdles. The recommended technique in the long run seeks to shut the space among theoretical knowledge and real-international software, establishing up the possibility for a supply chain that adopts blockchain era more efficaciously and sustainably.

The limitations of the literatures are given in Table I.

TABLE I. LIMITATIONS OF LITERATURE

Source	Limitation	How to Overcome
[17]	Expert judgment may be biased, and experts may fail to identify certain challenges because they do not think of them as significant barriers to success.	To minimize bias and omissions, employ pluralism of voices from various experts and compare conclusions to current research, implementing empirical evidence.
[18]	Convincing of suppliers may be hampered by internal organizational issues and inclusive of data sharing.	There is a need to develop and ensure the sharing of data through blockchain with the goal of rebuilding trust in sharing data between organizations.
[19]	There are biases in Gray Delphi method and the relationships between CSFs can be complex, detailed, subtleties may not be identified.	Number different studies and use both qualitative and quantitative research to focus on a broader range of factors affecting FSC.

IV. Proposed Blockchain Integration in Supply Chain Management

The technique employed in this examine initiates with the meticulous collection of critical sales order information sourced from Walmart's tremendous transaction facts, encapsulating pivotal details along with order ID, dates, customer identity, and complete product statistics. Following this initial section, robust information encryption strategies, specially leveraging the AES algorithm, are judiciously applied to make stronger the safety of touchy facts which includes customer identities and geographical records, thereby safeguarding privateness and confidentiality in the course of the complete deliver chain manner. Subsequent to the encryption system, a meticulous crafting of blockchain architecture ensues, harnessing the robust skills of the Ethereum blockchain renowned for its decentralized framework and sensible agreement functionalities, thereby fortifying the transparency and resilience of the supply chain infrastructure. Integral to this system is the strategic improvement of smart contracts designed to orchestrate seamless transactions, data encryption/decryption methods, and stringent access manipulate mechanisms in the blockchain community, thereby imposing predefined policies and authorizations pivotal for ensuring supply chain integrity. The implementation phase entails the configuration of nodes meticulously, fostering an environment conducive to strong facts garage, validation, and get admission to manipulate, capitalizing on the inherently fault-tolerant and resilient structure inherent within Ethereum's peer-to-peer community infrastructure. Post-integration, the encrypted sales order data seamlessly will become part of the blockchain network fabric, with each transaction meticulously tracked and securely saved throughout distributed nodes, as a consequence ensuring redundancy, records integrity, and utmost confidentiality paramount to the sanctity of deliver chain operations. The incorporation of rigorous access control mechanisms and stringent authentication protocols further fortifies the safety and privateness paradigm, ensuring that simplest duly legal employees are endowed with decryption keys or privileged get entry to, as a result fostering a fortified ecosystem bolstered via current blockchain era. Fig. 1 shows the Overall Architecture of the Proposed System.

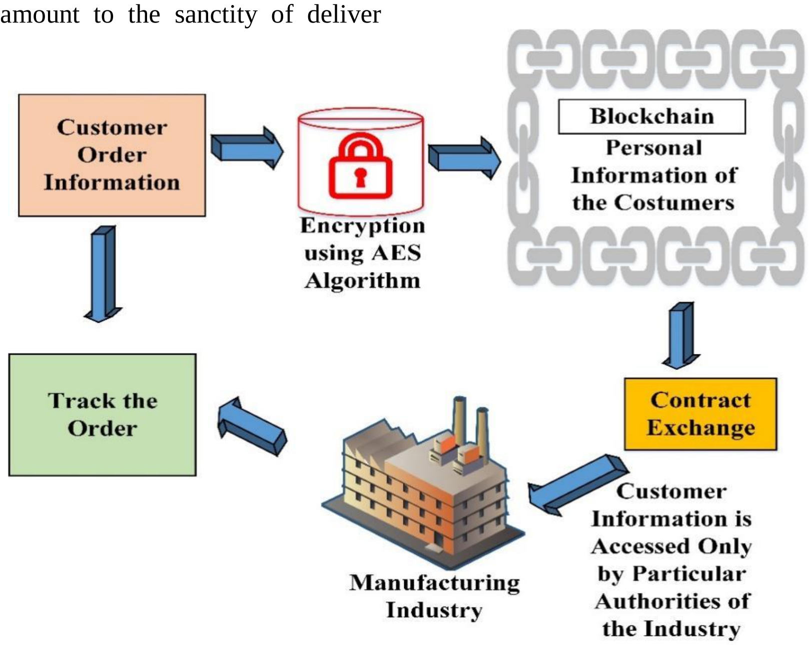

Customer Order Information
Encryption using AES Algorithm
Blockchain
Personal Information of the Costumers
Contract Exchange
Customer Information is Accessed Only by Particular Authorities of the Industry
Manufacturing Industry
Track the Order

Fig. 1. Overall architecture of the proposed system.

A. Data Collection

The dataset being shared includes giant income order statistics that were extracted from Walmart's extensive transaction statistics. It consists of all of the vital statistics, which include the order ID, the dates of the order and shipping, the consumer's identity, geographical data (United States, city, and nation), and specified product facts (call, type). Every access inside the dataset corresponds to a distinct income transaction, imparting a wealth of facts this is crucial for interpreting the complex dynamics of the supply chain and purchaser interactions within the retail environment. Furthermore, this observation targets to shed mild on the interesting opportunities of blockchain era in enhancing accountability, effectiveness, and safety all through the numerous delivery chain tiers via the prism of Walmart's transactional facts [20].

B. Data Encryption with AES Algorithm

In the context of deliver chain management, information encryption is critical for making certain the safety and privacy of personal records. AES uses a symmetrical key for decryption in addition to encryption, working with constant-length facts blocks. Blocks of plaintext statistics, typically 128 bits in size, are fed into the AES algorithm and converted thru a chain of encryption rounds. In order to efficiently obscure the original information, these rounds entail crucial enlargement, alternative, diversifications, and mixing operations achieved in a selected order. The initiation of the encryption key, which controls how plaintext blocks are converted into ciphertext, starts offevolved the encryption system. Every encryption spherical makes use of a different key thanks to the key growth process, which turns the initial encryption key into a series of spherical keys.

Sensitive facts fields in our dataset, along with consumer names and region records, are encrypted earlier than being recorded inside the blockchain. For example, the AES technique is used to transform the customer's name area, that's represented as a string of letters, into ciphertext. In a comparable vein, location facts that includes the nation, city, and country is encrypted to shield against manipulation or illegal get admission to. The observation makes certain that only those with authorization possessing the decryption key might also decode the encrypted records by way of encrypting those critical regions. Depending on the required level of security, a random encryption key with a period of 128, 192, or 256 bits is generated as a part of the encryption method. The plaintext statistics blocks are in the end encrypted the use of this encryption key and the AES method. The resultant ciphertext, which incorporates place statistics and encrypted consumer names, is then accurately saved on the blockchain, defensive personal records from malevolent use or illegal get admission to.

C. Implementation of Blockchain Design

A thorough technique is required whilst designing and deploying a blockchain infrastructure which addresses the need to store encrypted customer records and income order records. The look at chooses the Ethereum blockchain for this have a look at because of its adaptability, balance, and wealthy clever agreement capabilities. Each of the interconnecting blocks that make up our blockchain shape consists of encrypted sales order statistics alongside associated records. Transparency, immutability, and decentralization are upheld through the shape, ensuring the security and integrity of records this is saved. Within the blockchain network, clever contracts are essential to the coordination of transactions, statistics encryption/decryption processes, and get entry to control structures. By automating the enforcement of present norms and regulations, these self-executing contracts reduce the want for 542 and improve the productiveness of operations. Smart contracts are cautiously crafted within our Ethereum-based blockchain to govern every side of the supply chain management technique, which includes order processing, privateness protection, and get entry to control. Smart contracts provide for the steady garage of encrypted data, the validation of transactions, and the enforcement of access privileges in accordance with pre-installed tips and authorizations.

The setup of nodes for storage of records, confirmation, and get entry to manipulate is an essential step in the blockchain community implementation technique. Because of Ethereum's decentralized structure, fault tolerance and resilience are extended as coordinated variations of the blockchain are maintained through nodes at some stage in the network. Nodes are in charge of distributing sparkling blocks around the community, carrying out smart contracts, and verifying transactions. Nodes reach a consensus on the legitimacy of transactions and the inclusion of latest blocks to the blockchain using consensus strategies like Proof of Work (PoW) or Proof of Stake (PoS). The blockchain network's typical security is progressed and the opportunity of remoted factors of failure is reduced in step with this disbursed consensus approach. Robust cryptographic techniques, like AES, are used to first encrypt patron facts and sales order information. After that, the generated ciphertext is blanketed into transaction payloads and despatched to the Ethereum community to be blanketed in blocks. In order to assure that best the ones people with the essential decryption keys may get right of entry to and decode the encrypted information, smart contracts are in charge of approving and finishing these transactions. Furthermore, clever contracts' integrated get entry to manage mechanisms enforce rights and permissions, prohibiting unauthorized events from gaining access to or altering touchy statistics. Fig. 2 shows the Blockchain Implementation Architecture.

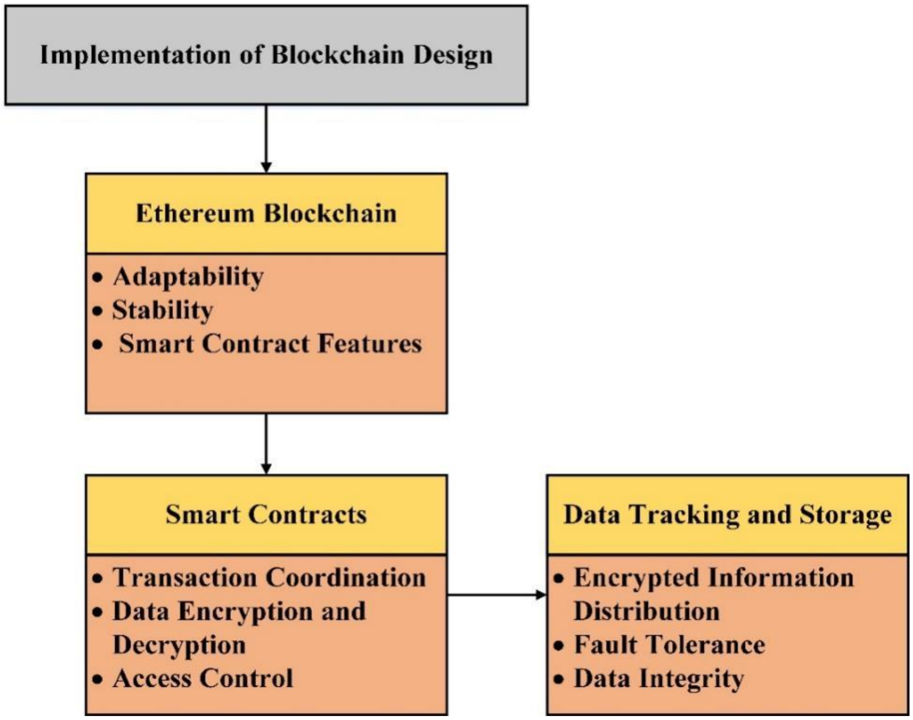

Implementation of Blockchain Design
Ethereum Blockchain
• Adaptability
• Stability
• Smart Contract Features
Smart Contracts
• Transaction Coordination
• Data Encryption and Decryption
• Access Control
Data Tracking and Storage
• Encrypted Information Distribution
• Fault Tolerance
• Data Integrity

Fig. 2. Blockchain implementation architecture.

The structure of the blockchain network makes positive that encrypted data is dispersed throughout several nodes, improving fault tolerance and redundancy. Every node keeps an encrypted replica of the blockchain, which makes retrieval of records and validation less difficult. Data integrity is maintained the usage of cryptographic hashing techniques, which permit nodes to verify the consistency of information stored. Moreover, the blockchain's immutability guarantees that encrypted statistics cannot be changed or tampered with as soon as it's miles saved, making sure the integrity and validity of client and sales order facts. Implementing Ethereum-based clever contracts involves setting up positive capabilities and common sense inside the agreement code to handle information encryption and decryption methods. Encrypted statistics and decryption keys from structures or users with permission are despatched to clever contracts as enter parameters. By the usage of decryption methods to free up the facts that has been encrypted, those contracts make certain that best individuals with permission may view the plaintext information. Smart agreement-included get right of entry to control techniques put into effect authentication and authorization requirements, limiting unwanted usage of confidential facts. Solidity, the Turing-entire programming language utilized by Ethereum, lets in developers to include complex encryption and decryption good judgment into smart contracts, ensuring robust protection protocols for the duration of the blockchain network.

Ethereum has a peer-to-peer community design wherein nodes are configured for records storage and validation, with every node keeping an archive of the blockchain ledger. Nodes ensure that everybody at the community is in settlement on the blockchain's contemporary popularity through validating transactions and carrying out smart contracts. Nodes can upload extra blocks to the blockchain with the aid of together deciding on the authenticity of transactions the use of strategies like PoW or PoS. The blockchain community's protection and resilience are improved by way of this decentralized validation manner, which reduces the opportunity of malicious attempts or single points of failure. The blockchain community's get entry to manage structures impose authentication and authorization requirements to control get admission to encrypted statistics. Access manipulates common sense incorporated in smart contracts establishes roles, credentials, and verification protocols, making sure that touchy facts might also handiest be accessed by legal individuals or systems. Cryptographic signatures, digital credentials, and multi-factor authentication protocols are examples of authentication structures that provide strong safety against undesirable entry. The blockchain network strictly controls access to records by enforcing access control policies within smart contracts, enhancing security and privacy.

D. Data Tracking and Storage

For income order records to be integrated with the blockchain community, the blockchain platform and the cutting-edge facts assets need to create a continuing interplay. Before being despatched to the blockchain network, the sales order data which incorporates the order ID, order date, shipping date, consumer records, product facts, and sales metrics is encrypted making use of sturdy cryptographic techniques like AES. Every sales order transaction is precisely documented at the blockchain as a brand-new block, with the encrypted statistics payloads appropriately saved interior. By the usage of this approach, the blockchain operates as an unchangeable ledger, making it viable to follow income order transactions transparently and without interference throughout the supply chain. Within the blockchain community, the encrypted purchase order facts is correctly stored throughout dispersed nodes, guaranteeing statistics integrity, redundancy, and secrecy. Because each node in the community incorporates an exact replica of the blockchain ledger, tolerance to failure and resilience are increased. Nodes use consensus techniques to determine among themselves whether or not additional transactions and blocks are valid, ensuring that simplest encrypted and authenticated statistics gets uploaded to the blockchain. This distributed storage layout spreads the encrypted statistics over several nodes, decreasing the opportunity of data loss or modification by using hostile parties. The observe guarantees the integrity and protection of sales order records on the safe, decentralized blockchain community by using following nice practices for blockchain facts garage and encryption.

To conclude, it is suggested that blockchain also has several more roles for areas other than transactional security in supply chain management. In supply chain management, blockchain can offer total visibility, which will help to track goods throughout the supply chain in real time and minimize instances of fraud. It can also make the stock control more efficient as it can provide an uninterruputed and secure record of the stock level and any movements of products thus minimizing on errors, excessive stocking, and running out of stock. In procurement, smart contracts can enable automation of the procurement processes by ordering more stocks and or restocking whenever certain conditions are met. Also in the vendor relationships blockchain technology underlines the improved trust and cooperation through recording all the interactions, contracts, and payments in the ledger avoiding conflicts. In the same manner, blockchain can also help provide an unalterable method of logging quality checks and certifications for compliance purposes to improve business relations with vendors. In conclusion, the use of blockchain can potentially enhance the functional supply chain areas by automating and enhancing the security and verification of supply chain activities.

V. Results and Discussion

In this section, the end result and discussion of the proposed model are given. The method commences with complete information collection from Walmart's transaction information, taking pictures important sales order details like order ID, dates, customer statistics, and product specifics, forming the muse for blockchain integration in supply chain control. Following data acquisition, robust encryption strategies, significantly leveraging AES, are implemented to protect sensitive data such as consumer identities and geographical information, ensuring privacy at some point in the supply chain procedure. Integral to this system is the development of smart contracts orchestrating transactions, encryption/decryption approaches, and get right of entry to controls, enforcing predefined policies to preserve supply chain integrity. Implementation involves configuring nodes to facilitate sturdy facts garage, validation, and access control inside Ethereum's fault-tolerant peer-to-peer network infrastructure. Post-integration, encrypted income order facts seamlessly integrates into the blockchain, with each transaction meticulously tracked and securely saved across dispensed nodes, ensuring redundancy, integrity, and confidentiality. The incorporation of stringent get admission to control mechanisms and authentication protocols similarly fortifies safety and privacy, proscribing get right of entry to authorized employees and bolstering the atmosphere's resilience.

TABLE II. Memory Usage

Algorithm	Memory Used (mb)
ABE [21]	0.107
RSA [22]	0.186
Proposed AES	0.0088

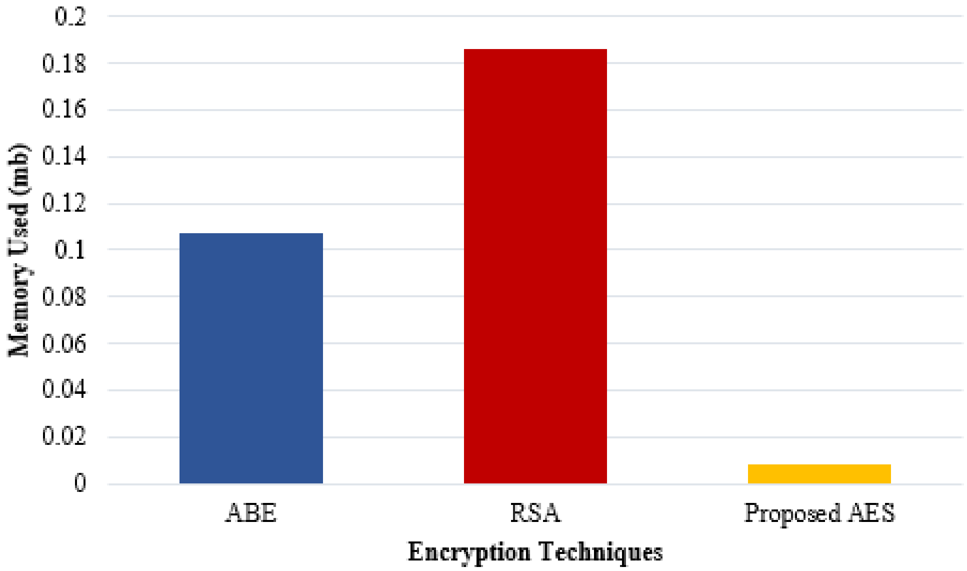


Fig. 3. Memory usage.

Table II and Fig. 3 presents a comparative evaluation of memory area consumption for extraordinary encryption algorithms, such as Attribute-Based Encryption (ABE), RSA (Rivest-Shamir-Adleman), and the proposed AES (Advanced Encryption Standard). The reminiscence utilization, measured in megabytes (mb), is indexed for each set of rules, showcasing their respective performance in phrases of memory intake. ABE, which is known for its versatility in getting right of entry to control mechanisms based on attributes, utilizes zero.107 mb of memory, indicating a mild degree of memory usage. In assessment, RSA, an extensively used uneven encryption set of rules, reveals a higher reminiscence footprint of 0.186 mb, suggesting pretty better memory requirements as compared to ABE. Notably, the proposed AES algorithm, regarded for its efficiency, simplicity, and strong encryption competencies, demonstrates a drastic decrease in memory usage at 0.0088 mb, making it the most memory-efficient encryption algorithm among the three.

TABLE III. Comparison of Encryption and Decryption Time with Different Encryption Algorithms

Algorithm	Encryption Time (sec)	Decryption Time (sec)
ABE [21]	7.5	5.2
RSA [22]	6.7	7.3
Proposed AES	2.8	3.2

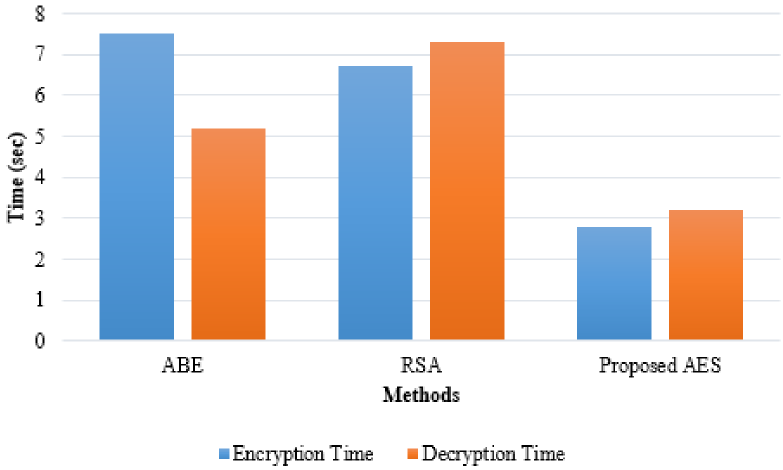


Fig. 4. Comparison of encryption and decryption time with different encryption algorithms.

Table III and Fig. 4 gives a complete evaluation of encryption and decryption instances associated with distinctive encryption algorithms, together with ABE, RSA, and the proposed AES. The Table showcases the time taken for encryption and decryption operations in seconds for every set of guidelines, offering insights into their respective efficiency in phrases of processing pace. ABE, famed for its characteristic-based get entry to manipulate mechanisms, demonstrates an encryption time of 7.5 seconds and a decryption time of 5.2 seconds, indicating slight overall performance in phrases of processing velocity. On the other hand, RSA, a standard asymmetric encryption algorithm, exhibits an encryption time of 6.7 seconds and a decryption time of 7.3 seconds, reflecting slightly faster encryption but slower decryption in comparison to ABE. Notably, the reposed AES algorithm, known for its velocity and efficiency in encryption and decryption, outperforms both ABE and RSA, with encryption and decryption instances of 2.8 seconds and 3.2 seconds, respectively. This highlights AES as the most efficient algorithm in terms of processing speed among the three.

TABLE IV. ENCRYPTION TIME VS. KEY NUMBERS OF DIFFERENT ALGORITHMS

Keys Number	ABE (ms)	RSA (ms)	Proposed AES (ms)
1	3.10	1.3	0.06
2	4.97	1.78	0.36
3	5.99	2.84	1.2
4	6.88	5.88	2.4

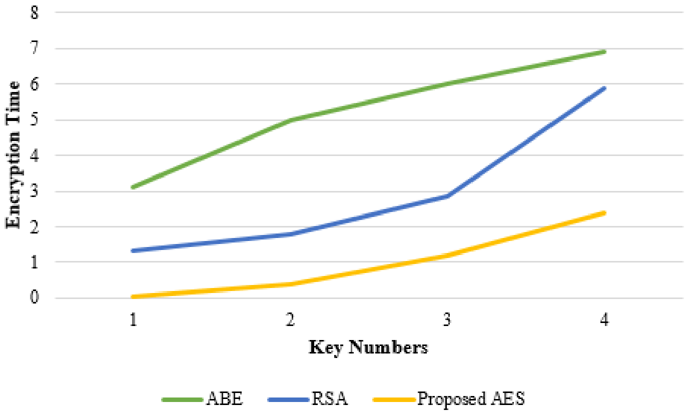


Fig. 5. Encryption time vs. key number of different algorithms.

The Table IV and Fig. 5 illustrates the encryption time in milliseconds for every set of rules primarily based on varying key numbers, ranging from 1 to 4 keys. Each row corresponds to a particular quantity of keys, even as the columns constitute the encryption time in milliseconds for ABE, RSA, and the proposed AES algorithm, respectively. For ABE, encryption time will increase progressively because the range of keys rises, indicating a linear dating between encryption time and key range. Similarly, RSA exhibits a proportional boom in encryption time with the addition of keys, although the price of growth seems barely steeper as compared to ABE. In comparison, the proposed AES algorithm demonstrates a drastic decrease in encryption instances throughout all key numbers, showcasing its performance and scalability in handling encryption responsibilities despite multiple keys.

TABLE V. COMPARISON OF BUFFER TIME WITH DIFFERENT ENCRYPTION ALGORITHMS

Algorithm	Buffer Time (mb)
ABE [21]	0.154
RSA [22]	0.165
Proposed AES	0.147

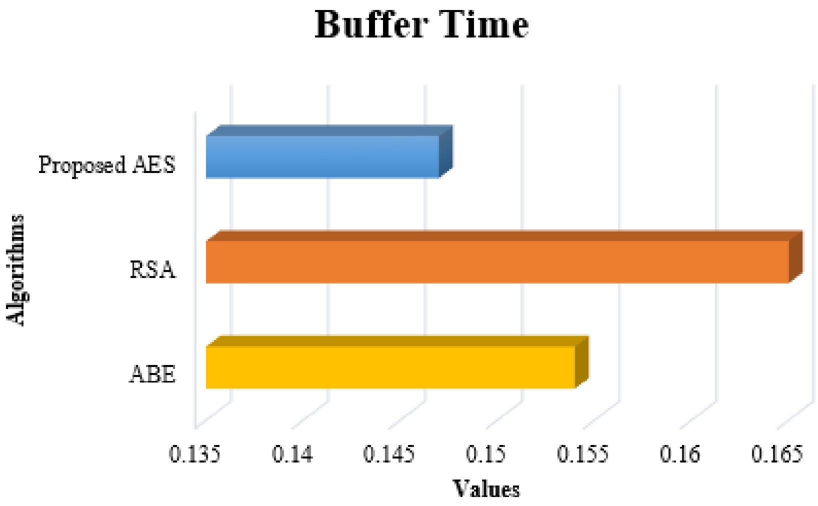


Fig. 6. Comparison of buffer time with different encryption algorithms.

Table V and Fig. 6 affords a comparative analysis of buffer time, measured in megabytes (mb), throughout exceptional encryption algorithms, including ABE, RSA, and the proposed AES. Buffer time refers to the amount of reminiscence space ate up via each algorithm in the course of encryption methods. Each row in the table represents a specific encryption algorithm, whilst the corresponding column displays the buffer time in megabytes for ABE, RSA, and the proposed AES algorithm, respectively. In comparison, ABE and RSA showcase slightly better buffer instances of 0.154 mb and 0.165 mb, respectively.

TABLE VI. SECURITY COMPARISON OF DIFFERENT ALGORITHMS

Algorithm	Security
ABE [21]	14
RSA [22]	19
Hybrid ABE-DES	22
Proposed AES	45

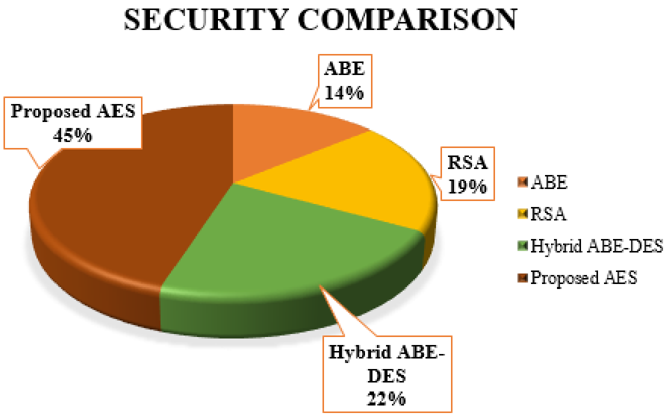


Fig. 7. Security comparison of different algorithms.

Table VI and Fig. 7 each row inside the desk represents a specific encryption set of rules, while the corresponding column presents the assigned protection score. In comparison, ABE and RSA showcase comparatively decreased safety ratings of 14 and 19, respectively. The Hybrid ABE-DES algorithm falls in between, with a security score of 22.

A. *Discussion*

The technique starts with thorough information collection from Walmart's transaction information, followed by using the implementation of sturdy encryption techniques the usage of AES to protect touchy data for the duration of the delivery chain method. Subsequently, the blockchain architecture is meticulously designed, leveraging Ethereum's decentralized framework and clever agreement talents to decorate transparency and resilience. The integration of encrypted income order facts into the blockchain network guarantees redundancy, integrity, and confidentiality, with stringent get entry to control mechanisms similarly fortifying protection and privateness.

The outcomes show that the new form of AES entropy is better than previous methods of ABE and RSA regarding the buffer time with 0.147MB being higher than 0.154MB and 0.165 MB respectively. Such a reduction in buffer time shows that there is an enhancement of efficiency in data processing, which plays a significant role in line with the real-time supply chain management systems. Although both ABE and RSA have strong encryption capacities, they are likely to result in high latency and the extra use of computational power. The above proposed AES method has the strength of delivering good encryption to meet the demands of client data and other geographic information as well as a faster encryption and decryption than other algorithms which makes it suitable for large volumes of transactions common in the supply chain networks. This improvement makes the system to be more efficient and scalable in handling the requirements of today’s supply chain processes.

The integration of blockchain in supply chain confronts a number of real-life impediments and hurdles. First, the adoption of blockchain is not easy as most of the supply chain network still operates with centralized databases which pose integration issues and cost-effectiveness. Interconnectivity between these systems and blockchain solutions such as Ethereum might entail substantial modifications on the technical level. Furthermore, the factors of the scalability and flexibility of blockchain are an issue that could be an issue, especially for the extensively large volume and global supply chains that may overwork the blockchain processing power and storage. Another issue is the potential high cost of establishing and sustaining decentralized nodes and smart contracts that may be out of reach for some businesses. It is also true that security is both a strength and a weakness where, while the use of blockchain improves the security of data stored in a network, smart contract errors or weak encryption algorithms can put the system at risk. Last, there may be reluctance by organizations to embrace blockchain due to its decentralized nature a move that may not be embraced by individuals within notable supply chain management organizations especially those who are used to the traditional means of operations.

VI. CONCLUSION AND FUTURE WORK

The use of blockchain technology in the supply chain is an innovation towards improving on the supply chain operations. This paper shows that the supply chain can be improved by implementing blockchain technology, specifically using Ethereum, where transparency in supply chain, decentralisation and smart contracts may enhance data security and supply chain functionality. The AES algorithm for encrypting information that are deemed sensitive including the clients' names and geographic details, adds high degrees of privacy and protection. In addition to that the adoption of the PayPal's peer-to-peer network enhances the transparency as well as the solidity of the entire supply chain in allowing for efficient, secure and accurate transactions while at the same time ensuring proper controlling of access. Blockchain approach also successfully solves some of the major problems like maintenance of supply chain by integrating encrypted sales order data with supply chain system. The employed encryption method demonstrates enhanced efficiency to RSA and ABE where especially the encryption and decryption time is considered. The technique that was created in Python also elevates security within supply chain functions while actively responding to the consciousness changes that are rampant within supply chain management. There is scope for further research relating to this paper as highlighted next and below: First, the additional use of other emergent trends in supply chain management like artificial intelligence, the internet of things in combination with blockchain technology might add more value to the situation and help to predict the supply chain problems better. However, it is also vital to understand whether such a solution would work in larger and more complex supply chains would be important and further research on this solution. Additional research could also work on the effects of a variety of applications of blockchain platforms and various encryption styles on effects and security. In addition, identifying and solving threat and risks factors concerning regulations and compliances related to the use of blockchain technology in different parts of the world will be crucial for its adoption. Lastly, real-life supply chain examples and pilot adoption can be used as follows in order to give practical examples of the effectiveness of using blockchain in supply chain management and to put a light on the practical issues and implementation barriers.

REFERENCES

[1] E. Purwaningsih, M. Muslikh, S. Suhaeri, and B. Basrowi, "Utilizing blockchain technology in enhancing supply chain efficiency and export performance, and its implications on the financial performance of SMEs," Uncertain Supply Chain Management, vol. 12, no. 1, pp. 449–460, 2024.

[2] F. Zhang and W. Song, "Sustainability risk assessment of blockchain adoption in sustainable supply chain: An integrated method," Computers & Industrial Engineering, vol. 171, p. 108378, 2022.

[3] Y. Liu, W. Fang, T. Feng, and M. Xi, "Blockchain technology adoption and supply chain resilience: exploring the role of transformational supply chain leadership," Supply Chain Management: An International Journal, 2024.

[4] S. Fernandez-Vazquez, R. Rosillo, D. De la Fuente, and J. Puente, "Blockchain in sustainable supply chain management: an application of the analytical hierarchical process (AHP) methodology," Business Process Management Journal, vol. 28, no. 5/6, pp. 1277–1300, 2022.

[5] S. Balasubramani, R. Dhanalakshmi, L. Kavisankar, K. Ramesh, S. Saritha, and D. Pandey, "Revolutionizing Supply Chain With Machine Learning and Blockchain Integration," in Utilization of AI Technology in Supply Chain Management, IGI Global, 2024, pp. 113–125.

[6] N. Challa, "Blockchain Integration in Supply Chain Management: A Comprehensive Analysis of B2B Implications".

[7] C. L. Tan, Z. Tei, S. F. Yeo, K.-H. Lai, A. Kumar, and L. Chung, "Nexus among blockchain visibility, supply chain integration and supply chain performance in the digital transformation era," Industrial Management & Data Systems, vol. 123, no. 1, pp. 229–252, 2023.

[8] B. Alawi, M. M. S. Al Mubarak, and A. Hamdan, "Blockchain evaluation framework for supply chain management: a decision-making approach," in Supply Chain Forum: An International Journal, Taylor & Francis, 2022, pp. 212–226.

[9] T. Gürpinar, M. Henke, and R. Ashraf, "Integrating blockchain technology in supply chain management–a process model with evidence from current implementation projects," 2024.

[10] S. Yousefi and B. M. Tosarkani, "An analytical approach for evaluating the impact of blockchain technology on sustainable supply chain performance," International Journal of Production Economics, vol. 246, p. 108429, 2022.

[11] A. Vaezi, E. Rabbani, and S. A. Yazdian, "Blockchain-integrated sustainable supplier selection and order allocation: A hybrid BWM-MULTIMOORA and bi-objective programming approach," Journal of Cleaner Production, p. 141216, 2024.

[12] V. K. Manupati, T. Schoenherr, M. Ramkumar, S. M. Wagner, S. K. Pabba, and R. Inder Raj Singh, "A blockchain-based approach for a multi-echelon sustainable supply chain," International Journal of Production Research, vol. 58, no. 7, pp. 2222–2241, 2020.

[13] R. Cole, M. Stevenson, and J. Aitken, "Blockchain technology: implications for operations and supply chain management," Supply chain management: An international journal, vol. 24, no. 4, pp. 469–483, 2019.

[14] A. Rijanto, "Blockchain technology roles to overcome accounting, accountability and assurance barriers in supply chain finance," Asian Review of Accounting, 2024.

[15] G. Blossey, J. Eisenhardt, and G. Hahn, "Blockchain technology in supply chain management: An application perspective," 2019.

[16] U. K. Suganda, H. A. Buchory, and Z. Aripin, "ACCEPTANCE OF BLOCKCHAIN TECHNOLOGY IN SUPPLY CHAIN MANAGEMENT IN INDONESIA: AN INTEGRATED MODEL FROM THE PERSPECTIVE OF SUPPLY CHAIN PROFESSIONALS FOR SUSTAINABILITY," KRIEZ ACADEMY: Journal of development and community service, vol. 1, no. 2, pp. 33–51, 2024.

[17] S. Khan, M. K. Kaushik, R. Kumar, and W. Khan, "Investigating the barriers of blockchain technology integrated food supply chain: a BWM approach," Benchmarking: An International Journal, vol. 30, no. 3, pp. 713–735, 2023.

[18] C. Cozzio, G. Viglia, L. Lemarie, and S. Cerutti, "Toward an integration of blockchain technology in the food supply chain," Journal of Business Research, vol. 162, p. 113909, 2023.

[19] R. Singh, S. Khan, J. Dsilva, and P. Centobelli, "Blockchain integrated IOT for Food Supply Chain: A grey based Delphi-DEMA$^{\text{TEL}}$ approach," Applied Sciences, vol. 13, no. 2, p. 1079, 2023.

[20] "Complete Exploratory Data Analysis of Walmart Data | Kaggle." Accessed: Mar. 23, 2024. [Online]. Available: https://www.kaggle.com/code/gautammourya/complete-exploratory-data-analysis-of-walmart-data

[21] Y. Jiang, X. Xu, and F. Xiao, "Attribute-based encryption with blockchain protection scheme for electronic health records," IEEE Transactions on Network and Service Management, vol. 19, no. 4, pp. 3884–3895, 2022.

[22] N. A. Ugochukwu, S. Goyal, A. S. Rajawat, S. M. Islam, J. He, and M. Aslam, "An innovative blockchain-based secured logistics management architecture: utilizing an RSA asymmetric encryption method," Mathematics, vol. 10, no. 24, p. 4670, 2022.

Enhanced Quantitative Financial Analysis Using CNN-LSTM Cross-Stitch Hybrid Networks for Feature Integration

Dr. Taviti Naidu Gongada[1], B. Kumar Babu[2], Janjhyam Venkata Naga Ramesh[3],
P. N. V. Syamala Rao M[4], Dr. K. Aanandha Saravanan[5], K Swetha[6], Dr Mano Ashish Tripathi[7]
Assistant Professor, Dept of Operations-GITAM School of Business, GITAM (Deemed to be) University, Visakhapatnam, India[1]
Assistant Professor, Department of CSE-SOT, GITAM (Deemed to be University) Hyderabad, India[2]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[3].
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[3].
Assistant Professor, Department of Computer Science and Engineering, SRM University, Amaravati, AP, India[4]
Associate Professor, Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India[5]
Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India[6]
Department of Humanities and Social Sciences, Motilal Nehru National Institute of Technology, Allahabad, India[7]

Abstract*—This research paper provides innovative approaches to support financial prediction, or it is a different kind of economic prediction that extends over collecting different economic information. Financial prediction is a concept that has been employed. The present study offers a unique approach to predicting finances by integrating many financial issues utilizing a cross-stitch hybrid approach. The method uses information from several financial databases, including market data, corporate reports, and macroeconomic indicators, to create a comprehensive dataset. Employing MinMax normalization the features are equally scaled to provide uniform input for the algorithm. The combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) systems form the basis of the framework. To understand the time-dependent nature of financial information, LSTM networks (long short-term memory) are utilized to record and simulate the temporal interactions and patterns. Concurrently, spatial features are extracted using CNNs; these components help identify patterns that are difficult to identify with conventional techniques. Better handling of risks, more optimal approaches to investing, and more informed decision-making are made possible by the enhanced forecasting potential that this method—which is described above—offers. Potential pilot studies will focus on innovative uses in financial decision-making and advancements in cross-stitching structure. This paper proposes a sophisticated approach that can help stakeholders, such as investors, analysts of data, and other financial intermediaries, traverse the complexities of financial markets.

Keywords*—*Cross-Stitch Hybrid Networks; predictive modelling; LSTM networks; convolutional neural networks; financial analysis

I. Introduction

The system of assessing financial markets and investment possibilities the usage of statistical and mathematical techniques is known as quantitative monetary analysis [1]. Stochastic models take uncertainty and unpredictability into attention, whereas econometric models—which include time collection models and regression analysis—assist in determining correlations among monetary variables [2]. Complex patterns in statistics are captured via gadgets gaining knowledge of and deep getting to know strategies like LSTM networks and Convolutional Neural Networks. Predictive overall performance is advanced when gadget mastering algorithms are combined with traditional econometric strategies [3].

Applications consist of portfolio optimization to select the top-rated asset blend, credit scoring to decide creditworthiness, and algorithmic buying and selling, which makes use of quantitative techniques for automated trading techniques [4]. Among the blessings of quantitative evaluation are expanded hazard minimization, objectivity, efficiency, and accuracy. But there are boundaries to conquer, such as the intricacy of the facts, overfitting of the model, and the requirement to regulate changing markets. As well known, the quantitative economic analysis combines economics, mathematics, and information technology to enhance market forecasting, risk management, and funding picks.

DL techniques are very beneficial for financial facts analysis due to the fact they can model noisy, complicated, and high-dimensional time-series statistics, particularly RNNs [5]. RNNs, in comparison to traditional neural networks, are ready with recurrent connections that allow them to leverage the version's memory feature to seize temporal relationships via remembering and making use of earlier facts to modern-day output computations. To introduce nonlinearity and capture complex patterns, input variables in RNNs are mixed with matching weights and a bias time period, exceeded via an aggregate feature, after which processed via a nonlinear activation characteristic [6]. RNNs can control the sequential structure of economic facts due to the fact to this layout, which makes them suitable for tasks like price forecasting, trend evaluation, and volatility prediction.

In financial analysis, the CNN-LSTM Cross-Stitch is an effective aggregate that combines the temporal modelling skills of LSTM networks with the spatial characteristic extraction competencies of CNNs [7]. By identifying localized features, these layers correctly lower the dimensionality of the incoming information at the same time as emphasizing vital components. These spatial factors are then processed via the LSTM thing, which then information the temporal linkages and lengthy-term dependencies present in sequential monetary records. The LSTM network is an effective device for modelling sequential statistics by using reminiscence cells and gating mechanisms. The CNN-LSTM Cross-Stitch gives a synergistic mixture that allows for a complete approach to financial assessment.

- To improve quantitative economic research, the paper suggests a system that combines CNNs for time dependence and LSTMs for spatial feature extraction.
- It allows for a comprehensive assessment of economic performance by integrating a variety of financial data, such as macroeconomic indicators, market data, and business financial statements.
- By ensuring appropriate feature scaling through the application of Min-Max normalization, high-magnitude features are not allowed to predominate during model training.
- Through the utilization of CNNs and LSTMs' respective advantages, this structure offers an effective instrument for examining intricate financial information.
- The study's architecture efficiently captures both temporal and geographic trends in financial data, enabling precise market trend forecasting.

The paper's structure is organized as follows: Section II conducts a literature review, summarizing existing research relevant to the study's topic. Section III articulates the problem statement, addressed in the study. In Section IV, the proposed methodology is presented, outlining the approach and techniques employed to address the problem statement. Section V presents the findings of the study. Finally, Section VI concludes the paper, discussing limitations, and suggesting avenues for future research.

II. Literature Review

In order to train a BPN for predicting corporation disasters, the hybrid economic analysis model provided in this paper makes use of financial ratios as its main inputs and consists of both static and technique evaluation models [8]. The model, which gives a higher prediction charge, outperforms discriminant analysis, choice timber, and a solo BPN in the usage of four datasets of Taiwanese enterprises. The version's education on region-particular records, potential problems in generalizing to different industries or areas, issues with complexity and interpretability, the opportunity of overfitting, the reliance on fantastic historical facts for technique evaluation, and the capability omission of important external monetary factors should, nevertheless, limit the version's applicability.

The examine makes use of a hybrid simulation method that combines a multi-duration simulation model that integrates a MRPII machine in Microsoft Excel with a manufacturing surroundings in ProModel and makes use of Microsoft Visual Basic to synchronize agenda dissemination and inventory updates to investigate the damaging consequences of different accounting strategies on mentioned profits throughout lean manufacturing implementation [9]. The use of simulation techniques, which may not accurately replicate the intricacies of actual lifestyles, potential biases in the choice of accounting techniques, and the emphasis on certain inventory reduction prices which won't practice to other groups or operational contexts are a number of the drawbacks.

There are three steps to the method: what first, preprocessing and denoising the data using WT; second, prediction of deconstructed data using SVM, RNN, and NB; and third, incorporating these predictions into the GA and WA methods [10]. On lots of information types, this version performs drastically higher than benchmark techniques and unmarried strategies. However, its computational requirements and complexity may additionally restriction realistic packages and may require heavy parameter modifications Moreover, even if the model performed well on the tested data, it is not immediately sensitive to external variables affecting stock prices such as political developments or prevailing economic conditions.

The take a look at provides a sturdy framework for excessive-accuracy inventory rate forecasting based on DL of-based regression techniques [11]. It makes use of ancient inventory rate facts from an Indian agency listed at the NSE that were recorded at five-minute intervals. The framework consists of five LSTM techniques and four CNN techniques. The models are proven and examined the use of metrics for execution time and RMSE. Although exhibiting superb consequences, the approach has drawbacks inclusive of dependence on extraordinarily distinct information, extensive processing overhead, and viable non-generalizability to different stocks or market instances. Additionally, the techniques might not fully recollect outdoor variables like geopolitical events or economic news, and the evaluation metrics won't completely mirror real-global trading dynamics.

In order to dynamically compare financial dangers and strategies, the research proposes a flexible DFA model framework for the nonlife insurance marketplace [12]. It does this by means of integrating not unusual components determined in many DFA models. The framework's trustworthy implementation and amendment to precise enterprise needs are made viable by way of its mathematical description. The method isn't without flaws, although, it including the issue of combining disparate principles, which requires an excessive stage of information and processing power and can make it less available to smaller organizations. Personalization can take a whole lot of time, and relying an excessive amount on previous statistics ought to lead to biases if ancient patterns aren't dependable indicators of future conduct. Furthermore, the efficacy of the model is contingent upon the quality of the accessible data and may fail to account for certain elements pertinent to particular categories of nonlife insurance or distinct nearby market dynamics.

In order to expect time collection, the have a look at looks into BiLSTM networks and compares them to unidirectional LSTM and ARIMA techniques [13]. It is hypothesised that BiLSTMs, with its dual-directional facts processing competencies, can boom prediction accuracy by using advanced training abilities. The look at its findings guides the concept that, despite the fact that taking longer to reap equilibrium, BiLSTMs carry out more correctly than both ARIMA and unidirectional LSTM techniques. BiLSTMs perform higher, but they take longer and demand extra pc power, that could limit their use in actual-time packages. Furthermore, the study does not investigate hybrid models or other sophisticated neural networks, and its findings could not apply to all kinds of time series data. Furthermore, features of interpretability and practical implementation both crucial for real-world applications are not fully explored.

The paper's suggested approach entails a methodical examination and arrangement of the historical evolution of financial analysis, emphasizing the contributions of international experts and identifying the most important non-financial and financial aspects affecting company value. By contrasting theoretical valuations with real market results, it critically evaluates conventional valuation techniques, pointing out inconsistencies and possible weaknesses. The study makes recommendations for further research to increase these valuation methodologies' precision and dependability. The potential risk of flaws in historical literature, the simplicity of the explanation at the expense of the basic relationships among variables, and the possibility of extraneous factors not included in the computations all suggest that the approach may not be flawless. Furthermore, the changes would undoubtedly need appropriate empirical confirmation. The research is broad and might not go deep enough in certain areas, therefore it might provide thorough answers to financial assessment and valuation procedures in some areas but not in others.

The research articles under evaluation examine a range of cutting-edge approaches for anticipating financial results and tackling major issues in financial analysis and forecasting. When applied to Taiwanese companies, the first look at's hybrid monetary evaluation model which mixes monetary ratios with BPN shows higher accuracy in predicting firm screw ups than greater traditional models like discriminant evaluation and selection bushes [14]. It highlights the financial effects however has limits in phrases of version realism and generalizability.

III. Problem Statement

Current approaches have notable shortcomings that affect their capacity for generalization and economic usefulness, despite advancements in financial projections and risk forecasting. In hybridized models of economics, for example, ratios of finance and BPN combine to forecast company losses more accurately than previous models [8]. However, these frameworks have problems expanding across sectors and geographical areas, and they may overfit historical data. Correspondingly, the simulation techniques used to assess lean manufacturing and accounting methods are limited by their incapacity to accurately represent the intricacies of the actual world and may induce biases because of their exclusive focus on particular inventory reduction expenses [9]. In addition, models for dynamic forecasting that use BiLSTM networks provide better prediction accuracy but have restricted application to other time series data sources and high processing costs [13]. These drawbacks underscore the need for more adaptable, less data-dependent, and computationally efficient models.

IV. Proposed Methodology

The software of deep learning particularly people who employ neural networks has ushered in a new generation of comprehension and forecasting within the subject of financial statistics analysis. CNNs and LSTM networks have turn out to be outstanding equipment amongst those cutting-edge methodologies, especially effective at dealing with the complicated and time-established nature of monetary markets. CNNs had been remarkably a success in their version to monetary time collection evaluation, regardless of their authentic development for image popularity packages. These networks are especially good at routinely spotting capabilities and trends in statistics, together with changes in volume, rate, or different marketplace indicators.

By maintaining memory for extended periods of time, LSTM networks function at the side of CNNs to address the temporal component of economic statistics. LSTM networks, as adverse to traditional recurrent neural networks, are capable of capturing long-variety relationships which can be critical for modelling complex financial time series because they alleviate the vanishing gradient trouble. Fig. 1 shows the architecture of the proposed financial analysis. LSTM networks can perceive recurrent patterns, adjust to shifting marketplace situations, and predict future traits with awesome accuracy by getting to know from past statistics. For jobs like chance management, portfolio optimization, and market forecasting, this makes them very useful. In essence, the mixture of CNNs with LSTM networks gives hitherto unheard-of insights and prediction strength, as a result representing a paradigm alternate inside the observe of monetary information. With the increasing quantity and complexity of monetary facts, these sophisticated DL will truly turn out to be more and more crucial for comprehending and navigating the complexities of the world's monetary markets.

A. Data Collection

With statistics from economic statements, essential ratios covering profitability, leverage, and performance, as well as market records on stock expenses and alternate volumes, it presents comprehensive know-how of how nicely a firm is performing when it comes to marketplace trends. To make sure thorough evaluation and nicely-knowledgeable decision-making, users are entreated to verify information veracity, use caution whilst interpreting ratios, and seek advice from unique sources for methodological clarity [15].

B. Data Pre-Processing

A data preparation approach called Min-Max normalization, every so often referred to as Min-Max scaling, is used to scale numerical capabilities to a given range, typically among zero and 1. The following is the components for MinMax normalization:

$$x_m = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

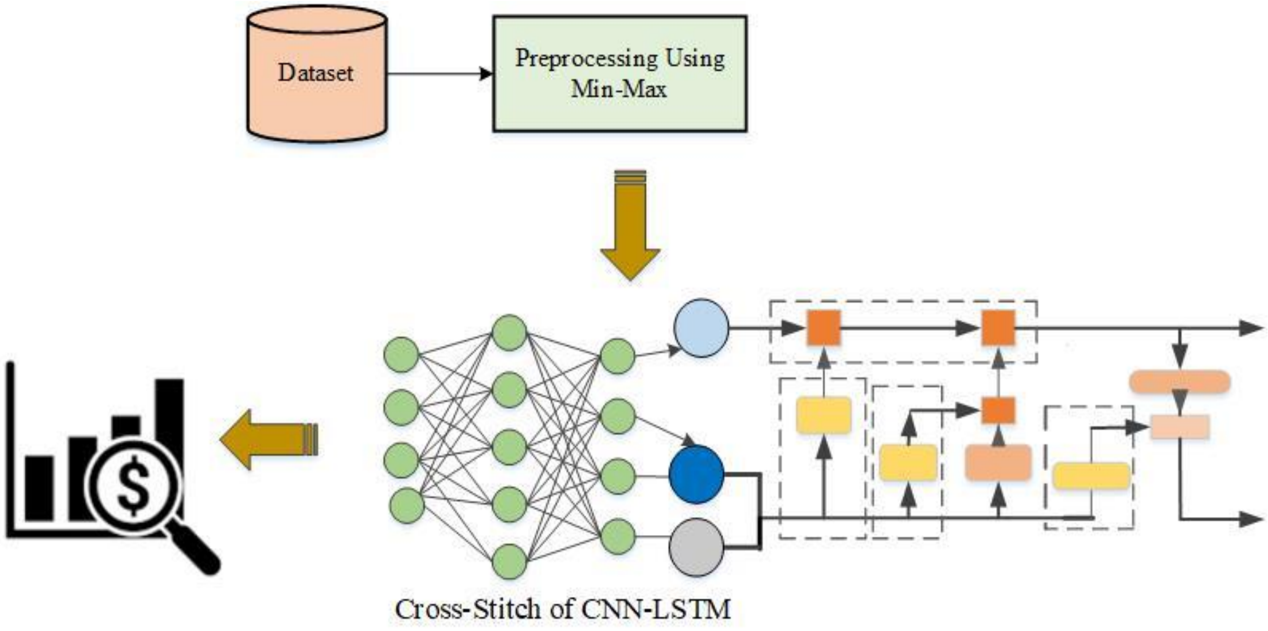


Fig. 1. The architecture of the proposed financial analysis.

Certain device ML which can be sensitive to function magnitudes may also benefit from this variation, which maintains the unique distribution of the records even as ensuring that each capability is at the same scale. Furthermore, Min-Max normalization helps avoid functions with higher scales from overwhelming people with decreased scales in the course of model training, making the facts less difficult to recognize. It's essential to remember that MinMax normalization is susceptible to outliers and could not work successfully in instances in which the records include excessive values.

1) Data cleaning: The proposed research's data cleaning procedure includes filling in the values that are missing employing approaches like imputation, which replaces missing data points with statistical metrics (such as mean, median, and mode), forward or backward fill, and linear or polynomial interpolation. Rows or columns containing an excessive amount of data that is missing may be removed if interpolation is not practical. The Z-Score, which is described in Eq. (2), and the IQR (Interquartile Range) Technique are two techniques for outlier detection that may discover and handle outliers by establishing threshold. To lessen the impact of extreme numbers, winsorization is used, and to preserve data integrity, clipping is used to cap values at a specified threshold.

$$Zscore\ Z = (X - \mu)/\sigma \tag{2}$$

Data Transformation: In data analysis, data transformation is the procedure of transforming unprocessed data into a structure that is more suited for analysis. To do this, data must be modified to increase its accuracy, uniformity, and usefulness. Data spans and patterns are adjusted by methods including scaling, standardization, and normalization, which facilitate analysis and interpretation. Transformation techniques including Box-Cox conversion, log transformation, and differencing help to eliminate patterns and stabilize dispersion. By verifying that the data satisfies the presumptions of mathematical frameworks and computations, this stage is essential for producing insights that are more reliable and precise. The Data transformation for the proposed model is explained in Eq. (3).

$$Y(\lambda) = \frac{Y^{\lambda}-1}{\lambda} \tag{3}$$

C. Utilizing CNNs for Advanced Feature Extraction

CNNs are powerful equipment for monetary information evaluation characteristic extraction, in particular in relation to identifying spatial relationships gift within the facts. Convolutional, pooling, and absolutely linked layers are most of the layers that generally make up a CNN's structure (see Fig. 2).

Convolutional Layers: These layers use a hard and fast of filters to carry out convolution operations on the enter information to be able to extract neighbourhood functions.

Pooling Layers: To lessen the dimensionality of the characteristic maps, pooling layers like max-pooling are used after the convolutional layers.

Fully connected Layers: After the feature maps are pooled, completely connected layers combine the functions right into a single, all-encompassing picture. By clarifying the connections between exceptional extracted factors, this included illustration makes it easier to perform complex analysis activities and helps to offer a deeper knowledge of the economic records this is being examined.

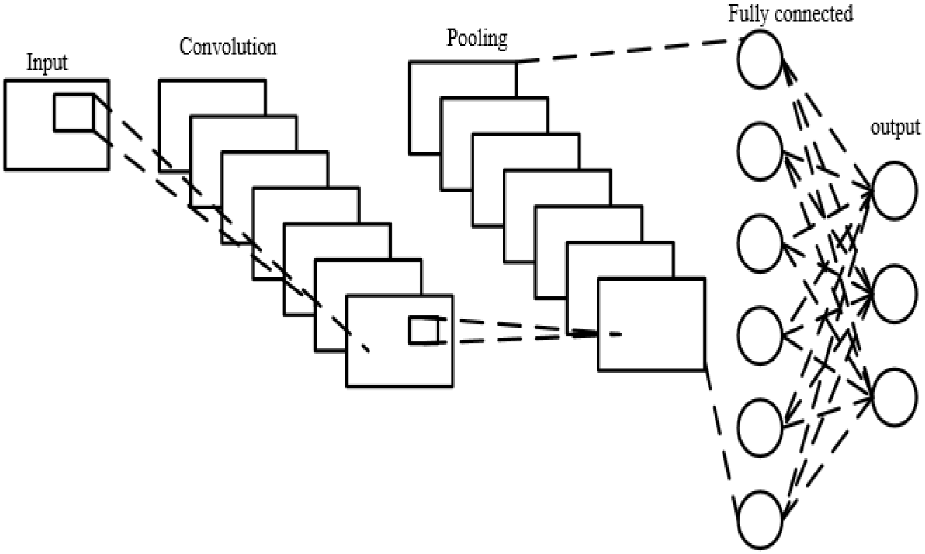


Fig. 2. CNNs for advanced feature extraction.

The CNN issue captures spatial features from the enters monetary data. Let XXX denote the input records, (X) represents the output characteristic map after convolution, and (X) denotes the feature map after activation. The Eq. (4) governing CNN function extraction are as follows:

$$Fc(X) = X * WcFc(X) = X * Wc \tag{4}$$

$$Hc(X) = Activation(Fc(X) + bc)Hc(X) = Activation(Fc(X) + bc) \tag{5}$$

Where, $*$ denotes the convolution operation, Wc represents the CNN weights, and bc denotes the bias term.

D. Forecasting Market Dynamics with LSTM Networks

An important development in sequence learning is represented by LSTM networks, which provide an answer to some of the problems that conventional Recurrent Neural Networks (RNNs) face, most notably the vanishing gradient problem. These networks are highly valuable in many fields, such as financial time series research, because they are specifically designed to be excellent at capturing the long-term relationships present in sequential data.

Memory cells, which can hold information for long stretches of time, are essential to the architecture of LSTM networks. Three fundamental gates control each memory cell:

The input gate regulates the amount of fresh data that enters the memory cell and controls how often it is updated with new information.

Forget Gate: Determines how much of the history kept in the memory cell should be erased, allowing the network to adjust flexibly to shifting circumstances.

Output Gate: Specifies how much data is taken out of the memory cell to be used in further calculations or outputs, making sure that only pertinent data is retained.

LSTM networks excel in economic time collection research because they're top at figuring out traits and temporal dependencies that emerge over the years. Because of their capability to simulate the dynamics of sequential statistics, analysts are highly capable of deciphering the complicated interactions that shape financial markets, gaining the capability to attract conclusions which might be realistic and enhance investment and danger control overall performance.

The input gate, neglect gate, output gate, and candidate cellular cost are the four essential additives of an LSTM unit. Memory cells may be calculated the usage of, and the end result based on those additives.

$$i_t = \sigma_g(W_i x_t + R_i h_t - 1 + b\ _i) \tag{6}$$

$$f_t = \sigma_g(W_f \mathrm{x}_t + \mathrm{R}_f \mathrm{h}_t - 1 + b_f) \tag{7}$$

$$g_t = \sigma c\big(W_g x_t + R_g h_t - 1 + b_g\big) \tag{8}$$

$$o_t = \sigma_g(W_o x_t + R_o h_t - 1 + b_o) \tag{9}$$

In Eq. (6) to Eq. (9) where, R_i , R_g , R_f , R_o odenotes the weights matrices for the preceding brief-term state W_i, W_f, W_g ,W_o ,are the burden matrices within the modern input kingdom. Thus, the cutting-edge output depends at the modern-day long-time period nation, modern-day input, and the preceding state.

E. The Cross-Stitch of CNN-LSTM for Financial Analysis

While LSTM networks are good at shooting temporal dependencies and long-term patterns, CNNs are true at finding spatial linkages within economic facts. A thorough method for comprehending and forecasting market conduct is offered by combining those two structures.

As a characteristic extractor, the CNN aspect reveals localized patterns inside the monetary facts enter. CNNs observe filters to input information thru convolutional layers, which are represented by using Eq. (1) and Eq. (2), with a purpose to extract pertinent functions. Metrics like stock charges and trading volumes have geographical correlations, tendencies, and anomalies which can be highlighted with the aid of these layers. By concentrating at the most critical additives of the facts, pooling layers enhance computing performance by way of similarly refining the derived functions. In addition, LSTM networks deal with the temporal thing. The first step in the CNN-LSTM prediction procedure is statistics input, which is then normalized using z-score standardization. After setting the network's initial weights and biases, the input data are sent through the pooling and convolution layers in order to extract features. The LSTM layer then computes the convolution layer's output to produce the output value. The fully connected layer receives this output value and uses it for additional processing. The procedure entails figuring out the fault and determining if the end condition is satisfied. Error backpropagation keeps improving the model if it doesn't. After training, the version is saved, and forecasted enter records is normalized before being fed into the CNN-LSTM model, which is then skilled for prediction. After that, the standardized output is changed back to its authentic value, concluding the forecasting manner of financial facts processing, making it viable to perceive lengthy-time period patterns and connections. LSTM networks, which can be provided with memory cells and gates, are able to successfully deal with and keep facts across long sequences. The network can modify to moving market situations because the neglect gate manages the retention of previous facts while the input gate controls the inflow of clean statistics into the reminiscence mobile. The output gate controls the extraction of useful statistics for extra calculations or forecasts, ensuring that the community makes choices primarily based completely on applicable information. Fig. 3 shows LSTM networks.

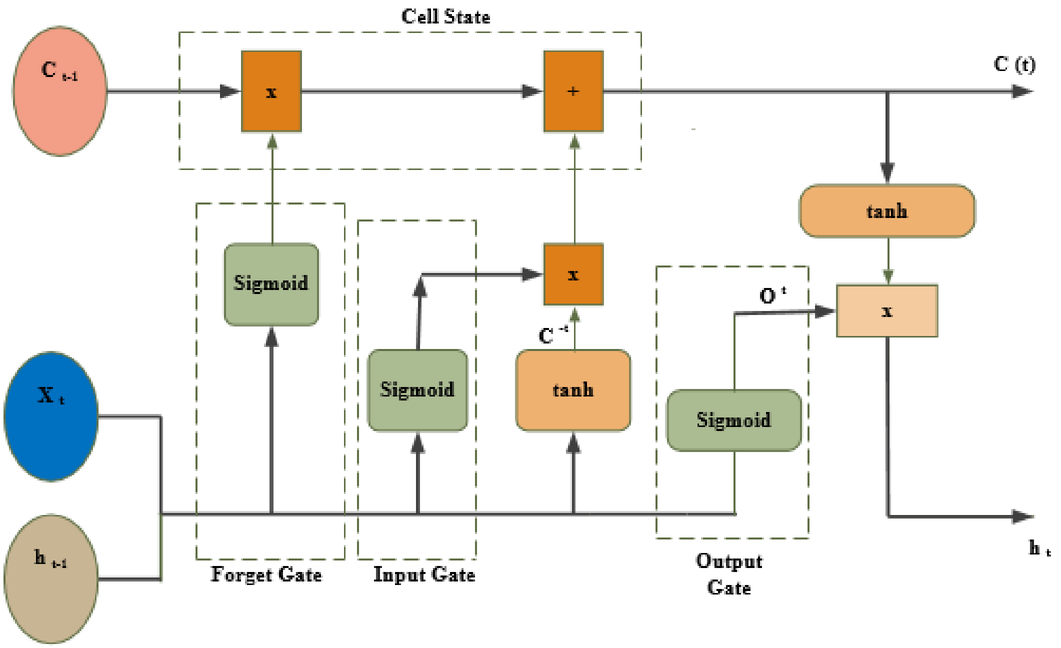


Fig. 3. LSTM networks.

A complete framework for predicting marketplace dynamics is created by analysts by using merging CNNs with LSTM networks. The characteristic extraction technique of the CNN is represented by Eq. (1) and Eq. (2), whereas the reminiscence cellular and gate operations describe the functioning of LSTM networks. This empowers stakeholders to successfully control risks, make properly-knowledgeable selections, and optimize investment strategies based totally on thorough understandings of market dynamics. In addition to making, it easier to apprehend the complicated relationships that form financial markets, the incorporated CNN-LSTM structure improves selection-making talents, which in flip improves funding performance and risk management approaches as shown in the Fig. 4.

The first step within the CNN-LSTM prediction system is facts input, which is then normalized by the usage of z-score standardization. After putting the community's initial weights and biases, the input data are despatched via the pooling and convolution layers to be able to extract capabilities. The LSTM layer then computes the convolution layer's output to supply the output price. The related layer receives this output cost and makes use of it for extra processing. The manner entails figuring out the fault and determining if the end circumstance is happy. Error backpropagation maintains enhancing the model if it doesn't. After training, the version is saved, and forecasted enter records is normalized before being fed into the CNN-LSTM model, that's then trained for prediction. After that, the standardized output is changed back to its unique cost.

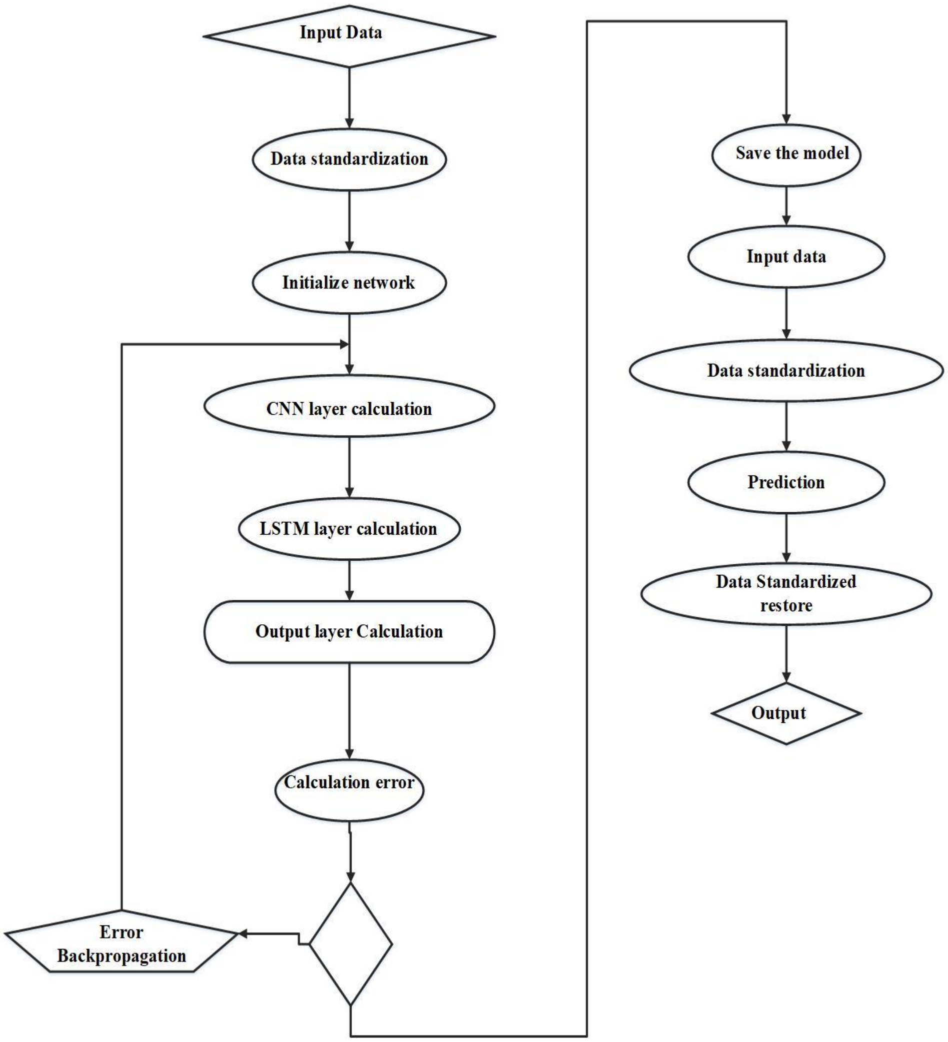


Fig. 4. CNN-LSTM financial analysis process.

V. Result and Discussion

The findings of the proposed study display that the move-sew hybrid network-based method for function integration in quantitative economic evaluation works nicely. Cross-stitch network structures function much better than standard approaches for capturing complicated connections between financial variables, as demonstrated by thorough testing on many financial datasets. To make the algorithms more dependable for economic forecasting and analysis, they undergo instruction utilizing this technique, which results in algorithms that are more resistant, accurate, and typically able to represent actual circumstances across market situations. The performance's inventiveness serves as proof of the cross-stitch hybrid networks' capacity to upend established wisdom and contribute new perspectives to economic modelling.

A. *Analysis of Proposed Financial Dataset*

The Table I shows an example of a data set with financial metrics for sales and profits of various products in different segments and countries It categorizes products into electronics, furniture, clothing, home goods, etc. Identifies various services or products. While the "product" column names a specific product under review, such as smartphones, office chairs, jackets, or laptops. The "Discount Band" column displays the amount of discount applied to each item, none from above, affecting the overall pricing strategy. The "Units Sold" column provides the number of units sold, which is a key figure for determining revenue. "Operating cost" refers to the cost to produce each item, while "selling price" refers to the cost of each item sold, providing the basis for calculating total revenue. The "total sales" column calculates revenue all accounted for before discount rates are applied, providing a lead in terms of revenue potential.

"Discount" represents the cash value of all price discounts offered to customers, which includes the most recent revenue total shown in the "Sales" column, which is the total sales less discounts out of it is an important calculation. Finally, the "Profit" section calculates net income by subtracting COGS from sales, and shows the economic performance of each item. This table gives a clear idea of how various factors such as discounts, production costs, and pricing strategies affect overall profitability in different markets and product segments there.

TABLE I. FINANCIAL ANALYSIS DATA SET EVALUATION

Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit
Electronics	USA	Smartphone	High	500	$200	$350	$175,000	$10,000	$165,000	$100,000	$65,000
Furniture	Germany	Office Chair	Low	300	$100	$200	$60,000	$3,000	$57,000	$30,000	$27,000
Clothing	France	Jacket	Medium	400	$50	$120	$48,000	$4,800	$43,200	$20,000	$23,200
Electronics	Japan	Laptop	None	150	$500	$800	$120,000	$0	$120,000	$75,000	$45,000
Home Goods	Canada	Blender	Medium	600	$25	$70	$42,000	$2,100	$39,900	$15,000	$24,90

B. Training and Testing Accuracy

In Fig. 5, the version's early getting to know section is indicated by the starting factors of the education (blue line) and testing (orange line) accuracies, which can be about 0.3 and 0.4, respectively. The accuracy for both datasets booms dramatically in the course of the course of the epochs, demonstrating the model's ability to research from and generalize from the facts.

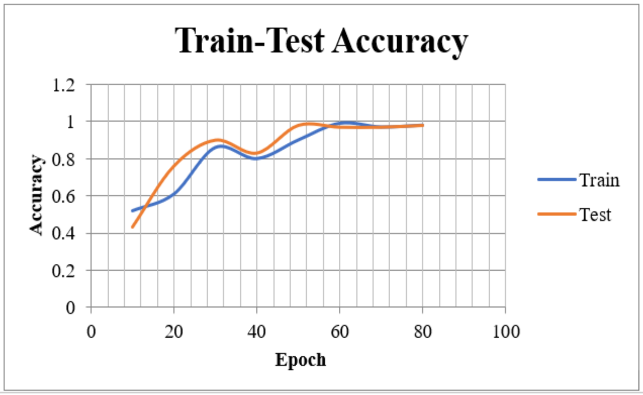


Fig. 5. Train-Test accuracy of CNN-LSTM model.

The checking out accuracy in brief exceeds the training accuracy at epoch 20, indicating a length of progressed generalization before they converge and settle. The version's balance is indicated by way of the 2 traces' convergence at epoch forty, at which point the training and checking out accuracies both technique 0.9 and stay solid.

C. Traing and Testing Loss

The CNN-LSTM model's training and testing loss across 50 epochs is shown in the Fig. 6, which offers insights into the model's generalization capacity and learning efficiency. As the model starts to learn from the data, it is predicted that the testing loss (red line) and training loss (blue line) would start out high, starting at approximately 0.8. For both datasets, there is a sharp drop in loss as training goes on, suggesting that the model is rapidly becoming more efficient by reducing prediction error. The losses for the training and testing datasets substantially converge by epoch 10, falling to less than 0.3.

The little difference between the training and testing losses shows that the model has stabilized, indicating that it has learnt from the training data and can generalize well to new data. To provide high performance and precise predictions, the CNN-LSTM architecture must be resilient and reliable to capture and model the intricate patterns seen in financial data. This is demonstrated by the low and steady loss values in later epochs.

D. Performance Comparison

Table II presents a detailed comparison of different machine learning models and hybrid architectures based on key performance metrics: accuracy, mean absolute error (MAE), root mean square error (RMSE), and Short-term memory models known to capture the time dependence in sequential data achieve 80% accuracy at 85% accuracy. This performs well in terms of error reduction, with 7.18% MAE and 9.14% RMSE, explaining a large proportion of data variability, as indicated by its 92% R^2 score XGBoost, the engineer for gradient-enhancing, outperforms LSTM in accuracy (87%) and precision (90.6%), but with a higher error rate—15.04%; -MAE, 27.02% RMSE— shows that it struggles more with stability but its low 33% R^2 score further highlights its limitations in explaining the variance in the data. The Support Vector Machine (SVM) model, although effective in some cases, performs poorly in this comparison, with an accuracy of 78%, precision of 82%, and the highest error—MAE of 15.61%; and RMSE of 29.33% with R2 score of 21% indicates very low explanatory power.

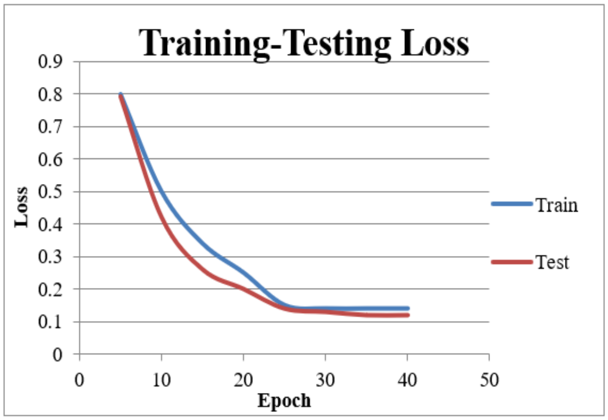


Fig. 6. Training-Testing loss of CNN-LSTM model.

TABLE II. PERFORMANCE METRICS FOR VARIOUS MACHINE LEARNING MODELS IN QUANTITATIVE FINANCIAL ANALYSIS

Model	Accuracy (%)	MAE (%)	RMSE (%)	Precision (%)	R² (%)
LSTM	80	7.18	9.14	85	92.0
XGBoost	87	15.04	27.02	90.6	33.0
SVM	78	15.61	29.33	82	21.0
LSTM+XGBoost	85	8.04	13.93	88	82.0
LSTM-Transformer Attention Hybrid	87	6.05	9.03	87	88.0
LSTM-CNN Cross-Stitch Hybrid Networks	90	5.0	8.0	92	92

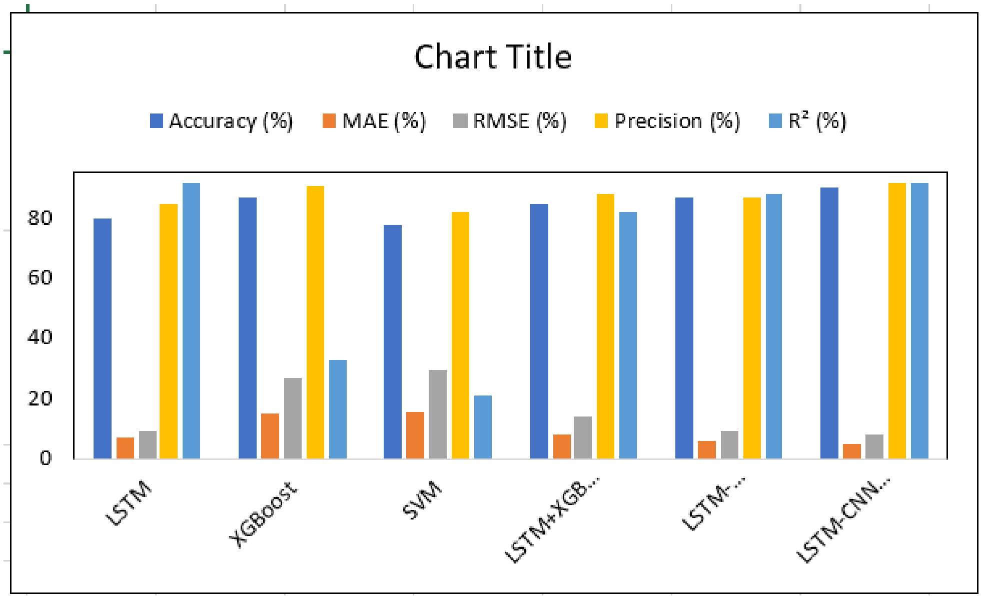


Fig. 7. Performance metrics overview of the proposed model.

The hybrid LSTM+XGBoost model [16] combines the temporal learning of LSTM with the predictive capabilities of XGBoost, yielding balanced results with 85% accuracy and 88% accuracy, as well as improved error metrics (MAE 8.04% and RMSE 13.93% and R^2 is 82 % . The LSTM-Transformer Attention Hybrid model improves LSTM by adding attention, resulting in 87% accuracy, 87% accuracy, error reduction (6.05% MAE, RMSE 9.03%), and R^2 a it comes to an 88% LSTM-CNN Cross-Stitch Hybrid Networks model , in which the LSTM Ne class combines CNN with spatial feature extraction, and stands out in the highest performance: 90% accuracy, 92%; accuracy, and the lowest error rate MAE of 5.0%, RMSE of 8.0%, with 92% R^2, the most robust and effective in this comparison.

In Fig. 7 the graph uses Accuracy, MAE, RMSE, Precision, and R2 to illustrate the outcomes of several approaches. With the greatest Accuracy (90%) and Precision (92%), the LSTM-CNN Cross-Stitch Hybrid Networks function better than the others, however, XGBoost performs worse on all measures. Results for the LSTM+XGBoost and LSTM-Transformer Attention Hybrid models are matched.

E. Performance Evaluation Metrics

To evaluate the performance of the model, this study uses the following metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R²).

The Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (10)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and is the number of observations.

The Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted and actual values. It is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (11)$$

The Coefficient of Determination (R²) indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \qquad (12)$$

The model correctly predicts outcomes 90% of the time, reflecting high overall prediction reliability.

$$\text{Accuracy (ACC)} = \frac{True\ Positive\ \ +True\ Negative}{TP+TN+FP+FN} \tag{13}$$

Of all predicted positive outcomes, 92% are true positives, indicating effective identification of relevant cases.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{14}$$

TABLE III. PERFORMANCE METRICS OF THE PROPOSED MODEL

Accuracy (%)	90
MAE (%)	5.0
RMSE (%)	8.0
Precision (%)	92
R² (%)	92

The performance parameters of the suggested approach are presented in Table III. It can achieve 90% accuracy with a mere 5.0% mean absolute error (MAE), resulting in excellent performance forecast statistics with tiny average errors. Along with the 92% accuracy, which indicates that 92% of instances can be correctly recognized, and the 92% R², which indicates a near-perfect fit and excellent explanation capacity, the 8.0% RMSE indicates that the framework is a resilient error controller.

The performance evaluations of the proposed model, including accuracy, precision, mean absolute error (MAE), root mean square error (RMSE), and R2, are displayed in Fig. 8, a bar chart. The framework is very dependable on all criteria, exhibiting excellent precision, accuracy, and R2 along with low MAE and RMSE levels.

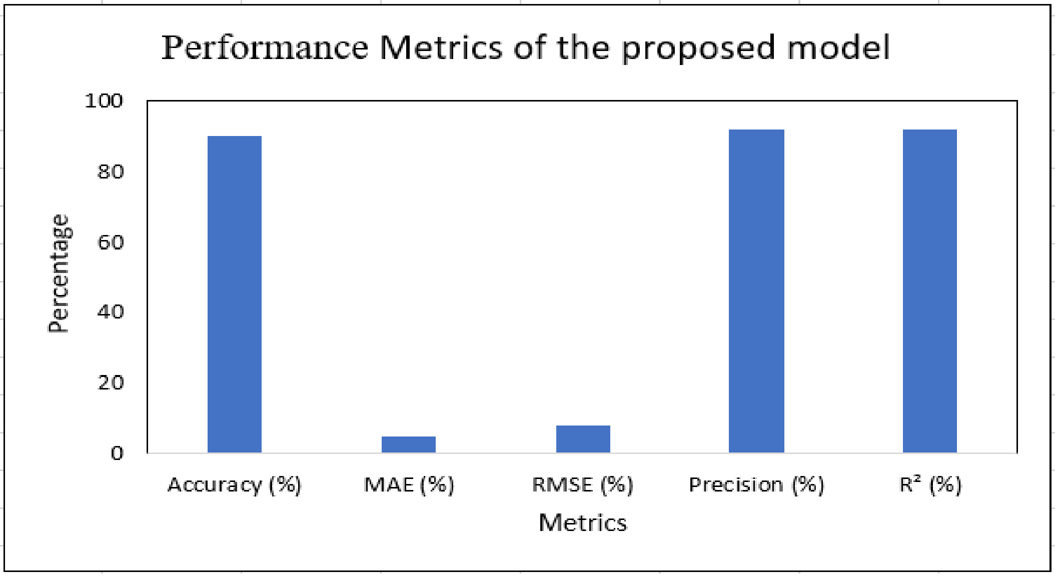


Fig. 8. Performance metrics overview of the proposed model.

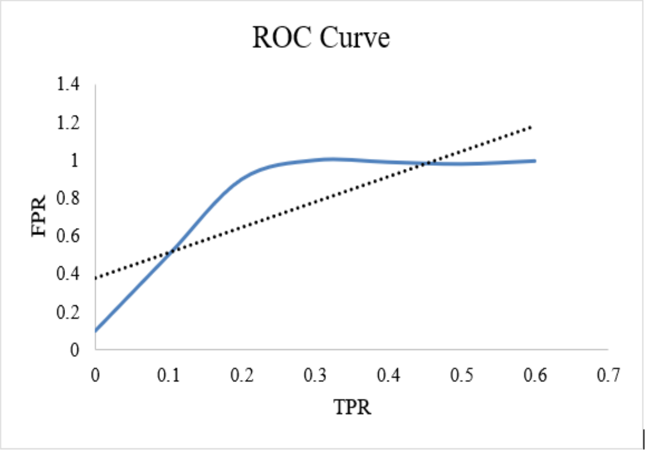


Fig. 9. ROC curve of CNN-LSTM model.

The Fig. 9, represented by the ROC (Receiver Operating Characteristic) curve in the graph. The performance of the model is shown by the blue line, while the performance of a random classifier is shown by the diagonal dotted line, which acts as a baseline. A model with exceptional discriminative capacity would have a sharp ascent towards the top-left corner of the ROC curve, signifying high sensitivity and minimal fall-out. Nonetheless, the graph's ROC curve shows a progressive rise in FPR along with a rise in TPR that begins at the origin (0,0) and moves towards the right. The model improves at first, but as the TPR rises, the FPR rises as well, indicating that the model has difficulty making accurate class distinctions. A TPR of 0.4 to 0.5 is where the performance peaks, while an FPR of more than 0.8 denotes subpar performance. This ROC curve study shows that although the CNN-LSTM model can identify certain trends in the financial data, more data or fine-tuning may be needed to improve the model's overall prediction accuracy and its capacity to distinguish between classes.

F. *Discussion*

The results of the study show that the proposed CNN-LSTM Cross-Stitch Hybrid Network improves the accuracy and reliability of economic forecasts. Compared with traditional models, this hybrid approach exhibits better performance, providing an accuracy of 90% and a lower error rate 5.0% MAE and 8.0% RMSE, indicating a higher model accuracy 92 % and R^2 scores 92% in the capturing power structures are robust in economic terms. Also further emphasizes the difficulty in explaining data variance, making it a powerful tool for economic forecasting. Despite this strength, ROC curve analysis reveals areas of improvement, suggesting that whether the model still struggles with class differences in some cases. These findings highlight the potential of improved hybrid networks in adapting economic paradigms, although further fine-tuning is required to be effective in different contexts.

VI. Conclusion and Future Work

The technologies have allowed researchers to effectively combine spatial and temporal data into a model known as Cross-Stitch Hybrid Networks that provides the most accurate and efficient predictions of stock market trends. The framework is a cross between LSTM and CNN. The comprehensiveness of the data collection is crucial since financial time series differ greatly from one another. Neural networks' capacity to learn functions and concentrate on the distribution of data rather than just input properties throughout the analogy-making process has made this feasible. This work highlights the potential for further study and improvement in this field as well as the ability of sophisticated neural networks to handle the difficulties of financial prediction. The results of the research support the idea that feature aggregation becomes increasingly significant at the highest levels of model efficiency, and that this significance is particularly evident in domains like the stock market. Systems are demonstrated to have an effective pooling and incorporation of learned representation using the Cross-Stitch mechanism, which results in a framework that not only makes more accurate predictions but also scales better and performs well in a variety of scenarios. This section shows that the accuracy and range of the input information have a direct impact on the model's efficacy, indicating the need for full datasets in financial evaluation. To validate the framework's generality and adaptability, additional research might investigate its use in areas linked to various financial goals, such as identifying hazards and optimization of portfolios. To improve the Cross-stitch Combination Networks system's strength and generalization, future work will add additional economic attributes to it and see if it can be applied to other financial sectors, specifically managing risks and portfolio optimization. The information showed that financial executives and analysts may benefit greatly from the Cross-Stitch Hybrid Networks approach, which would enable them to make better-informed and more effective judgments. Cross-stitch hybrid Networks with temporal and spatial components combined represent a major advancement in quantitative financial evaluation. The inability of conventional models to identify correlations among financial time series data—like fluctuations in the public price of bitcoin, for example—is a prevalent problem. For instance, it has been discovered that concentrating on advances in technology with cross-stitching hybrid networks—a combination of CNNs and LSTMs—in sophisticated peer-to-peer lending systems, improves forecast outcomes.

References

[1] J. Wang, J. Wang, W. Fang, and H. Niu, “Financial Time Series Prediction Using Elman Recurrent Random Neural Networks,” Computational Intelligence and Neuroscience, vol. 2016, p. e4742515, May 2016, doi: 10.1155/2016/4742515.

[2] S. Zaheer et al., “A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model,” Mathematics, vol. 11, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/math11030590.

[3] “Stock Price Forecast Based on LSTM Neural Network | SpringerLink.” Accessed: May 24, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93351-1_32

[4] “Financial Analysis by Return on Equity (ROE) and Return on Asset (ROA) - A Comparative Study of HUL and ITC by CMA(Dr.) Ashok Panigrahi, Kushal Vachhani :: SSRN.” Accessed: May 24, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3940100

[5] B. Lakshmi, “A study on financial performance evaluation using DuPont analysis in select automobile companies,” vol. IX, Jan. 2019.

[6] H.-S. Kim, “A Study of Financial Performance using DuPont Analysis in Food Distribution Market,” Culinary Science & Hospitality Research, vol. 22, pp. 52–60, Sep. 2016, doi: 10.20878/cshr.2016.22.6.005.

[7] “A graph-based CNN-LSTM stock price prediction algorithm with leading indicators | Multimedia Systems.” Accessed: May 24, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s00530-021-00758-w

[8] S.-M. Huang, C.-F. Tsai, D. C. Yen, and Y.-L. Cheng, “A hybrid financial analysis model for business failure prediction,” Expert Systems with Applications, vol. 35, no. 3, pp. 1034–1040, Oct. 2008, doi: 10.1016/j.eswa.2007.08.040.

[9] D. J. Meade, S. Kumar, and A. Houshyar, “Financial analysis of a theoretical lean manufacturing implementation using hybrid simulation modeling,” Journal of Manufacturing Systems, vol. 25, no. 2, pp. 137–152, Jan. 2006, doi: 10.1016/S0278-6125(06)80039-7.

[10] D. Wu, X. Wang, and S. Wu, “A hybrid method based on extreme learning machine and wavelet transform denoising for stock prediction,” Entropy, vol. 23, no. 4, p. 440, 2021.

[11] S. Mehtab and J. Sen, “Analysis and Forecasting of Financial Time Series Using CNN and LSTM-Based Deep Learning Models,” in Advances in Distributed Computing and Machine Learning, vol. 302, J. P. Sahoo, A. K. Tripathy, M. Mohanty, K.-C. Li, and A. K. Nayak, Eds., in Lecture Notes in Networks and Systems, vol. 302. , Singapore: Springer Singapore, 2022, pp. 405–423. doi: 10.1007/978-981-16-4807-6_39.

[12] M. Eling and S. D. Marek, “Do Underwriting Cycles Matter? An Analysis Based on Dynamic Financial Analysis,” Table of, p. 131, 2012.

[13] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM,” Nov. 21, 2019, arXiv: arXiv:1911.09512. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/1911.09512

[14] Z. Shang, “The Research of Financial Forecasting and Valuation Models,” presented at the 2021 International Conference on Enterprise Management and Economic Development (ICEMED 2021), Atlantis Press, Jun. 2021, pp. 70–76. doi: 10.2991/aebmr.k.210601.012.

[15] “Finalcial analysis quantitative data.” Accessed: May 24, 2024. [Online]. Available: https://www.kaggle.com/datasets/foridurrahman/finalcial-analysis-quantitative-data

[16] J. Zhou, “Predicting Stock Price by Using Attention-Based Hybrid LSTM Model,” Asian Journal of Basic Science & Research, vol. 6, no. 2, pp. 145–158, 2024.

Enhanced Early Detection of Oral Squamous Cell Carcinoma via Transfer Learning and Ensemble Deep Learning on Histopathological Images

Gurjot Kaur[1], Sheifali Gupta[2], Ashraf Osman Ibrahim[3]*, Salil bharany[4]*,
Marwa Anwar Ibrahim Elghazawy[5], Hadia Abdelgader Osman[6], Ali Ahmed[7]
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1, 2, 4]
Department of Computer and Information Sciences, Universiti Teknologi Petronas, Seri Iskandar, Malaysia[3]
Positive Computing Research Center, Emerging and Digital Technologies Institute[3]
Computer Department, Applied College, Northern Border University, Arar, Saudi Arabia[5, 6]
Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, 21589, Saudi Arabia[7]

***Abstract*—Oral Squamous Cell Carcinoma (OSCC) is one main kind of oral cancer; early diagnosis is rather important to increase patient survival chances. This study investigates the application of advanced deep learning techniques including transfer learning and ensemble learning to increase the accuracy of oral squamous cell cancer (OSCC) diagnosis using histopathological image analysis. Two transfer learning models, EfficientNetB3 and ResNet50, support the suggested method to extract suitable features from the histopathological images. Both models permit fine-tuning to improve their classification accuracy. On tests taken after the initial training, the EfficientNetB3 model scored 96.15%. Later on, training ResNet50 yielded a test accuracy of 91.40%. Weighted voting merged several models into an ensemble model designed to maximize the strengths of each network. With a test accuracy of 98.59% and a training accuracy of 99.34%, the ensemble model showed notably higher performance than the values obtained by the individual models. Divided into OSCC and standard categories, the collection has 5,192 extremely well-resolved images. The images were used to create training, validation, and testing sets. We used this method to consistently evaluate the model's performance and reduce overfitting. Furthermore, the ensemble model proved to be quite accurate with recall and F1 scoring, thereby proving its capacity to routinely identify OSCC images. Both groups produced ROC curves, and the area under the curve (AUC) demonstrated excellent model performance. Transfer learning and ensemble learning are used together in this study to show that OSCC can be found early and consistently in histopathology images. The findings reveal that the recommended strategy could be a consistent tool to assist pathologists in the precise and timely detection of OSCC, thereby improving patient treatment and outcomes.**

Keywords*—Oral Squamous Cell Carcinoma (OSCC); histopathology images; transfer learning; ensemble learning; EfficientNetB3; ResNet50; deep learning; cancer detection; medical image analysis

I. Introduction

Oral Squamous Cell Carcinoma (OSCC) [1] is one of the most prevalent forms of cancer worldwide, accounting for over 90% of oral cancers. Usually affecting the squamous cells guarding the oral cavity, the disorder arises in malignant tumors that, if not found and treated quickly, can spread to other parts of the body. According to the World Health Organization (WHO), OSCC ranks in the top ten most common malignancies globally, with especially high prevalence in South Asia, Southeast Asia, and parts of Europe. OSCC has a significant global influence considering an estimated 300,000 new cases and around 145,000 deaths annually. The OSCC diagnosis primarily dictates its prognosis; early-stage diagnosis considerably increases the chances of successful therapy and survival [2]. Among the most often occurring risk factors of OSCC, a multifactorial etiopathogenesis including tobacco usage, alcohol intake, betel quid chewing, and human papillomavirus (HPV) infection stands. Growing emphasis on the importance of molecular changes and genetic predispositions in the evolution of OSCC has been observed in recent years. Although OSCC's molecular biology is now well known, the main diagnostic tool that remains is histopathological investigation of tissue biopsies. This method is formed by a pathologist's microscopic examination of tissue samples in search of malignant cells. Nevertheless, this standard diagnostic approach has important disadvantages including inter-observer variability, the time-consuming character of the operation, and the likelihood of misinterpretation coming from the subjective interpretation of histological features [3]. Early and correct OSCC detection is highly significant, thus the scientific and medical industries have worked hard to develop automated diagnostic tools that can help pathologists and minimize diagnosis mistakes. In this sense, applying machine learning (ML) and artificial intelligence (AI) in medical imaging has become an interesting road for innovation. The development of digital pathology and advances in image analysis methods have presented opportunities for the design of AI-driven diagnostic systems capable of very accurate and fast histological image interpretation [4]. OSCC presents varied clinical signs based on tumor sizes and location, such as erythroplakia, leukoplakia, non-healing ulcers, lumps, or persistent sores. Other symptoms include voice changes, dysphagia, pain, numbness, bleeding, loose teeth, and, in advanced stages, neck swelling due to lymph node metastasis. The etiology of OSCC is unclear, though major risk factors include tobacco use and alcohol consumption, which together significantly increase cancer risk [5]. High-risk human papillomavirus (HPV), particularly HPV-16, is also a major factor, especially for oropharyngeal cancers [6]. Additional risks include poor oral hygiene, environmental toxins, chronic trauma, a weakened immune system, and genetic predispositions. Understanding these factors is critical for early detection and prevention programs aimed at reducing OSCC incidence and mortality [7]. Treatments for OSCC combine a multimodal approach tailored to the tumor's stage, location, and patient health. Major treatments include surgery, radiation, and chemotherapy, often used together for optimal results [8]. Surgery, the primary treatment for localized OSCC, aims to remove the tumor with clear margins to minimize recurrence. Reconstructive surgery may follow to restore function and appearance after partial or full removal of affected parts, such as the tongue or jawbone. Radiation therapy, typically external beam, is used as a supplement to surgery or as a main treatment when surgery isn't viable [9]. Chemotherapy, often paired with radiation (chemoradiation), is reserved for advanced or metastatic cases and patients with recurrent OSCC. Targeted therapies and immunotherapies offer appealing adjunct treatments by focusing on molecules linked to tumor growth [10]. Supportive care, including nutrition, pain management, and speech therapy, helps maintain quality of life. Multidisciplinary tumor boards optimize treatment by balancing patient-specific factors with evidence-based guidelines and chosen treatment methods. In oncology, where early detection significantly influences patient prognosis, the diagnosis of Oral Squamous Cell Carcinoma (OSCC) provides a huge challenge. Sometimes subjectivity, variability in interpretation, and the need for specialist knowledge limit conventional diagnosis approaches such as histological study and clinical examination, therefore causing inconsistent and delayed diagnosis [11]. Although machine learning has advanced, present models for OSCC detection from histopathology images often depend on single-model architectures, which might not adequately reflect the complex and heterogeneous character of OSCC. Furthermore, confusing accurate detection results in variations in image quality and staining methods. More strong, accurate, and dependable diagnostics tools that can combine several models to improve detection possibilities are desperately needed. This work fills in this need by looking at the application of transfer learning and ensemble learning approaches to create a better, ensemble-based model for early and accurate OSCC identification [12]. The main goal of this work is to provide a strong and accurate strategy based on advanced deep learning [13] approaches more especially, transfer learning [14] and ensemble learning [15] for the detection of Oral Squamous Cell Carcinoma (OSCC). The work intends to extract discriminative characteristics from histopathology images using the strengths of two state-of-the-art pre-trained models, EfficientNetB3 and ResNet50. The study aims to raise the general classification accuracy and dependability of OSCC detection by optimizing these models and combining their outputs using an ensemble learning method. With an eye on measurements including accuracy, precision, recall, F1 score, and AUC-ROC, the study also seeks to evaluate the ensemble model against individual models. The ultimate aim is to develop an ensemble model significantly above present methods, thus providing a more consistent diagnostic tool for early OSCC diagnosis and aiding in improving patient outcomes.

The primary contributions of this research are:

- This work presents a new ensemble learning method based on two state-of-the-art pre-trained models, EfficientNetB3 and ResNet50, for Oral Squamous Cell Carcinoma (OSCC) identification from histopathology images.
- Specifically, with OSCC identification, the work shows medical image analysis transfer learning performance. The work uses transfer learning to overcome the difficulties provided by insufficient labeled data thus improving the generalizing power of the model by fine-tuning pre-trained models on histomorphology images.
- The study evaluates the proposed ensemble model exhaustively using significant performance measures including accuracy, precision, recall, F1 score, and AUC-ROC. Since the ensemble model trumps individual models, the results demonstrate that it offers a more consistent diagnosis tool for early OSCC identification. Development of the field of medical image analysis and improvement of oncology clinical results depend on this study.

II. Literature Review

The detection of OSCC has become even more critical considering the substantial fatality rate connected with late-stage diagnosis. Deep learning, notably in CNNs and Transfer Learning, has lately shown promise in boosting the accuracy of OSCC detection from histopathological images.

Fanizzi et al. (2024) [16] used an explainable CNN model to examine oropharyngeal squamous cell cancer, training on CT images of 499 OPSCC patients with an independent test set of 92. Using an Inception-V3 architecture, they achieved a 73.50% AUC, highlighting tumor locations via Grad-CAM and emphasizing CNN relevance in therapy.

Paramasivam et al. (2024) [17] used deep learning with three modified CNN architectures, including DENSENET-121 variants, to diagnose OSCC in 5,492 histopathological images, achieving 97.03% accuracy. The study highlighted AI's potential to reduce human errors and improve diagnostic accuracy, addressing the limitations of traditional mouth cancer diagnosis methods.

Weber et al. (2024) [18] studied Stimulated Raman Histology (SRH) with deep learning for Oral Squamous Cell Carcinoma categorization. Using a VGG19 CNN, the model achieved balanced accuracies of 0.90 (SRS) and 0.87 (SRH) from SRH images transformed to resemble H&E sections, highlighting AI's efficiency in intraoperative OSCC detection.

Albalowitz et al. (2024) [19] developed a deep learning model using EfficientNetB3 for OSCC diagnosis, achieving 99% accuracy, precision, recall, and F1-score. The study utilized 1,224 histograms from 230 individuals, leveraging data augmentation, regularization, and optimization. This work demonstrates the potential of deep learning in improving OSCC diagnostic accuracy.

Kumar et al. (2024) [20] examined deep learning for early OSCC identification from histopathology images. Using an enhanced Inception-Resnet-V2 CNN model, the study achieved 91.78% accuracy in distinguishing benign from malignant biopsy images. The paper highlights deep learning's potential in automating OSCC diagnosis and improving early clinical interventions.

Mishra et al. (2024) [21] explored CNNs for early OSCC identification, achieving 98.49% training accuracy, 86.89% validation accuracy, and 89.37% testing accuracy. The model's F1-score of 0.89 for both classes highlights CNNs' potential for improving OSCC screening protocols and patient outcomes, thus reducing fatality rates linked to late diagnosis.

Siddique et al. (2024) [22] used biopsy-derived histological images to explore oral cancer detection techniques, emphasizing image pre-processing's role in enhancing CNN model performance. Models including VGG16, VGG19, InceptionV3, AlexNet, and ResNet50 achieved respective accuracies of 84%, 82%, 67%, 76%, and 42%, showcasing deep learning's potential in cancer diagnosis.

SMira et al. (2024) [23] investigated early oral cancer detection using deep learning and smartphone-based imagery. They developed a resampling method and "center positioning" image-capturing technique to manage variability. The deep learning network achieved 83.0% sensitivity, 96.6% specificity, 84.3% accuracy, and 83.6% F1 score on 455 test images.

Deo et al. (2024) [24] studied the categorization of Oral Squamous Cell Carcinoma (OSCC) using a Vision Transformer (ViT) framework. The updated ViT model outperformed eight pre-trained deep learning models on a dataset of 4,946 images, achieving 97.78% accuracy, 96.72% specificity, and 98.80% sensitivity, proving ViT's efficacy on smaller datasets.

Soni et al. (2024) [25] studied early OSCC diagnosis using an improved EfficientNetB0 model and Dual Attention Network (DAN). Achieving 91.1% accuracy, 92.2% sensitivity, 91.0% specificity, and a 92.3% F1 score, EfficientNetB0 outperformed 17 pre-trained models, highlighting deep learning's clinical potential for early OSCC identification and therapy enhancement.

Deo et al. (2024) [26] developed a DL model for OSCC detection from histopathological images. Using 2D empirical wavelet transform and a ResNet50-DenseNet201 ensemble, binary classification achieved 92% accuracy. The method enhances diagnostic accuracy, reduces human error, and accelerates classification, demonstrating deep learning's potential to aid clinical decision-making.

Saraswathi et al. (2023) [27] examined AlexNet's classification of Oral Squamous Cell Cancer (OSCC) using 5,192 histopathological images. AlexNet outperformed ResNet, achieving 89% accuracy and 60% loss compared to ResNet's 78% accuracy and 82% loss. The study suggested future improvements through larger datasets, software applications, and modified AlexNet models.

Ahmad et al. (2024) [28] studied AI approaches for OSCC detection via histopathological images. Three methods were compared: Gabor+CatBoost, ResNet50+CatBoost, and a hybrid Gabor+ResNet50+CatBoost. The hybrid achieved 94.92% accuracy, 95.51% precision, and 94.9% F1 score. PCA reduced feature dimensionality, improving performance for accurate OSCC diagnosis.

Nagarajan et al. (2023) [29] developed a deep learning system to diagnose Oral Squamous Cell Carcinoma from histopathology images, using a Modified Gorilla Troops Optimizer. MobileNetV3 achieved the highest accuracy (0.89), which increased to 0.95 after optimization. This method demonstrates swarm intelligence's potential to improve OSCC detection accuracy.

Begum et al. (2023) [30] use deep learning models to automatically detect OSCC from histological images, employing transfer learning and layer modification. Among four pre-trained CNN models (NASNet Large, InceptionNet, Xception, DenseNet 201), DenseNet 201 achieved the highest accuracy of 91.25%, showcasing DL-based advancements in OSCC diagnosis and treatment.

III. Methodology

A. Dataset Distribution

This work uses a dataset [31] consisting of 5,192 histopathology images, carefully selected to identify Normal tissues and Oral Squamous Cell Carcinoma (OSCC). As Fig. 1 shows, the dataset is almost balanced with 2,698 OSCC images (52%) and 2,494 Normal images (48%). The right pie chart shows the percentage distribution and the left bar chart shows the image count in every class. This balanced set guarantees a fair training approach, therefore reducing possible categorization bias. Three subsets comprising 3,634 images for training, 779 images for validation, and 779 images for testing comprise the dataset. Using a strong and well-organized dataset, this section helps deep learning models to be developed, refined, and evaluated for effective OSCC detection.

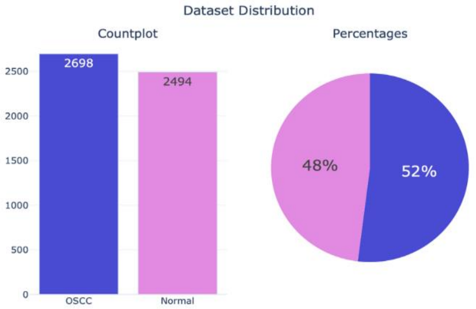


Fig. 1. Dataset distribution.

B. Input Dataset

The normal tissue samples in Fig. 2 show a well-organized structure typical of healthy oral epithelium. The cells have exactly defined borders and firmly packed layers that match their consistent size and form. The way the compartments are arranged is logical there are no obvious anomalies or deviations. There is no hyperchromatism, which would indicate non-malignant tissue; the nuclei are constant in size. Conversely, the Oral Squamous Cell Carcinoma (OSCC) samples in Fig. 3 reveal a disturbance in cellular architecture. OSCC sample cells are irregular, bigger, and show pleomorphism that is, a great range in size and form. Commonly an indication of cancer, the hyperchromatic nature of the nuclei results from the higher DNA content, which seems darker. Furthermore, the OSCC images show abnormal keratinization and uneven stratification of the epithelium with layers that are no longer clearly defined or ordered. These pathological abnormalities are important in histopathological diagnosis since they mirror the aggressive character of OSCC, in which disorganized, malignant cells replace the normal, ordered tissue.

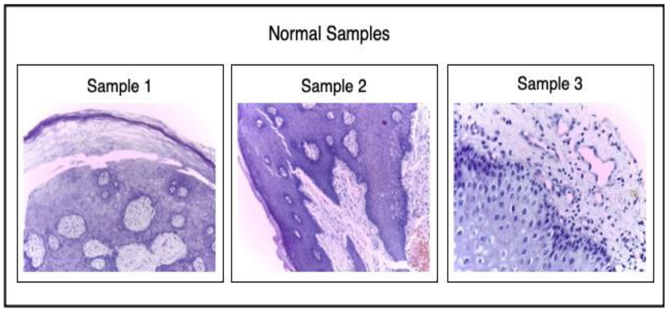


Fig. 2. Dataset samples of normal images.

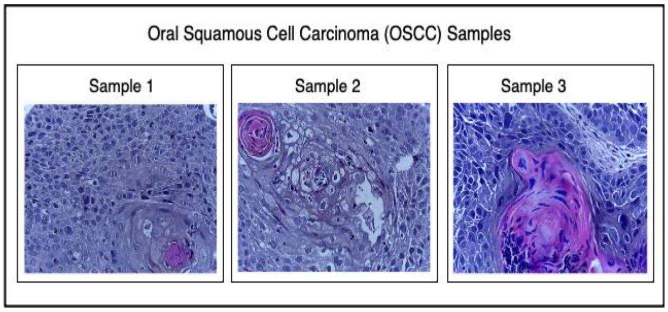


Fig. 3. Dataset samples of OSCC images.

C. Data Preprocessing and Augmentation

Preprocessing is a crucial step that greatly influences the performance of a deep-learning model developed to classify OSCC from histopathological images. Essential for getting the dataset ready for efficient training, validation, and testing are image resizing, normalizing, and augmenting preprocessing techniques [32].

1) Image resizing: Every image is reduced to a consistent dimension suitable for input into the deep learning models thereby guaranteeing consistency over the dataset. This scaling method converts every image into such size since many CNN designs demand each image to be a specific size usually (224×224). This phase guarantees that the model manages images of the same resolution, therefore facilitating efficient training and reducing computer complexity.

2) Normalization: Normalization is carried out to rescale the pixel values of the images to a standardized range, usually ranging from 0 to 1. The process of normalizing can be quantitatively expressed in (1):

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where, X represents the original pixel values, min(X) is the minimum pixel value, and max(X) is the maximum pixel value. By normalizing the input values to a similar range, this step enhances the convergence speed and stability of the training process.

3) Augmentation: Data augmentation provides artificial unpredictability and increases the training dataset size. Another uses horizontal flipping, rotation, zooming, and shifting among other techniques. Mathematically, augmentation can be stated in (2):

$$X' = T(X) \quad (2)$$

Where X' is the augmented image and T is the transformation carried out, say rotation by a certain angle θ or scaling by a factor s. Exposing the model to several image transformations during training, this stage is vital for improving its robustness and lowering overfitting. Using their integration into the model pipeline, these preprocessing actions guarantee the well-preparedness of the dataset, so facilitating the deep learning model to achieve high accuracy in OSCC detection and generalization capability.

IV. Proposed Methodology

Fig. 4 demonstrates the proposed methodology for the detection of Oral Squamous Cell Carcinoma (OSCC). The present study is based on a combination of two pre-trained deep learning models: EfficientNetB3 and ResNet50. These models utilize transfer learning techniques and are fine-tuned to increase their performance on the OSCC dataset. The outputs of various models are then integrated using ensemble learning, leading to an enhanced classification performance. Below is a thorough description of the proposed methodology.

A. EfficientNetB3 Architecture

EfficientNetB3 is a deep learning model known for balancing accuracy and efficiency by scaling network depth, width, and resolution, as shown in Fig. 5. It uses a 3x3 convolutional layer followed by MBConv layers, depthwise separable convolutions, and squeeze-and-excitation (SE) blocks for feature extraction. Larger kernel sizes (5x5) in deeper layers help capture contextual information for OSCC differentiation. After feature extraction, the model classifies Normal and OSCC tissues. Its computational efficiency and performance make it ideal for medical imaging, especially when combined with ResNet50 in an ensemble model for OSCC detection.

EfficientNet-B3 is a scalable model that balances network depth, width, and resolution for optimal performance. The scaling is governed by a compound coefficient ϕ, with the scaling laws given in (3):

$$d = \alpha^{\phi}, w = \beta^{\phi}, r = \gamma^{\phi} \quad (3)$$

where d, w, and r represent the depth, width, and resolution of the network, respectively, and α, β, γ are constants determined through grid search. The output of EfficientNet-B3 is passed through a global average pooling layer, followed by a dense layer to produce the final classification output shown in (4):

$$Z_{efficientnet} = softmax\left(W_{efficientnet}.f_{global} + b_{efficientnet}\right) \quad (4)$$

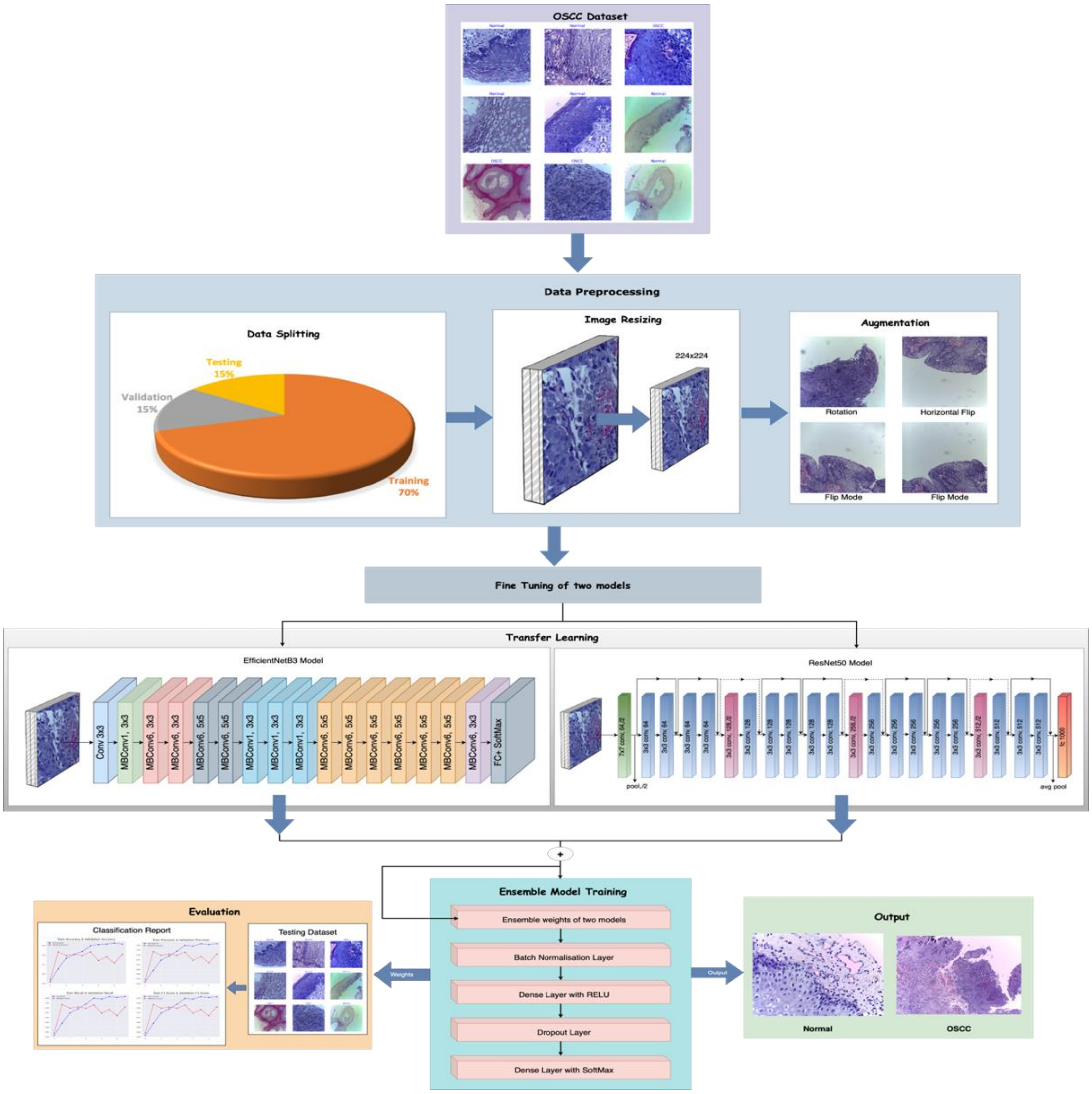


Fig. 4. Proposed methodology.

Where, $Z_{efficientnet}$ represents the predicted probabilities, f_{global} is the feature vector after global pooling, and $W_{efficientnet}$ and $b_{efficientnet}$ are the weights and biases of the final classification layer.

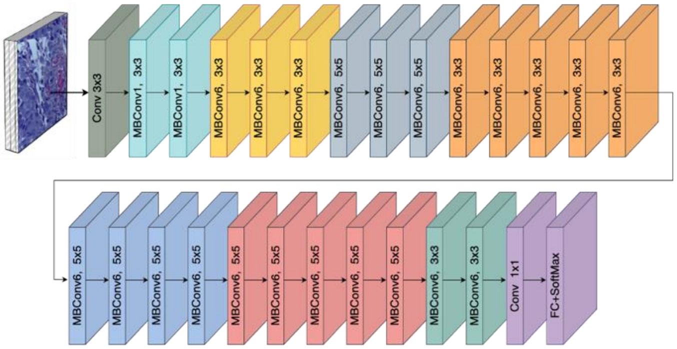


Fig. 5. EfficientNetB3 architecture.

B. *Resnet50 Architecture*

Fig. 6 shows the ResNet50 architecture, a deep CNN designed to overcome challenges in deep network training with 50 convolutional layers, batch normalization, and ReLU activations. Residual connections prevent vanishing gradients, allowing deeper networks. The architecture starts with a 7x7 convolution and max-pooling, followed by four stages with residual blocks and filters ranging from 64 to 512. Spatial dimensions are reduced via stride-2 convolutions for efficiency. The final section includes global average pooling and a fully connected layer. For OSCC detection, the last layer classifies images into Normal and OSCC, helping detect subtle variations in oral cancer.

ResNet50 is a residual network that employs skip connections to address the problem of vanishing gradients in deep networks. The fundamental building block of ResNet50 is the residual block, defined mathematically as in (5):

$$y = F(x, \{W_i\}) + x \tag{5}$$

Where, x is the input, $F(x,\{W_i\})$ represents the convolutional operation with weight parameters $\{W_i\}$, and y is the output of the residual block. The network consists of 50 layers, including convolutional, batch normalization, and ReLU activation layers. The final output of the ResNet50 model, after passing through a global average pooling layer, is given by (6):

$$Z_{resnet} = softmax\left(W_{resnet}.Y_{global} + b_{resnet}\right) \quad (6)$$

Where, Z_{resnet} is the predicted class probabilities, W_{resnet} and b_{resnet} are the weights and biases of the final dense layer, and Y_{global} is the output from the global pooling layer.

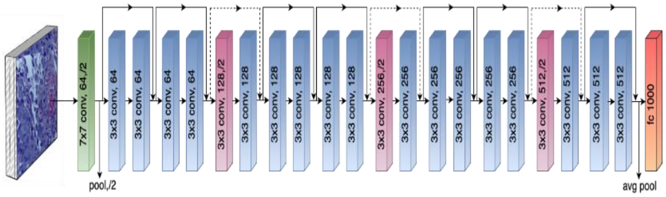


Fig. 6. ResNet50 architecture.

Transfer learning significantly enhanced OSCC detection accuracy by utilizing pre-trained models like EfficientNetB3 and ResNet50, which were fine-tuned on the OSCC dataset. These models transfer learned features, such as texture and shape, from large datasets, reducing the need for extensive labeled data and improving feature generalization. Fine-tuning helps the models capture OSCC-specific patterns, preventing overfitting on small datasets and accelerating training convergence, allowing for efficient and precise differentiation between normal and malignant tissues. This technique proved essential for optimizing accuracy and training speed in medical image analysis.

C. *Ensemble Learning*

The ensemble model with the architecture seen in Fig. 7 combines the predictions of ResNet50 and EfficientNet-B7 to leverage the strengths of both architectures, thereby enhancing overall classification accuracy.

1) Ensemble model design: The ensemble approach is implemented using a weighted average method, where the final prediction is a weighted combination of the individual models' outputs. The output predictions from ResNet50 and EfficientNet-B7 are combined using a weighted average approach shown in (7):

$$Z_{ensemble} = W_{resnet}.Z_{resnet} + W_{efficientnet}.Z_{efficientnet} \quad (7)$$

where $Z_{ensemble}$ is the final ensemble prediction, and $W_{resnet}, W_{efficientnet}$ are the weights assigned to the ResNet50 and EfficientNet-B3 outputs, respectively. These weights are optimized based on the validation performance of each model.

2) Ensemble model algorithm: This algorithm outlines the process of training individual models, obtaining their predictions, combining those using weighted averages or voting, and optionally training a meta-learner to refine the ensemble's predictions. The Algorithm of the ensemble model is shown in Table I.

TABLE I. ALGORITHM OF ENSEMBLE MODEL

```
Algorithm 1: Ensemble Model for Oral Squamous Cell Carcinoma Detection
Input: Histopathological images $D_{train}, D_{val}, D_{test}$ , number of epochs $E$, batch size $B$, learning rate $\eta$.
Output: Trained ensemble model for Oral Squamous Cell Carcinoma detection.
Step 1: Model Initialization
    1.1 Initialize EfficientNetB3 and ResNet50 models with pre-trained ImageNet weights.
    1.2 Configure the final dense layers for detection of two classes (Normal and OSCC).
Step 2: Model Training
    2.1 For each epoch $e$ from 1 to $E$:
    2.2 For each batch $b$ in $D_{train}$:
    2.2.1 Perform forward propagation through EfficientNetB3 to obtain $Z_{efficientnet.}$
    2.2.2 Perform forward propagation through ResNet50 to obtain $Z_{resnet}$.
    2.2.3 Combine the outputs using the weighted average:
            $Z_{ensemble} = W_{resnet}.Z_{resnet} + W_{efficientnet}.Z_{efficientnet}$
    2.2.4 Compute the cross-entropy loss $\mathcal{L}(Z_{ensemble}, Y)$ where $Y$ is the true label.
    2.2.5 Backpropagate the loss and update the weights using the Adamax optimizer with learning rate $\eta$.
Step 3: Model Validation
    3.1 Evaluate the ensemble model on $D_{val}$ after each epoch.
    3.2 Adjust the weights $W_{efficientnet}$ and $W_{resnet}$ based on validation performance.
Step 4: Model Evaluation
    4.1 After completing training, evaluate the model on $D_{test}$.
    4.2 Compute the final accuracy, precision, recall, and F1-score.
Step 5: Model Deployment
    5.1 Save the trained ensemble model for future use.
```

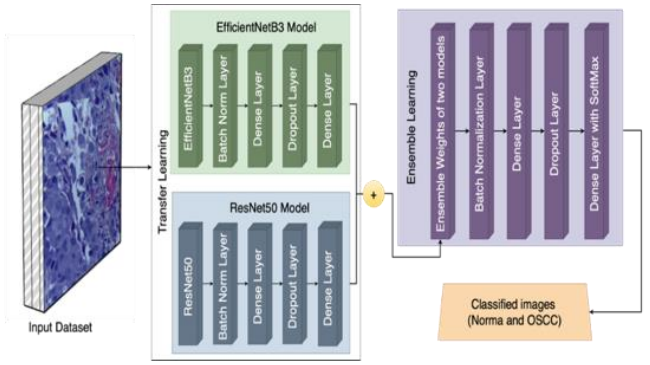


Fig. 7. Ensemble learning architecture.

V. RESULTS

A. *Model Evaluation*

The work aimed at the identification of Oral Squamous Cell Carcinoma (OSCC) using a deep learning approach combining the strengths of two strong models EfficientNetB3 and ResNet50 by constructing the Ensemble model for improved results. Divided into training, validation, and testing, the models were developed using a set including 5,192 histopathology images. The results reveal the degree of increase in OSCC detection classification accuracy resulting from the combined approach. The Confusion Matrix, Performance criteria, and State of the art Comparison are fully described below.

1) EfficientNet-B3's accuracy and loss analysis: The examination of the EfficientNetB3 model's training and validation accuracy, along with its loss metrics across numerous epochs, is shown in Fig. 8 and Table II. The model demonstrates strong generalization and learning capacity, with an early drop in training loss and gradual reduction in both training and validation loss, ultimately converging to low values. The model quickly reaches 90% accuracy early in training, with validation accuracy following suit. Both accuracy measures level off in later stages, indicating the model's peak performance and suitability for reliable predictions, without overfitting.

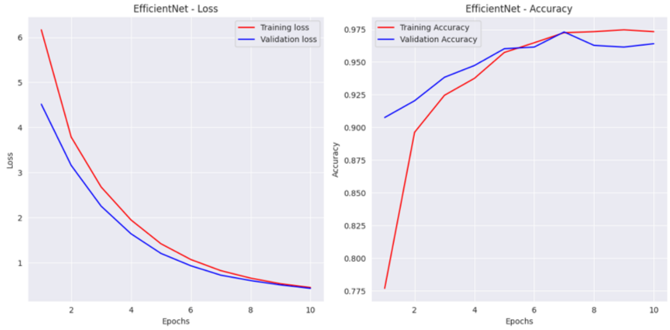


Fig. 8. Loss and accuracy analysis of EfficientNetB3 architecture.

The precise performance per epoch is shown in Table II. It shows that with every epoch the model reduces loss and steadily gains accuracy. The strong learning capacity and adaptability of the EfficientNetB3 model to fresh data are shown by this constant improvement and convergence in both accuracy and loss values [33-36].

TABLE II. EFFICIENTNETB3 MODELS PERFORMANCE PER EPOCH

Epoch	Time(sec)	Loss	Accuracy	Validation Loss	Validation Accuracy
1	131	6.1664	0.7768	4.5153	0.9076
2	54	3.7852	0.8963	3.1570	0.9204
3	54	2.6791	0.9246	2.2538	0.9384
4	53	1.9470	0.9375	1.6400	0.9474
5	54	1.4193	0.9573	1.2048	0.9602
6	54	1.0688	0.9648	0.9303	0.9615
7	54	0.8223	0.9725	0.7222	0.9730
8	54	0.6560	0.9733	0.6015	0.9628
9	54	0.5315	0.9747	0.5054	0.9615
10	54	0.4487	0.9733	0.4286	0.9641

2) ResNet-50's accuracy and loss analysis: The ResNet50 model was trained for 10 epochs, with performance metrics in Table III and accuracy and loss curves in Fig. 9. The training and validation losses steadily decreased, showing effective learning, with a slight generalization gap around the 7th epoch. Training loss stabilized after the 4th epoch, while training accuracy approached 95% by the 10th epoch. Validation accuracy improved gradually, staying slightly below training accuracy. The close alignment of both accuracy curves indicates minimal overfitting and strong generalization. This demonstrates the ResNet50 model's suitability for the current classification task.

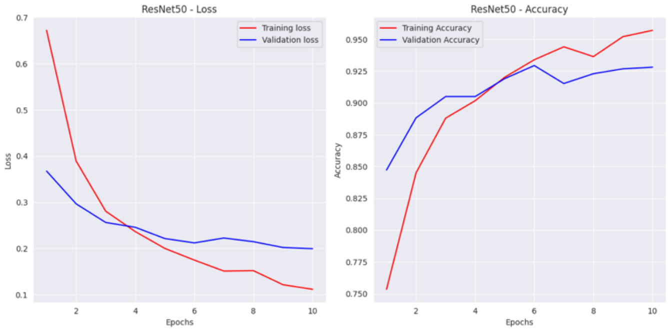


Fig. 9. Loss and accuracy analysis of ResNet50 architecture.

Table III, which illustrates the performance measures for each epoch, shows a constant increase in both training and validation accuracy together with related declines in loss. This development implies that in every epoch the parameters of the model are being tuned somewhat successfully.

TABLE III. RESNET50 MODELS PERFORMANCE PER EPOCH

Epoch	Time(sec)	Loss	Accuracy	Validation Loss	Validation Accuracy
1	43	0.6721	0.7534	0.3674	0.8472
2	36	0.3891	0.8451	0.2965	0.8883
3	36	0.2806	0.8880	0.2562	0.9050
4	36	0.2365	0.9018	0.2457	0.9050
5	37	0.2000	0.9202	0.2214	0.9191
6	41	0.1747	0.9340	0.2120	0.9294
7	36	0.1509	0.9441	0.2225	0.9153
8	37	0.1518	0.9364	0.2146	0.9230
9	35	0.1213	0.9521	0.2020	0.9268
10	36	0.1115	0.9571	0.1994	0.9281

3) Ensemble model's performance analysis: Table IV summarizes the Ensemble model's performance over ten epochs, with corresponding accuracy and loss graphs shown in Fig. 10. The model effectively aggregates multiple classifiers to boost prediction accuracy and reduce generalization error. Training and validation losses consistently decrease, though validation loss shows minor swings, suggesting potential overfitting. Accuracy reaches 99% for training, but validation accuracy varies slightly, indicating room for improvement. Precision and recall graphs in Fig. 11 reveal high precision and steadily improving recall, with validation trends reflecting minor fluctuations. The F1-score and ROC curve in Fig. 12 offer a comprehensive performance view, showing the model's balance between precision and recall and strong discriminative ability. The ROC curve achieves near-perfect AUC scores of 1.00 for both Normal and OSCC classes, highlighting the model's effectiveness in classification with minimal errors, and demonstrating strong reliability.

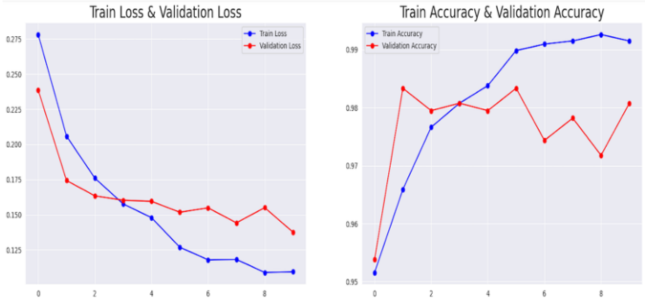


Fig. 10. Loss and accuracy analysis of ensemble model.

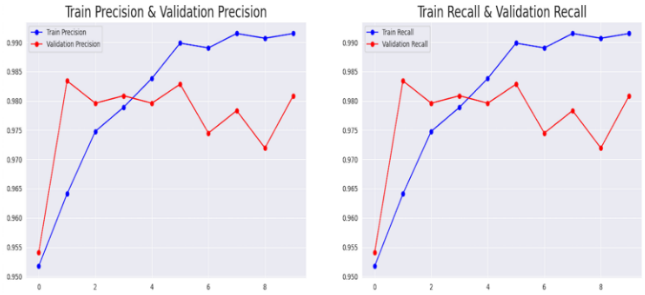


Fig. 11. Precision and recall analysis of ensemble model.

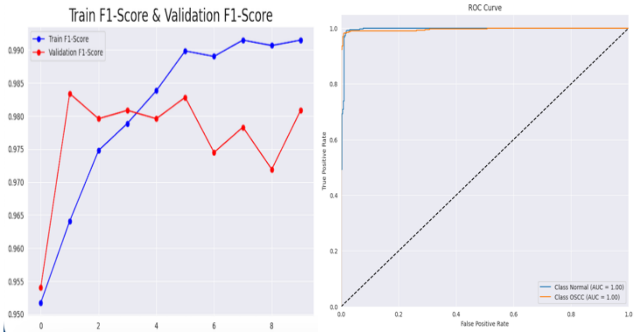


Fig. 12. F1-score and ROC analysis of ensemble model.

Table IV shows the performance measures per epoch, demonstrating constant loss and accuracy improvement. The Ensemble model's efficiency in this classification problem is shown by its ability to preserve great accuracy while reducing loss over epochs.

TABLE IV. ENSEMBLE MODELS PERFORMANCE PER EPOCH

Epoch	Time(sec)	Loss	Accuracy	Validation Loss	Validation Accuracy
1	138	0.1015	0.9854	0.1266	0.9718
2	62	0.0768	0.9907	0.1210	0.9769
3	62	0.0721	0.9923	0.1068	0.9807
4	62	0.0664	0.9934	0.1081	0.9820
5	62	0.0646	0.9951	0.1081	0.9820
6	62	0.0634	0.9942	0.1085	0.9820
7	62	0.0592	0.9940	0.1113	0.9795
8	62	0.0620	0.9926	0.1086	0.9782
9	62	0.0620	0.9926	0.1086	0.9782
10	62	0.0545	0.9934	0.0936	0.9835

B. *Confusion Matrix*

The three models' performance EfficientNetB3, ResNet50, and the Ensemble Model is evaluated here based on the confusion matrices in Fig. 13, 14, and 15. In Fig. 13 the EfficientNetB3 model displays excellent detection performance, properly classifying 385 Normal instances and 384 OSCC cases. However, it misclassifies 9 Normal cases as OSCC and 21 OSCC cases as Normal, demonstrating a minor bias toward misidentifying OSCC as Normal. Despite these flaws, the model exhibits a high level of accuracy overall.

In Fig. 14, the ResNet50 model accurately detects 349 Normal cases and 363 OSCC cases. With 25 Normal instances mistakenly categorized as OSCC and 42 OSCC cases incorrectly identified as Normal, it reveals a higher number of misclassifications than EfficientNetB3. This suggests that although ResNet50 performs well, it finds greater difficulty differentiating the two groups than EfficientNetB3.

Among the three, the Ensemble model in Fig. 15, which integrates the capabilities of EfficientNetB3 and ResNet50 showcases the best performance. With only three Normal cases misclassified as OSCC and eight OSCC cases misclassified as Normal, it correctly classifies 307 Normal cases and 369 Normal cases. The much smaller number of misclassifications implies that the Ensemble model efficiently lowers the mistakes observed in the individual models, hence producing better general accuracy.

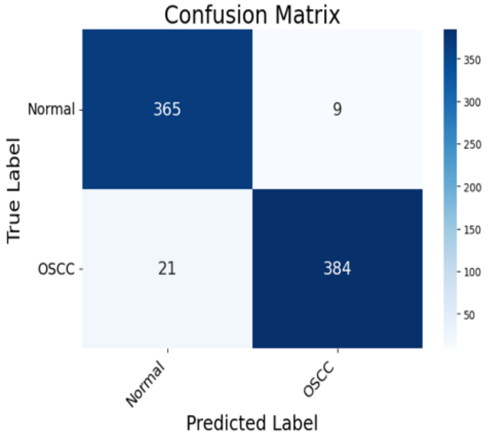


Fig. 13. Confusion matrix of EfficientNetB3.

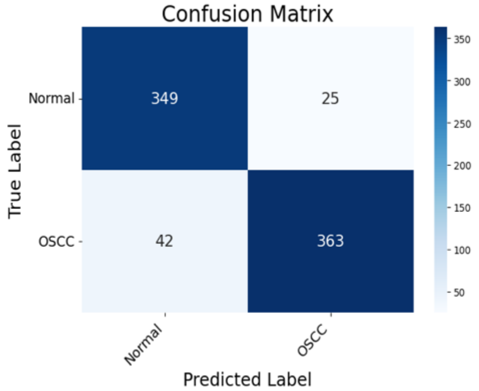


Fig. 14. Confusion matrix of ResNet50.

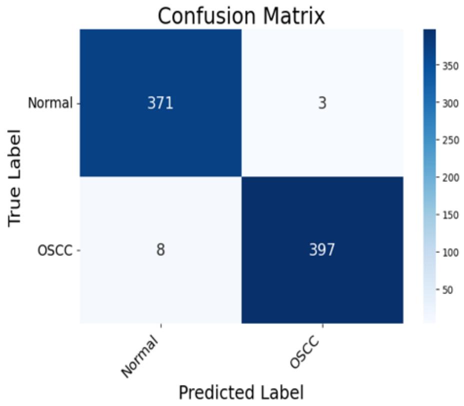


Fig. 15. Confusion matrix of ensemble model.

C. Performance Parameters

Table V presents an extensive three-model performance parameter comparison- precision, recall, F1-score, support, and accuracy across EfficientNetB3, ResNet50, and the Ensemble Model. EfficientNetB3 shows great precision and recall reflecting its great capacity to properly identify positive instances with low false positives and missed real positives. ResNet50 struggles more in differentiating between classes and demonstrates somewhat lower precision and recall even if it is still efficient. Though ResNet50 trails somewhat behind EfficientNetB3, the F1-scores for both models show a reasonable mix of precision and recall. Reflecting its better capacity to balance precision and recall while minimizing classification mistakes, the Ensemble model beats both individual models by obtaining the highest precision, recall, and F1 score. Furthermore, the Ensemble model has the best accuracy, which emphasizes the fact that it can appropriately classify most of the cases. All models show consistent support that guarantees these measures fairly represent performance over a balanced dataset. Combining the strengths of EfficientNetB3 and ResNet50 to provide excellent classification performance, the Ensemble model shows overall to be the most dependable and efficient classifier.

D. State-of-the-Art Comparison

Table VI presents a state-of-the-art comparison of the proposed Ensemble model with EfficientNetB3, ResNet50, and other top approaches in OSCC detection. The Ensemble model outperforms in F1 scores and general accuracy by aggregating model strengths, setting a new benchmark in medical imaging classification tasks.

TABLE V. PERFORMANCE PARAMETERS OF EFFICIENTNETB3, RESNET50 AND ENSEMBLE MODEL

Model	Classes	Precision	Recall	F1-Score	Support	Accuracy
EfficientNetB3	Normal	0.95%	0.98%	0.96%	374	0.96%
	OSCC	0.98%	0.95%	0.96%	405	
ResNet 50	Normal	0.89%	0.93%	0.91%	374	0.91%
	OSCC	0.94%	0.90%	0.92%	405	
Ensemble Model	Normal	0.98%	0.99%	0.99%	374	0.99%
	OSCC	0.99%	0.98%	0.99%	405	

TABLE VI. STATE-OF-THE-ART COMPARISON

Reference No.	Image Type	Technique	Images Count	Accuracy
(2024) [16]	CT images	Inception-V3	499	73.50%
(2024) [17]	Histopathology images	DenseNet121	5192	97.02%
(2024) [18]	Raman Histology	VGG-19	21,703	0.90%
(2024) [19]	Histopathological images	EfficientNetB3	1224	99%
(2024) [20]	Histological images	InceptionV3,Xception,InceptionResNetV2, NASAnet	5685	89.3%,89.5%,91.78%,90.8%
(2024) [21]	Histological images	Convolutional Neural Network	5192	98.49%, 86.89%, 89.37%
(2024) [22]	histopathological images	VGG16,VGG19,InceptionV3, AlexNet, ResNet50	1224	84%,82%,67%,76%,42%
(2024) [23]	Smartphone based images	Deep Learning	760	84.3%
(2024) [24]	Histological images	Deep Learning Models	4946	97.78%
(2024) [25]	histopathological images	DL-CNN	1224	91.1%
(2024) [26]	histopathological images	Deep Learning Models	696	0.92%
(2023) [27]	Histological images	AlexNet	5192	89%
(2024) [28]	Histological images	AI based approaches	5192	94.92%
(2023) [29]	histopathological images	MobileNetV3, InceptionV2, EfficientNetB3	5192	0.89%,0.88%,0.52%
(2023) [30]	Histopathological images	DL-CNN models	1224	91.25%
Proposed Model	Histopathological images	Ensemble Learning	5192	99.34%

VI. Conclusion

This work efficiently applied ensemble learning and transfer learning approaches for the identification of Oral Squamous Cell Carcinoma (OSCC) from histopathology photos. By combining the best features of pre-trained models such as EfficientNetB3 and ResNet50 via ensemble learning, the proposed approach shows astonishing accuracy and robustness. Outliving individual models, the last ensemble model unequivocally displayed increasing generality and accuracy. The detailed study of the performance of the deep learning models in medical image processing highlights its promise since it reveals how to routinely raise accuracy and reduce loss during training and validation datasets. Especially the EfficientNetB3 model has shown remarkable performance measures, thereby addressing the problems of early OSCC detection as a required need for timely and effective treatment. The requirement of incorporating modern deep-learning architectures with ensemble learning techniques to provide reliable and efficient diagnostic tools is underlined in this work. Apart from improving detection accuracy, the proposed method provides a structure suitable for other kinds of cancer and medical image analysis uses. Using more varied datasets and investigating other deep-learning approaches will help to improve the performance of the model, hence producing strong, real-time diagnostic systems for clinical use.

VII. Limitations and Future Scope

Although the suggested transfer learning and ensemble learning method greatly increased the accuracy of OSCC detection, some constraints have to be mentioned. First of all, compared to larger-scale medical imaging datasets, the dataset size is somewhat tiny even if it is enough for the present work. This may restrict the model's generalizability to unprocessed data from many populations. Second, depending on pre-trained models such as EfficientNetB3 and ResNet50 implies that more specific medical datasets for OSCC could help to improve the model's performance even further. Including a bigger, more varied dataset in the next projects could help the model to be more resilient. A deeper understanding of model predictions may also come from investigating more sophisticated ensemble methodologies and including explainable artificial intelligence approaches. Another important development of this work would be implementing the model in actual clinical environments and verifying its efficacy among several institutions.

References

[1] S. Y. Yang et al., “Histopathology-based diagnosis of oral squamous cell carcinoma using deep learning,” J. Dent. Res., vol. 101, no. 11, pp. 1321–1327, 2022.

[2] C. Carreras-Torras and C. Gay-Escoda, “Techniques for early diagnosis of oral squamous cell carcinoma: Systematic review,” Med. Oral Patol. Oral Cir. Bucal, vol. 20, no. 3, pp. e305-15, 2015.

[3] S. R. Larsen, J. Johansen, J. A. Sørensen, and A. Krogdahl, “The prognostic significance of histological features in oral squamous cell carcinoma,” J. Oral Pathol. Med., vol. 38, no. 8, pp. 657–662, 2009.

[4] A. K. Shaw et al., “Diagnostic accuracy of salivary biomarkers in detecting early oral squamous cell carcinoma: A systematic review and meta-analysis,” Asian Pac. J. Cancer Prev., vol. 23, no. 5, pp. 1483–1495, 2022.

[5] B. F. Adeyemi and B. Kolude, “Clinical presentation of oral squamous cell carcinoma,” Niger. Postgrad. Med. J., vol. 20, no. 2, pp. 108–110, 2013.

[6] L. Nokovitch et al., “Oral cavity squamous cell carcinoma risk factors: State of the art,” J. Clin. Med., vol. 12, no. 9, 2023.

[7] C. Seethalakshmi, “Early detection of oral squamous cell carcinoma (OSCC)-Role of genetics: A literature review,” Journal of clinical and diagnostic research, vol. 7, no. 8, 2013.

[8] P. Hollows, P. G. McAndrew, and M. G. Perini, “Delays in the referral and treatment of oral squamous cell carcinoma,” Br. Dent. J., vol. 188, no. 5, pp. 262–265, 2000.

[9] J. J. Sciubba, “Oral cancer: the importance of early diagnosis and treatment,” American journal of clinical dermatology, vol. 2, pp. 239–251, 2001.

[10] C. E. Palme, P. J. Gullane, and R. W. Gilbert, “Current treatment options in squamous cell carcinoma of the oral cavity,” Surg. Oncol. Clin. N. Am., vol. 13, no. 1, pp. 47–70, 2004.

[11] D. K. Das, S. Bose, A. K. Maiti, B. Mitra, G. Mukherjee, and P. K. Dutta, “Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis,” Tissue Cell, vol. 53, pp. 111–119, 2018.

[12] S. Arora, A. Matta, N. K. Shukla, S. V. S. Deo, and R. Ralhan, “Identification of differentially expressed genes in oral squamous cell carcinoma. Molecular Carcinogenesis,” vol. 42, pp. 97–108, 2005.

[13] M. Das, R. Dash, and S. K. Mishra, “Automatic detection of oral squamous cell carcinoma from histopathological images of oral mucosa using deep convolutional neural network,” Int. J. Environ. Res. Public Health, vol. 20, no. 3, 2023.

[14] R. Marzouk et al., “Deep transfer learning driven oral cancer detection and classification model,” Comput. Mater. Contin., vol. 73, no. 2, pp. 3905–3920, 2022.

[15] S. Krishna, J. Lavanya, G. Kavya, N. Prasamya, and Swapna, “Oral cancer diagnosis using deep learning for early detection,” in 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022.

[16] A. Fanizzi et al., “Explainable prediction model for the human papillomavirus status in patients with oropharyngeal squamous cell carcinoma using CNN on CT images,” Sci. Rep., vol. 14, no. 1, p. 14276, 2024.

[17] M. E. Paramasivam, B. S. Sriganesh, and S. Sureshkrishna, “Oral cancer detection using convolutional neural network,” in 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), 2024.

[18] A. Weber et al., “AI-based detection of oral squamous cell carcinoma with Raman Histology,” Cancers (Basel), vol. 16, no. 4, p. 689, 2024.

[19] E. Albalawi et al., “Oral squamous cell carcinoma detection using EfficientNet on histopathological images,” Front. Med. (Lausanne), vol. 10, p. 1349336, 2023.

[20] K. V. Kumar, S. Palakurthy, S. H. Balijadaddanala, and S. Reddy, “Early Detection and Diagnosis of Oral Cancer Using Deep Neural Network,” Journal of Computer Allied Intelligence, vol. 2, no. 02, pp. 22–34, 2024.

[21] A. Mishra, K. Srinivas, and A. Charan Kumari, “Empowering Oral Squamous Cell Carcinoma detection with deep learning: Insights from convolutional neural network analysis of histopathological images,” J. Tr., Chal. Art. Intell., vol. 1, no. 2, pp. 45–50, 2024.

[22] M. U. A. Siddique, S. M. Rabha, J. Periwal, N. Choudhury, and R. Mandal, “Early detection of oral cancer using image processing and computational techniques,” in Proceedings of the NIELIT’s International Conference on Communication, Electronics and Digital Technology, Singapore: Springer Nature Singapore, 2024, pp. 37–54.

[23] E. S. Mira et al., “Early diagnosis of oral cancer using image processing and artificial intelligence,” Fusion: Practice and Applications, vol. 14, no. 1, pp. 293–308, 2024.

[24] B. S. Deo, M. Pal, P. K. Panigrahi, and A. Pradhan, “Supremacy of attention-based transformer in oral cancer classification using histopathology images,” Int. J. Data Sci. Anal., 2024.

[25] A. Soni, P. K. Sethy, A. K. Dewangan, A. Nanthaamornphong, S. K. Behera, and B. Devi, “Enhancing oral squamous cell carcinoma detection:

a novel approach using improved EfficientNet architecture," BMC Oral Health, vol. 24, no. 1, p. 601, 2024.

[26] B. S. Deo, M. Pal, P. K. Panigrahi, and A. Pradhan, "An ensemble deep learning model with empirical wavelet transform feature for oral cancer histopathological image classification," Int. J. Data Sci. Anal., 2024.

[27] T. Saraswathi and V. M. Bhaskaran, "Classification of Oral Squamous Carcinoma Histopathological images using Alex Net," in 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), IEEE, 2023, pp. 637–643.

[28] M. Ahmad, M. I. Khattak, A. Jan, and I. U. Haq, A Novel Hybrid AI-Based System for Early Detection of Oral Squamous Cell Carcinoma via Histopathological Images.

[29] B. Nagarajan et al., "A deep learning framework with an intermediate layer using the swarm intelligence optimizer for diagnosing oral squamous cell carcinoma," Diagnostics (Basel), vol. 13, no. 22, p. 3461, 2023.

[30] S. H. Begum and P. Vidyullatha, "Deep Learning Model for Automatic Detection of Oral squamous cell carcinoma (OSCC) using Histopathological Images," International Journal of Computing and Digital Systems, 2023.

[31] A. F. Kebede, "Histopathologic Oral Cancer Detection using CNNs." 21-Jul-2021.

[32] B. Ananthakrishnan, A. Shaik, S. Kumar, S. O. Narendran, K. Mattu, and M. S. Kavitha, "Automated detection and classification of oral squamous cell carcinoma using deep neural networks," Diagnostics (Basel), vol. 13, no. 5, p. 918, 2023.

[33] S. Bharany and S. Sharma, "Intelligent green internet of things: An investigation," in Machine Learning, Blockchain, and Cyber Security in Smart Environments. Chapman and Hall/CRC, 2022, pp. 1–15.

[34] S. Badotra et al., "A DDoS Vulnerability Analysis System against Distributed SDN Controllers in a Cloud Computing Environment," Electronics, vol. 11, no. 19. MDPI AG, p. 3120, Sep. 29, 2022. doi: 10.3390/electronics11193120.

[35] K. Kaushik et al., "Multinomial Naive Bayesian Classifier Framework for Systematic Analysis of Smart IoT Devices," Sensors, vol. 22, no. 19. MDPI AG, p. 7318, Sep. 27, 2022. doi: 10.3390/s22197318.

[36] S. Bharany, K. Kaur, S. E. M. Eltaher, A. O. Ibrahim, S. Sharma, and M. M. M. A. Elsalam, "A Comparative Study of Cloud Data Portability Frameworks for Analyzing Object to NoSQL Database Mapping from ONDM's Perspective," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10. The Science and Information Organization, 2023. doi: 10.14569/ijacsa.2023.0141086.

Furniture Panel Processing Positioning Design Based on 3D Measurement and Depth Image Technology

Binglu Chen[1]*, Guanyu Chen[2], Qianqian Hu[3]
Institute of Science and Art, University of Wales Trinity Saint David, Swansea, SA1 6ED, United Kingdom[1]
School of Cultural Relics and Arts, Hebei Oriental University, Langfang, 065001, China[2]
School of Art and Design, Zhengzhou University of Light Industry, Zhengzhou, 450000, China[3]

***Abstract*—In recent years, furniture panel processing positioning based on computer vision technology has received increasing attention. A 3D measurement imaging technology based on laser scanning technology is proposed to address the significant environmental impact of traditional visual technology. Subsequently, deep image processing techniques are introduced to address the high image noise. In the experiment of measuring panel using 3D measurement technology, 14 measurement lines were taken every 10mm of the measurement length. The maximum measurement value was 204.62mm, the minimum measurement value was 204.37mm, and the manual measurement result was 204.5mm. 14 measurement lines were taken every 14mm of the measurement length. The maximum measurement value was 134.15mm, the minimum measurement value was 133.894mm, and the manual measurement was 134.1mm. 14 measurement lines were taken every 14mm of the measurement thickness. The maximum measurement value was 26.646mm, the minimum measurement value was 26.242mm, and the result of manual measurement was 26.5mm. The 3D imaging technology based on laser scanning is relatively accurate in measuring the 3D data of panels, which can be applied in the positioning and detection system of panel processing. In addition, the experiment compares depth image processing methods, verifying the effectiveness of the designed method. Meanwhile, this research also has certain reference significance for exploring the real-time positioning of other objects.**

Keywords*—Laser scanning; 3D measurement; deep image processing; positioning; computer vision

I. Introduction

In the current modern industrial automation and intelligent production, object positioning technology has become a key link. The stability and accuracy of positioning technology are crucial for production efficiency and product quality [1-2]. However, in actual positioning, object positioning systems often encounter anomalies, leading to a decrease in positioning accuracy or even failure. Common positioning faults include hardware failure, software failure, and environmental impact [3]. Among them, environmental factors are an important factor leading to abnormal visual positioning. For example, temperature, humidity, vibration, and other factors at the production site can affect the stability and accuracy of equipment [4-6]. In addition, external light sources, electromagnetic interference, etc. may also cause interference to the visual positioning system, leading to positioning failure. Object positioning technology is of great significance for furniture automation manufacturing. The pursuit of quality of life is also increasing day by day. In the home environment, furniture, as an important component of life, is increasingly attracting consumer attention in terms of quality, style, personalization, and other aspects. To meet consumer demand, the Chinese furniture market is undergoing profound changes, gradually developing towards high-end, branded, intelligent, and personalized directions [7-8]. In this process, the furniture panel industry plays an important role. Furniture panels are not only the basic materials that make up furniture, but also a key factor affecting the quality, appearance, and safety of furniture. A high-quality board not only gives furniture a beautiful appearance but also ensures its durability, providing consumers with a safe and reliable user experience. With the continuous expansion of the high-end furniture market, the furniture panel industry will also usher in more business opportunities. Meanwhile, this also puts higher requirements on the technical level, innovation ability, and product quality of furniture panel enterprises. In order to strengthen the automation and intelligent production of furniture panels in enterprises, more practitioners have added object positioning technology to the processing of furniture panels.

The scanning imaging and image preprocessing of objects exerts a crucial role in the precise positioning of objects. Yin T et al. seamlessly integrated RGB sensors into lidar-based 3D recognition to address the low resolution of lidar sensors. The proposed framework significantly improved the strong center point baseline and outperformed competitive fusion methods [9]. Jiang W et al. proposed a new high-speed 2D and 3D imaging scheme on the basis of traditional single-pixel imaging, which was difficult to achieve high imaging speed using traditional single-pixel imaging mechanisms. The new scheme performed 2D and 3D imaging on a rotating chopper with a speed of 4800rpm, with an imaging speed of up to two million fps, providing a powerful solution for high-speed imaging [10]. Pan X et al. built a deep generative prior method to address the difficulty in capturing rich image semantics. The results showed that this method helped to preserve the reconstruction information, resulting in more accurate and faithful reconstruction results of real images [11]. To optimize the limits of plug and play image restoration, Zhang K et al. inserted deep denoising priors as module parts into iterative algorithms ground on semi quadratic splitting to solve various image restoration problems. The results showed that the designed strategy with deep denoising prior was not only significantly superior to other model-based methods, but also more competitive [12]. Guo Y et al. proposed a fully variational depth network based on deep learning and high-resolution color images to address the edge misalignment and depth discontinuity in color images used for guiding depth image reconstruction. Under the guidance of high-resolution colors, the restored depth image was closest to the ground truth on the edge contours [13]. To assess the position and application of lightweight panels in Iranian furniture production, Khojasteh-Khosro S et al. built an analysis method on the basis of consumer behavior and preferences. The main evaluation criteria for purchasing furniture were product quality, design, price, environmental protection, and guaranteed functionality [14]. Thio V et al. proposed a new step detection strategy relying on the sine wave approximation of acceleration signals to address the noise and infinitely increasing position errors in sensors embedded in smart devices. This method could detect step scores and complete steps, thereby achieving continuous and real-time updates. The results demonstrated the feasibility of using the sine wave approximation method for step detection, and the expected benefits of integrating pedestrian prediction positioning with ultrasound systems [15]. To solve the difficulty in accurately identifying various types of defects in glued wood panels, Chen L C et al. adopted computer vision and deep learning to identify low, high, normal, long, and short defects. The new method achieves higher accuracy than other methods with excessive killing and escape rate analysis [16]. To address the enormous demand for low-cost and high-precision positioning and navigation solutions in emerging IoT applications, Feng D et al. built an indoor positioning system that combined inertial measurement units with ultra-wide bands by extending Kalman filters and unscented Kalman filters. The results indicated that the prior information provided by the inertial measurement unit could significantly suppress ultra wide-band observation errors [17]. In response to the problem that the currently released dataset for crowd counting and localization is too small to meet the requirements of supervised convolutional neural network algorithms, Wang Q et al. built a large-scale crowded crowd counting and localization dataset NWPU-Crowd, consisting of 5109 images and 2133375 annotated headers with dots and boxes. The results showed that the new dataset could meet the requirements of supervised convolutional neural network algorithms [18].

In summary, domestic and foreign researchers have put forward many discussions on the positioning of furniture panel processing, including the three-dimensional measurement technology. However, few scholars have combined 3D measurement technology based on laser scanning technology with depth image processing technology. Compared with the methods used in the study, although these methods may be enhanced in certain directions, they are still relatively weak in continuous and large-scale image processing. In response to this issue, a furniture panel processing and positioning design based on 3D measurement and depth image technology is proposed, aiming to apply different technologies to different stages of panel positioning to achieve optimal positioning results. Compared with other works in the same direction, the method proposed in the study has higher processing efficiency for sheet images due to the use of mature methods that have been validated for a long time, thus meeting the requirements for processing a large number of images. The innovation of the research lies in the combination of 3D measurement imaging technology and depth image processing technology, which helps to better locate the position of the panel.

II. Methods and Materials

Aiming at the high-precision measurement problem required for furniture panel processing positioning, a 3D measurement imaging technology based on laser scanning is introduced. On this basis, deep image processing technology is applied to process the images scanned by 3D measurement imaging technology, providing a position positioning basis for furniture panel processing.

A. 3D Measurement Imaging Technology Based on Laser Scanning

Currently, 3D measurement technology has been applied in many fields. This study adopts a 3D measurement technology based on laser scanning technology. The basic principle of laser measurement is to use the reflection of light for measurement. After the laser beam is emitted onto the object being measured, a portion of the light is reflected by the object, while another portion of the light is absorbed or scattered by the object. A laser measuring instrument obtains relevant information about the measured object by measuring the reflected light [19-20]. Laser scanning measurement utilizes an external motion platform to continuously scan laser lines through the target surface, thereby obtaining the overall contour of the object. To achieve real-time measurement of moving target objects, the laser triangulation method is used for measurement. The basic principle of laser triangulation is to irradiate a beam of laser at a specific angle onto the surface of the measured object, and then observe the laser spot on the measured object from a different angle. The position of the light spot varies with the height of the measured object [21-22]. The photodetector measures the position of the spot imaging, which can determine the height of the laser irradiation point on the object. The laser triangulation measurement technology utilizes optical reflection laws and the principle of similar triangles. Its vertical measurement is shown in Fig. 1.

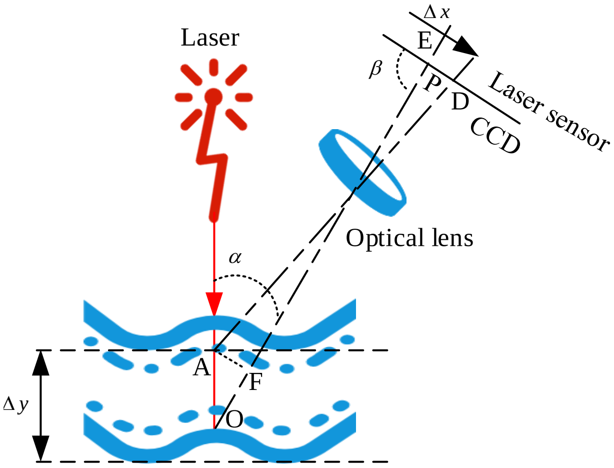


Fig. 1. Schematic diagram of laser vertical measurement.

In Fig. 1, A represents the contact point between the laser and the upper surface of the object, and O represents the contact point between the laser and the lower surface of the object. AP and AD represent the reflected light rays formed by a lens after being reflected by an object. Assuming the initial position of the object surface is in a certain plane, when the distance △ y moves to the lower surface, the incident point changes from position A to point O, and the corresponding image point changes from point D to position P. From the changes in the optical path structure, it can be inferred that the incidence points at different heights correspond to different imaging point positions. Based on this relationship, the spatial position information of any point on the surface of an object can be obtained by analyzing the geometric relationship between the image point and the incident point. Eq. (1) can be obtained from the similarity between triangle △AFB and △BDE.

$$\frac{AF}{DE}=\frac{FB}{BE} \tag{1}$$

In Eq. (1), A is the projection point of the laser axis on the object. D is the corresponding pixel. F,E are perpendicular feet of the vertical line that reflects light at point A and point D. B is the location of the intersection point between two lasers, that is, the intersection point between the laser and the optical lens. Eq. (2) can be obtained by transforming Eq. (1).

$$\frac{AO\cdot\sin\alpha}{DP\cdot\sin\beta}=\frac{OB-AO\cdot\cos\alpha}{BP+DP\cdot\cos\beta} \tag{2}$$

In Eq. (2), O signifies the intersection point between the vertical injection of the laser and the lower surface of the object. P is the image point position of the laser reflected light on the laser sensor. α is the angle between AO and OP. β is the angle formed between OP and the linear CCD after passing through the lens. Eq. (2) can be transformed to obtain Eq. (3).

$$AO=\frac{OB\times DP\times\sin\alpha}{BP\times\sin\alpha+DP\times\sin(\alpha+\beta)} \tag{3}$$

In oblique measurement, Eq. (4) can be obtained according to the ray projection relationship and the principle of similar triangles.

$$\frac{AF}{\sin\alpha}\times\frac{1}{DP\times\sin\beta}=\frac{OB}{BP+DP\times\cos\beta} \tag{4}$$

In Eq. (4), AF is the height of the incident point on the object. DP is the displacement between pixels. D and F are the image points corresponding to the positions of point A and point O. OB is the imaging objective distance of point O. BP is the image distance. Eq. (4) can be transformed to obtain Eq. (5).

$$AF=\frac{DP\times\sin\beta\times OB\times\sin\alpha}{BP+DP\times\cos\beta} \tag{5}$$

Laser triangulation scanning imaging technology is widely used in automated production. However, in actual factory operations, there are often problems that cannot detect and recognize objects. There are many reasons for such problems, as shown in Fig. 2.

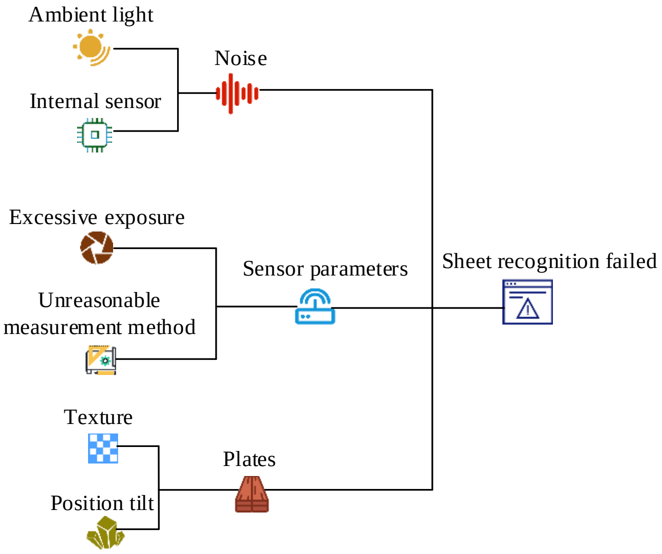


Fig. 2. System influencing factors.

Fig. 2 shows the factors that the system cannot detect and recognize objects, mainly including external noise, sensor working parameter settings, board position and texture, etc. In addition, the board position detection system may also fail to effectively scan and recognize objects, as shown in Fig. 3.

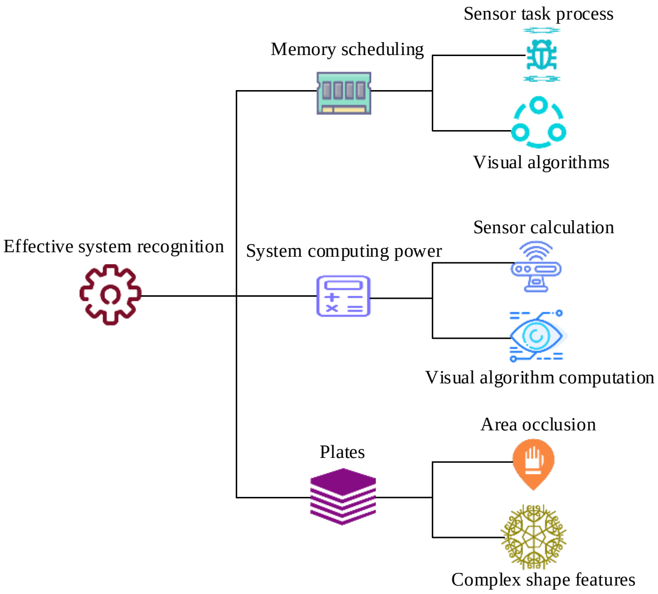


Fig. 3. The system logic.

Fig. 3 shows the factors that affect the effective identification of the board detection system. Among them, the board position detection system needs to integrate several functional requirements in the software composition structure, including hardware communication, visual processing, robot control, sensor data exchange, etc.

B. Design of Furniture Panel Positioning and Detection System Based on Depth Image Technology

Due to the interference of external environment during image collection, the collected board images often contain some noise and have poor image quality. Therefore, after the panel is scanned into images through three-dimensional measurement, it is necessary to preprocess the images [23]. Several processing algorithms for image enhancement and filtering denoising are studied, and suitable depth image preprocessing algorithms are selected based on their processing effects, effectively improving the plate recognition accuracy, as displayed in Fig. 4.

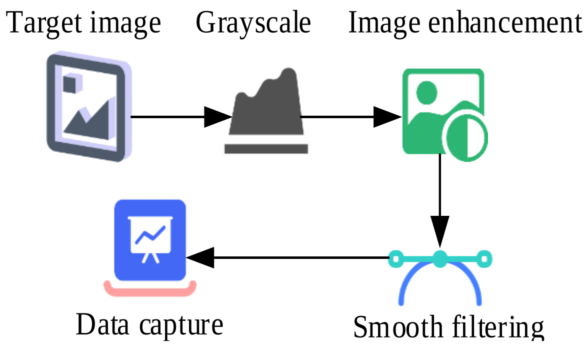


Fig. 4. Image processing and target extraction process.

Fig. 4 shows the preprocessing and target area extraction process of the panel image after three-dimensional measurement. In Fig. 4, image preprocessing is mainly aimed at suppressing or eliminating irrelevant signals in the image, enhancing the true information such as edges in the image. The processing methods include filtering denoising, contrast enhancement, etc. The scope of image enhancement processing is divided into spatial and frequency domain processing. The study adopts the spatial enhancement method. The spatial domain refers to the two-dimensional space of the image itself. This processing is performed on the image pixels, as shown in Eq. (6).

$$g(m,n)=T\left[f(m,n)\right] \tag{6}$$

In Eq. (6), (m,n) represents the current operating pixel point. $f(m,n)$ signifies the initial pixel grayscale value. $g(m,n)$ signifies the processed pixel grayscale value. T is the functional model of the processing method. Grayscale transformation is a common spatial domain image enhancement algorithm that can increase the grayscale distribution range of an image and make local details clearer. There are several commonly used enhancement methods. The image inversion method is shown in Eq. (7).

$$g(m,n)=L-T\left[f(m,n)\right]-1 \tag{7}$$

In Eq. (7), L represents the grayscale value of the original image. The linear transformation method is shown in Eq. (8).

$$g(m,n)=I_1-k\left[f(m,n)-M_1\right] \tag{8}$$

In Eq. (8), k represents the slope of the transformation function. I_1 represents the starting point of the processed pixel grayscale value interval. M_1 represents the starting point of the initial grayscale value interval for pixels. The slope k is shown in Eq. (9).

$$k=\frac{I_2-I_1}{M_2-M_1} \tag{9}$$

In Eq. (9), I_2 represents the endpoint of the processed pixel grayscale value interval. M_2 represents the endpoint of the initial grayscale value interval for pixels. To highlight the target of interest or certain regions in the image, segmented linear transformation can be used to expand the grayscale regions at different levels on the image, as shown in Eq. (10).

$$g(m,n)=\begin{cases}\dfrac{I_2-I_1}{M_2-M_1}\left[f(m,n)-M_1\right]+I_1 & M_1\le f(m,n)\le M_2\\ \dfrac{I_1}{M_1}\left[f(m,n)-M_1\right]+I_1 & 0\le f(m,n)\le M1\end{cases} \tag{10}$$

Another method for enhancing depth images of panel is histogram equalization. Histogram equalization is to adjust the probability distribution of pixels at each grayscale level to evenly distribute the overall grayscale range of pixels, in order to expand the dynamic range of pixel grayscale. Its expression is shown in Eq. (11).

$$T(j)=\int_0^j p_j(z)dz \tag{11}$$

In Eq. (11), $T(j)$ represents the probability density distribution function of j. z represents the integral variable. j represents the grayscale value. p_j represents the probability of processed pixels. The probability of pixels in each grayscale interval of the original image is shown in Eq. (12).

$$P(i)=\frac{n_i}{n} \tag{12}$$

In Eq. (12), $P(i)$ represents the probability of pixels in the original image. n_i represents the original image pixels. n represents the total pixels in the image. The grayscale cumulative histogram is shown in Eq. (13).

$$p_j=\sum_{i=0}^{L-1}P(i) \tag{13}$$

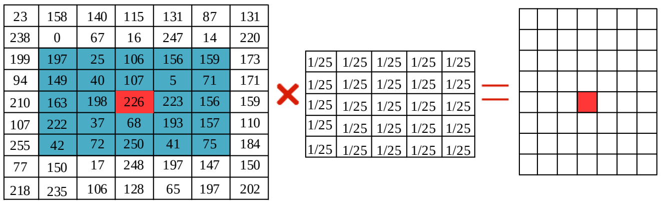
In Eq. (13), L signifies the pixel grayscale value . The grayscale value calculation is shown in Eq. (14).

$$j=(L-1)p_j \tag{14}$$

According to Eq. (11) to Eq. (14), the histogram after image processing can be obtained, as shown in Eq. (15).

$$P(j)=\frac{n_j}{n} \tag{15}$$

In Eq. (15), $P(j)$ is the probability density after image processing. n_j is the pixel point after image processing. During image acquisition, due to the low stability of the sampling system, it may be affected by external environmental interference, resulting in a significant amount of noise in the obtained image. To reduce the impact of these noises, the original image needs to be smoothed. In this study, mean filtering is used for denoising. It is a relatively simple and fast processing method in filtering algorithms, as shown in Fig. 5.



23	158	140	115	131	87	131
238	0	67	16	247	14	220
199	197	25	106	156	159	173
94	149	40	107	5	71	171
210	163	198	226	223	156	159
107	222	37	68	193	157	110
255	42	72	250	41	75	184
77	150	17	248	197	147	150
218	235	106	128	65	197	202

1/25	1/25	1/25	1/25	1/25
1/25	1/25	1/25	1/25	1/25
1/25	1/25	1/25	1/25	1/25
1/25	1/25	1/25	1/25	1/25
1/25	1/25	1/25	1/25	1/25

Fig. 5. Mean filtering denoising method.

Fig. 5 shows the mean filtering denoising method. The pixel value of any point in Fig. 5 is the mean of surrounding N/times M pixels, where the pixel value of the red point is the sum of the pixel values of the surrounding blue background area divided by 25. After image enhancement processing, a high-quality image is obtained. In subsequent processing, the target information is often only a part. For the entire image, only this part of the features need to be extracted and analyzed. The image obtained through image preprocessing can be used for panel detection and recognition. The detection system framework is shown in Fig. 6.

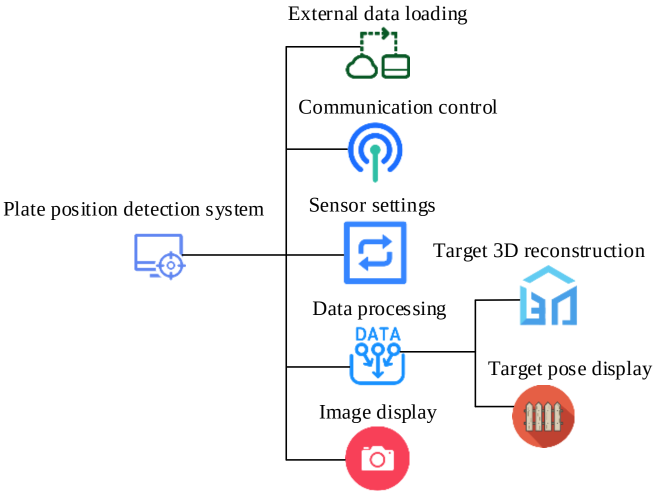


Fig. 6. Framework diagram of panel detection system.

Fig. 6 is the framework diagram of the board detection system. From Fig. 6, the system mainly includes several parts: external data loading, communication control, sensor settings, image acquisition and display, and depth image data processing. The interactive interface written by C#.NET integrates several functional modules of the software system, making the operation simple and efficient.

III. Results

To verify the effectiveness of the plate processing positioning and detection system, simulation analysis is conducted. Firstly, the relationship between laser sensors and imaging images in 3D measurement technology is analyzed. Subsequently, the depth image processing method is validated. Finally, the recognition accuracy of the plate positioning detection system is experimentally analyzed.

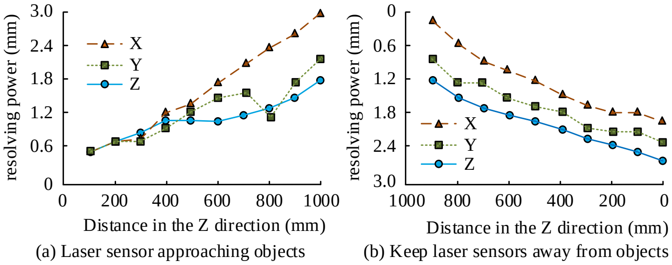
A. Analysis of Furniture Panel Processing Positioning Design Based on 3D Measurement Technology

This experiment is developed in the Visual Studio 2022 17.4 development environment, using C# as the programming language. The Halcon machine vision algorithm package processing function library and open graphics library 3D display tool are combined to build the system software framework. Table I displays the experimental parameter settings.

TABLE I. Experimental Parameter Settings

Parameter type	Index	Explain
Response time	10ms	/
Measuring range	250-650mm	X-axis
	200-1000mm	Z-axis
Laser wavelength	660nm	/
Laser line length	600×3mm	At a height of 850mm
Resolving power	0.6-1.8mm	X-axis
	1-4mm	Z-axis
Working voltage	20-35v	DC
Communication protocol	UDP	/

The images formed by 3D measurement technology are the basis for subsequent positioning and detection of panel processing. To verify the effectiveness of 3D measurement technology, experiments are conducted on various aspects of the technology, mainly focusing on the quantitative analysis between 3D measurement imaging technology on the basis of laser scanning and image formation. The relationship between the object distance of the sensor and the imaging resolution is shown in Fig. 7.



(a) Laser sensor approaching objects (b) Keep laser sensors away from objects

Fig. 7. The relationship between sensor distance and resolution.

In Fig. 7(a), when the laser sensor was close to the measured object, the resolution was higher. The resolution change speed in the X-direction was smaller than that in the Y and Z-axes. The length of the laser line in the X-direction depended on the distance Z between the measured object and the sensor. When the Z-axis distance was 1000mm, the laser line length reached the maximum value of 600mm, and the corresponding measurement resolution was also the lowest. In Fig. 7(b), the resolution decreased when the laser sensor was far away from the measured object. The resolution change speed in the X-direction was smaller than that in the Y and Z- axes. When the Z-axis distance was 900mm, which was the initial position, the corresponding measurement resolution was the highest. It indicates that when the sensor is further away from the target, it is easier to obtain more accurate images. The resolution is inversely proportional to the distance from the measurement object during measurement. The coordinate distribution of the measured object on the X and Z-axes is shown in Fig. 8.

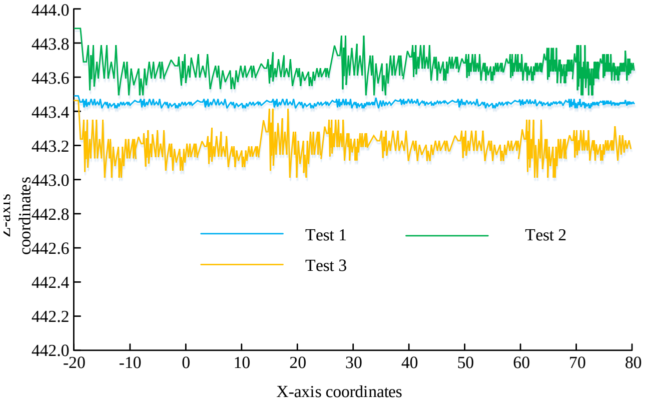


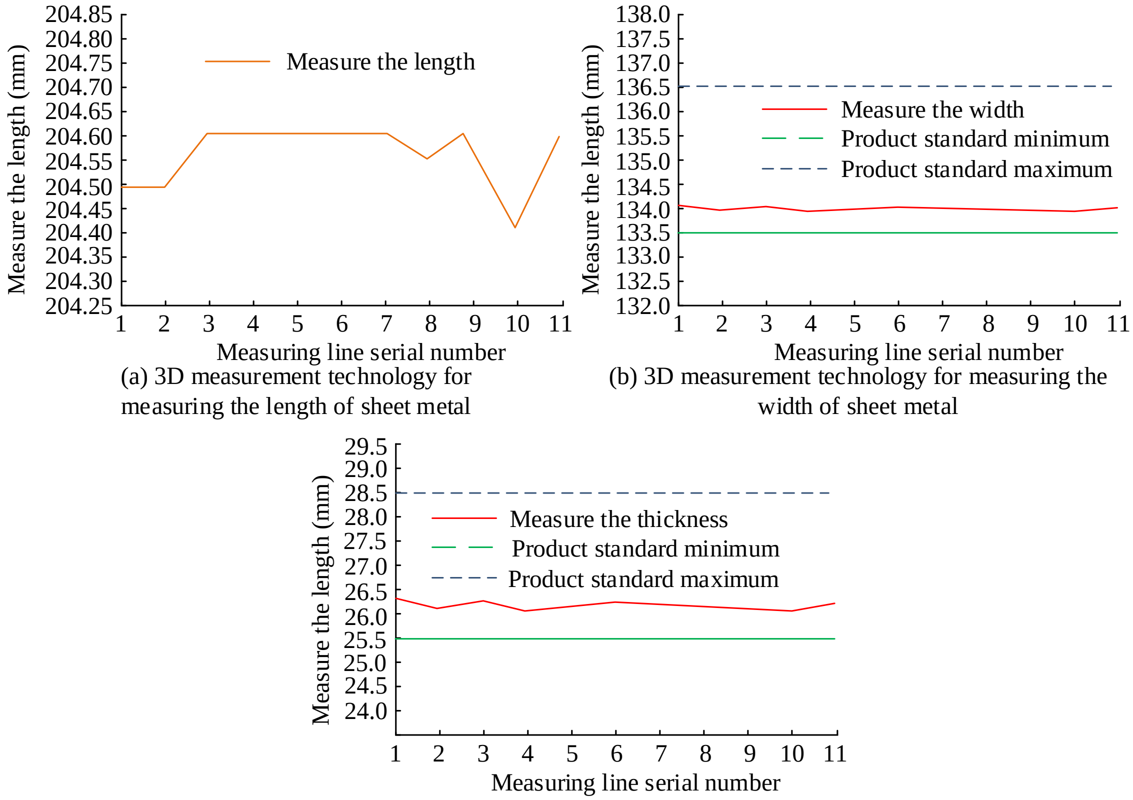
Fig. 8. Sensor single measurement test coordinates.

As shown in Fig. 8, the coordinate distribution of Test1, Test2, and Test3 on the Z-axis was relatively stable, with Test1 concentrated around 443.5, Test2 concentrated around 443.2, and Test3 concentrated around 443.7. The coordinate information on the X-axis and Z-axis is relatively complete, which can verify the connection status and good data measurement of the laser sensor. The actual measurement of panel using 3D measurement technology is shown in Fig. 9.

According to Fig. 9(a), 14 measurement lines were taken every 10mm interval. The maximum measurement value was 204.62mm, the minimum measurement value was 204.37mm, and the manual measurement result was 204.5mm. As shown in Fig. 9(b), a total of 14 measurement lines were taken every 14mm interval. The maximum measurement value was 134.15mm, the minimum measurement value was 133.894mm, and the manual measurement result was 134.1mm. As shown in Fig. 9(c), 14 measurement lines were taken every 14mm of the measurement thickness. The maximum measurement value was 26.646mm, the minimum measurement value was 26.242mm, and the manual measurement result was 26.5mm. The 3D imaging technology based on laser scanning is relatively accurate in measuring the 3D data of panel, which can be applied in the positioning and detection system of panel processing.

B. *Analysis of Furniture Panel Positioning and Detection System Based on Deep Image Technology*

Deep image processing technology can highlight the target area, suppress or eliminate irrelevant signals in the image, and enhance the true edge information in the image. Common processing techniques include grayscale transformation and histogram equalization. The latter is displayed in Fig. 10.

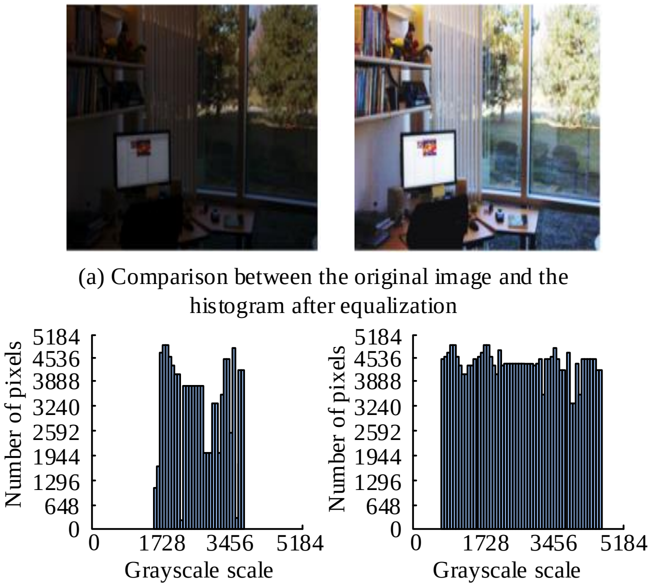


(a) 3D measurement technology for measuring the length of sheet metal

(b) 3D measurement technology for measuring the width of sheet metal

(c) 3D measurement technology for measuring the thickness of sheet metal

Fig. 9. 3D measurement technology for measuring the three-dimensional dimensions of panel.



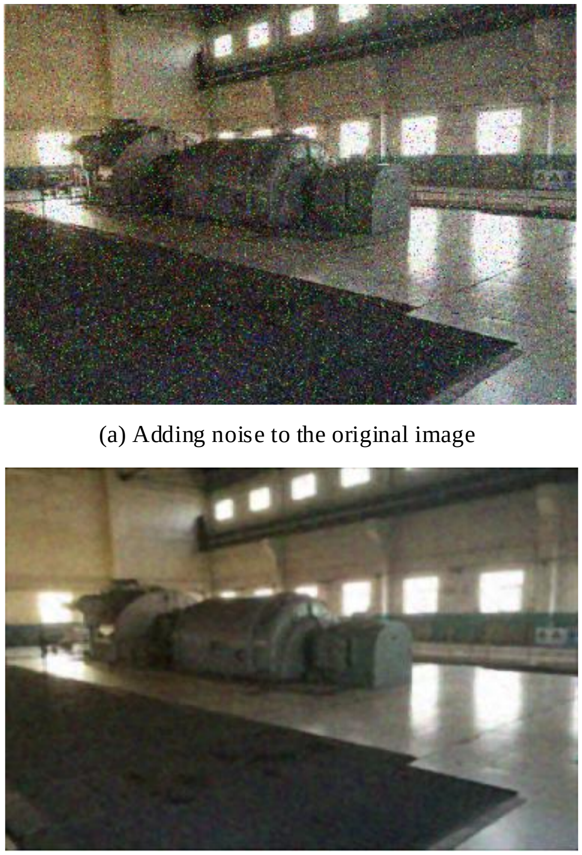
(a) Comparison between the original image and the histogram after equalization

(b) Comparison between initial histogram and equalization histogram

Fig. 10. Comparison chart of histogram equalization processing.

From Fig. 10(a), for some unclear detail features in the original image, after histogram equalization processing, the grayscale levels of these areas were richer. As shown in Fig. 10 (b), the histogram after equalization was greatly expanded. The original image had a grayscale range of 1720 to 3470, while the balanced image had a grayscale range of 700 to 5100. The grayscale range of areas with similar grayscale and occupying many pixels in the original image was widened, making small grayscale changes in large areas visible and making the image clearer. Experimental data shows that histogram equalization can enhance the visual effect of images and enhance the overall contrast of images. During image acquisition, due to the low stability of the sampling system, it may be affected by external environmental interference, resulting in much noise in the obtained image. This makes the subsequent processing and analysis of the image more complex. To reduce the impact of these noises, the original image needs to be smoothed and filtered. The mean filtering is used to process image noise, as shown in Fig. 11.

As shown in Fig. 11, the upper image is the original image with added noise, and the lower image is the image processed by mean filtering. The noise in the image below was significantly reduced. The experiments show that mean filtering can effectively remove noise. To analyze the positioning accuracy of the plate processing positioning detection system and verify the stability performance, the actual coordinate position under the sensor detection system is indirectly obtained by obtaining the position displayed by the robot at the specified position, and experimental data is measured accordingly. The data and statistical error distribution diagram are shown in Fig. 12.



(a) Adding noise to the original image

(b) Image processed by mean filtering with noise

Fig. 11. Comparison chart of mean filtering denoising.

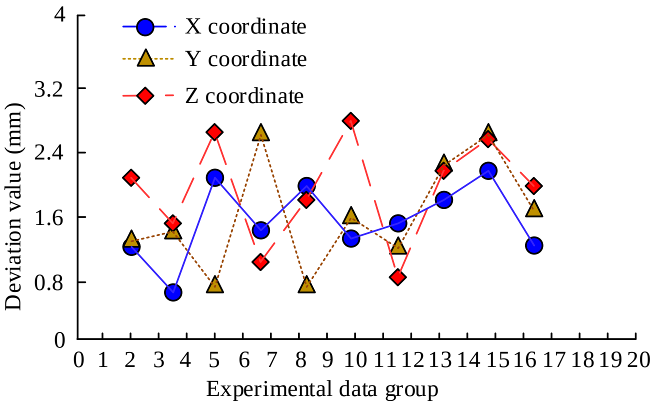


Fig. 12. Coordinate deviation diagram of the panel.

In Fig. 12, the maximum error of the experimental group in the X-axis was 2.2mm and the minimum was 0.6mm. The maximum error in the Y-axis was 2.6mm and the minimum was 0.8mm. The maximum error in the Z-axis was 2.8mm and the minimum was 0.8mm. The experimental data shows that the overall coordinate deviation of the experimental group can be controlled within 5mm, which can meet the positioning and detection requirements of the system.

IV. Discussion

In order to accurately locate furniture panels, a furniture panel processing and positioning system based on 3D measurement and depth imaging technology was studied and designed. The results showed that the experimental group of the detection system had a maximum error of 2.2mm and a minimum error of 0.6mm in the X direction. The maximum error in the Y direction is 2.6mm, and the minimum is 0.8mm. The maximum error in the Z direction is 2.8mm, and the minimum is 0.8mm. Similarly, scholars such as Fan J proposed the Chaos Cuckoo Search Algorithm (Ccsa) to solve image segmentation problems in computer vision and improve image accuracy. Unlike ordinary cuckoo search, chaotic mapping combines deterministic search and random verification. The experimental results show that compared with other existing methods, the proposed CCSA model improves accuracy and reduces uncertainty [21]. In contrast, the study uses grayscale transformation for image enhancement and employs mean filtering for denoising. From the perspective of image detection accuracy, the proposed method has slightly poor performance in image processing. The reason is that the method used in the study is greatly affected by image quality factors, and in the process of sheet metal processing, image quality is greatly affected by environmental factors such as surrounding sawdust and dust, resulting in relatively low image detection accuracy. Take a measurement line every 14mm and use 3D measurement technology for measurement. The maximum measurement result is 26.646mm, the minimum measurement value is 26.242mm, and the manual measurement result is 26.5mm. The research of Qian J and other scholars adopted a deep learning based color stripe projection contour technique for single absolute 3D shape measurement. The results show that this method can perform high-precision single frame absolute 3D shape measurement on complex objects [22]. This result is slightly higher than the 3D scanning imaging technology used in the study, indicating that the research can further optimize the scanning accuracy.

V. Conclusion

Traditional computer vision technology has a significant impact on furniture panel positioning due to environmental factors. Based on this, this study introduced furniture panel positioning based on 3D measurement and depth image processing technology. The results indicated that the resolution change rate of the image on the X-axis was smaller than that in the Y and Z-axes. When the Z-axis distance was 1000mm, the laser line length reached the maximum value of 600mm, and the corresponding measurement resolution was also the lowest. When the laser sensor moved away from the measured object, the resolution decreased, and the resolution change rate in the X-axis was slower than that in the Y and Z-axes. When the Z-axis distance was 900mm, which was the initial position, the corresponding measurement resolution was the highest. In addition, two sets of measurement tests were conducted on the object to observe its coordinate distribution. The coordinate distribution of Test1, Test2, and Test3 on the Z-axis was relatively stable, and the maximum error of the detection system in the X-axis was 2.2mm, and the minimum was 0.6mm. The maximum error in the Y-axis was 2.6mm and the minimum was 0.8mm. The maximum error in the Z-axis was 2.8mm and the minimum was 0.8mm. Experimental data showed that when the sensor was closer to the target, it was easier to obtain more accurate images. The resolution was inversely proportional to the distance from the measured object during measurement. The overall coordinate deviation of the experimental group was controlled within 5mm, which verified the effectiveness of the positioning detection system. Through the verification and analysis of the results, the design of furniture board processing positioning has been achieved, and the detection system can effectively identify and detect the board, which provides important technical support for the automation processing of the board. However, through the analysis of existing research results, it was found that although the image recognition detection error meets the requirements, there is still significant room for improvement. The proposed work aims to further address the factors that affect measurement accuracy in actual measurement operations, not only limited to environmental factors such as sawdust and dust, but also including scanning imaging hardware, system errors, and more advanced image processing technologies. In addition to upgrading hardware facilities, the most important aspect of the proposed work is to increase the universality of research. A detailed exploration was conducted on the positioning of furniture boards based on the processing environment, taking into account factors such as fiber optics and dust in the processing environment. However, there was insufficient exploration of object positioning in other environments. The future direction of work will further investigate the universality of the proposed method, such as verifying its localization ability under the influence of different factors such as temperature, humidity, light, and object shape.

References

[1] Tang Z, Jia S, Zhou C, Li B. 3D printing of highly sensitive and large-measurement-range flexible pressure sensors with a positive piezoresistive effect. ACS applied materials & interfaces, 2020, 12(25): 28669-28680.

[2] Jiahuan Wang, Haixiao Jia, Peifen Pan, et al. Research on The Technology of Man-Machine Collision Early Warning System In Tunnels Based On Bds High-Precision Positioning In Tunnel. Applied Computer Letters, 2023, 7(1)

[3] She X, Hongwei Z, Wang Z, Yan J. Feasibility study of asphalt pavement pothole properties measurement using 3D line laser technology. International Journal of Transportation Science and Technology, 2021, 10(1): 83-92.

[4] Bai L, Lundström O, Johansson H, Meybodi F, Arver B, Sandelin K, Brandberg Y. Clinical assessment of breast symmetry and aesthetic outcome: can 3D imaging be the gold standard. Journal of Plastic Surgery and Hand Surgery, 2023, 57(1-6): 145-152.

[5] Wang G, Ye J C, De Man B. Deep learning for tomographic image reconstruction. Nature machine intelligence, 2020, 2(12): 737-748.

[6] Jiahuan Wang, Haixiao Jia, Xuejiao Bai, et al. Research on The Location of Railway Train in Tunnel Based on Factor Graph Optimization. Applied Computer Letters, 2023, 7(1)

[7] Cui Y, Chen R, Chu W, Chen L., Tian D, Li, Y, Cao D. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(2): 722-739.

[8] Zengting Mu. Research on the Design of Wearable Life-Saving Furniture on Water Based on UCD. Applied Computer Letters, 2023, 7(2).

[9] Li P, Zhao H. Monocular 3d detection with geometric constraint embedding and semi-supervised training. IEEE Robotics and Automation Letters, 2021, 6(3): 5565-5572.

[10] Jiang W, Yin Y, Jiao J, Zhao X, Sun B. 2,000,000 fps 2D and 3D imaging of periodic or reproducible scenes with single-pixel detectors. Photonics Research, 2022, 10(9): 2157-2164.

[11] Pan X, Zhan X, Dai B, Lin D, Loy C C, Luo Pl. Exploiting deep generative prior for versatile image restoration and manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7474-7489.

[12] Zhang K, Li Y, Zuo W, Zhang L, Van Gool L, Timofte R. Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6360-6376.

[13] Guo Y, Xie S, Hu Y, Xu X. Color image guided depth image reconstruction based on a total variation network. JOSA A, 2024, 41(1): 19-28.

[14] Khojasteh-Khosro S, Shalbafan A, Thoemen H. Consumer behavior assessment regarding lightweight furniture as an environmentally-friendly product. Wood Material Science Engineering, 2022, 17(3): 192-201.

[15] Thio V, Aparicio J, Anonsen K B,Bekkeng J K, Booij W. Fusing of a continuous output PDR algorithm with an ultrasonic positioning system. IEEE Sensors Journal, 2021, 22(3): 2464-2474.

[16] Chen L C, Pardeshi M S, Lo W T,Sheu R K, Pai K C, Chen C Y, Tsai Y T. Edge-glued wooden panel defect detection using deep learning. Wood Science and Technology, 2022, 56(2): 477-507.

[17] Feng D, Wang C, He C, Zhuang Y. Kalman-filter-based integration of IMU and UWB for high-accuracy indoor positioning and navigation. IEEE Internet of Things Journal, 2020, 7(4): 3133-3146.

[18] Wang Q, Gao J, Lin W, Li X. NWPU-crowd: A large-scale benchmark for crowd counting and localization. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(6): 2141-2149.

[19] Çam G, Javaheri V, Heidarzadeh A. Advances in FSW and FSSW of dissimilar Al-alloy plates. Journal of Adhesion Science and Technology, 2023, 37(2): 162-194.

[20] Lai B F L, Lu R X Z, Davenport Huyer L,Kakinoki S, Yazbeck J, Wang E Y, Radisic M. A well plate‐based multiplexed platform for incorporation of organoids into an organ-on-a-chip system with a perfusable vasculature. Nature protocols, 2021, 16(4): 2158-2189.

[21] Hao R B, Lu Z Q, Ding H,Chen L Q. A nonlinear vibration isolator supported on a flexible plate: analysis and experiment. Nonlinear Dynamics, 2022, 108(2): 941-958.

[22] Premier A, GhaffarianHoseini A, GhaffarianHoseini A. Solar-powered smart urban furniture: preliminary investigation on limits and potentials of current designs. Smart and Sustainable Built Environment, 2022, 11(2): 334-345.

[23] Preethi P, Mamatha H R. Region-based convolutional neural network for segmenting text in epigraphical images Artificial Intelligence and applications. 2023, 1(2): 119-127.

Research and Implementation of Facial Expression Recognition Algorithm Based on Machine Learning

Xinjiu Xie*, Jinxue Huang
School of Modern Information Industry, Guangzhou College of Commerce, Guangzhou 511363, China

***Abstract*—Traditional information security management methods can provide a degree of personal information protection but remain vulnerable to issues such as data breaches and password theft. To bolster information security, facial expression recognition offers a promising alternative. To achieve efficient and accurate facial expression recognition, we propose a lightweight neural network algorithm called T-SNet (Teacher-Student Net). In our approach, the teacher model is an enhanced version of ResNet18, incorporating fine-grained feature extraction modules and pre-trained on the MS-Celeb-1M facial dataset. The student model uses the lightweight convolutional neural network ShuffleNetV2, with the model's accuracy further improved by optimizing the distillation loss function. This design carefully considers the key features of facial expressions, determines the most effective extraction techniques, and classifies and recognizes these features. To evaluate the performance of our algorithm, we conducted comparative experiments against state-of-the-art facial expression recognition methods. The results show that our approach outperforms existing methods in both recognition accuracy and efficiency.**

Keywords*—*Facial expression; expression recognition; convolutional neural network; deep learning

I. Introduction

Facial Expression Recognition (FER) technology has become increasingly prevalent across various fields, significantly enhancing human-computer interaction and automation systems. In healthcare, FER is utilized to diagnose and monitor mental health conditions by analyzing patients' emotional states. For example, individuals with depression or anxiety often display specific emotional characteristics that FER can help clinicians identify, leading to more personalized treatment plans. In marketing and retail, FER is employed to assess customer satisfaction and engagement by observing their reactions to products and advertisements. Businesses can use this technology to analyze emotional changes during shopping, allowing them to optimize product displays and advertising strategies, ultimately increasing conversion rates. Security systems leverage FER to detect suspicious behaviors and potential threats based on facial cues. In public spaces and critical facilities, FER can monitor the emotional state of crowds in real-time, quickly identifying abnormal behaviors to prevent potential security risks. In education, FER assists in evaluating students' comprehension and engagement during remote learning. By analyzing students' facial expressions during classes, teachers can better understand their attention levels and emotional states, allowing them to adjust teaching methods for improved educational outcomes.

The significance of FER lies in its ability to provide deeper insights into human emotions and intentions, which is crucial for enhancing communication and interaction between humans and machines. By accurately interpreting facial expressions, systems can respond more appropriately to users' needs, thereby improving user experience and efficiency. For instance, in customer service, FER-equipped automated systems can detect signs of customer frustration and promptly offer assistance, leading to increased satisfaction. These systems can analyze facial expressions in real-time during interactions with service representatives, immediately alerting the representative to take appropriate action when signs of confusion or dissatisfaction are detected, thus improving service quality.

Neural networks, especially deep learning models, have been instrumental in advancing facial expression recognition (FER). These models can automatically learn and extract complex features from facial images, outperforming traditional methods that rely on handcrafted features. For example, Convolutional Neural Networks (CNNs) are particularly effective at recognizing subtle facial expressions by capturing the spatial hierarchy of facial features. Through multiple convolutional and pooling layers, CNNs create hierarchical feature representations from raw images, which are essential for distinguishing various facial expressions. Moreover, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) enhance FER by incorporating temporal dynamics, enabling the analysis of expression sequences over time. These networks are well-suited for handling time-series data, as they capture dependencies between different time points—critical for understanding the dynamic nature of expression changes. The integration of neural networks into FER systems significantly enhances their accuracy and robustness, making these systems more reliable and practical for real-world applications. This technological progress expands the potential applications of facial expression recognition across various domains, driving further innovation in human-computer interaction technologies.

Section II introduces the current development status, and Section III presents the methods we propose. The superiority of our proposed method has been demonstrated through experiments in Section IV. Finally, a summary and outlook were made in Section V.

II. Literature Review

In this section, we introduce the development of recognition algorithms, and facial expression recognition algorithms, and summarize the research gaps.

*Corresponding Author.

A. Recognition Algorithm

With the advancement of technology, recognition algorithms have become increasingly sophisticated. These algorithms are not only prevalent in the field of image processing but are also applied across various other domains. Li et al. [1] proposed a method employing a differential evolution algorithm to optimize convolutional neural network (CNN) parameters for music emotion recognition tasks. Chen et al. [2] addressed the issues of slow convergence and weak generalization capability of CNNs in-vehicle feature recognition by proposing an improved bee colony algorithm to optimize CNN-based vehicle recognition strategies (ibsa-cnn). Li et al. [3] introduced a super-automatic algorithm combining non-local convolution and three-dimensional convolutional neural networks to address the shortcomings of missing critical feature information when processing long-time series video behavior features. Shi et al. [4] utilized data collected from multiple sensors measuring the Earth's magnetic field and employed a one-dimensional convolutional neural network algorithm for gesture recognition. Mao et al. [5] selected the Mel-frequency cepstral coefficient (MFCC) and filter bank (Fbank) as feature parameters to recognize English speech. Zhou et al. [6] proposed a method for lip print recognition based on convolutional neural networks.

In summary, recognition methods based on convolutional neural networks have been applied in a wide range of fields. Therefore, with the continuous development of technology, we can achieve more accurate facial expression recognition based on neural networks.

B. Facial Expression Recognition Algorithm

Currently, many scholars are dedicated to researching facial expression recognition algorithms. Zheng et al. [7] proposed a facial expression recognition method called TransformerKNN (TKNN), which integrates information about the state of the eyebrows and eyes in scenarios where the face is partially covered by a mask. Dong et al. [8] introduced a basic center regularization term based on the variance between basic centers to ensure that the learned expression features possess adequate discriminative capability. Wang et al. [9] proposed an Expression Complementary Disentanglement Network (ECDNet), which accomplishes the Facial Expression Disentanglement (FED) task during the face reconstruction process to handle all facial attributes in the disentanglement process. Yan et al. [10] introduced a new neonatal facial expression database for pain analysis. Win et al. [11] proposed a method for synthesizing complex facial expression images from learned expression representations without specifying emotion labels as input. Naveen et al. [12] addressed the problem of facial expression recognition under occlusion and proposed a robust facial expression recognition framework.

In summary, most current practice systems design a common practice path that all students must follow, lacking personalization. Due to the absence of necessary feedback, students' enthusiasm for practice is low, leading to low system utilization and poor results.

C. Research Gaps

The intelligent tutoring system proposed in this paper must address the following key challenges:

1) Effectiveness in capturing subtle differences: Current models struggle to effectively capture the nuanced differences between similar facial expressions, which is crucial for accurate facial expression recognition.

2) Complexity of recognition algorithms: Many existing facial expression recognition algorithms are overly complex due to their attempt to capture a wide range of features. This complexity often hinders performance, highlighting the need for a more lightweight model that can enhance recognition accuracy without unnecessary computational overhead.

These challenges are essential to consider in the development of facial expression recognition algorithms. The following sections of this paper will explore how deep learning methods can be utilized to achieve precise facial expression recognition while addressing these challenges.

III. Proposed Method

Our facial expression recognition model, designed using knowledge distillation, consists of a teacher model and a student model. The teacher model is an enhanced version of ResNet18, incorporating fine-grained feature extraction modules to capture deeper, multi-scale, and more detailed facial expression features. The student model utilizes the lightweight ShuffleNetV2 network, which maintains high recognition accuracy while being more computationally efficient.

The student model is trained and its parameters are updated by optimizing a distillation loss function. This function combines the probability distributions output by both the teacher and student models, enabling effective knowledge transfer. Through this process, the student model captures the essential knowledge from the teacher model, achieving high performance while remaining lightweight. Distillation training, therefore, facilitates the transfer of knowledge from a complex, high-performance model to a more efficient, lightweight model. The structure of this lightweight facial expression recognition network based on knowledge distillation is illustrated in Fig. 1.

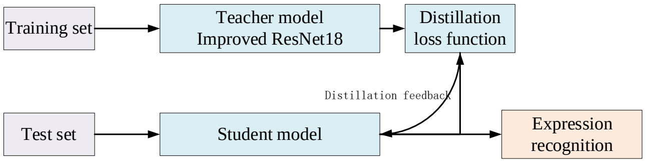


Fig. 1. Lightweight facial expression recognition network based on knowledge distillation.

In conclusion, our approach effectively tackles the challenge of distinguishing subtle differences between similar facial expressions while addressing the need for a more efficient and compact model. By utilizing knowledge distillation, we transfer the strengths of the teacher model to the student model, ensuring that the latter maintains high accuracy while being computationally efficient and suitable for real-time applications. This method offers a balanced solution, combining the depth and robustness of complex models with the speed and efficiency of lightweight models, making it an ideal choice for facial expression recognition tasks.

A. *Fine-grained Feature Extraction Module (FFE)*

To tackle the challenge of small intra-class variations and the difficulty in extracting features from different facial expression categories, we designed a fine-grained feature extraction module inspired by Res2Net [13]. This module enables the extraction of multi-scale features at a fine-grained level from critical facial regions within a single basic block, as illustrated in Fig. 2.

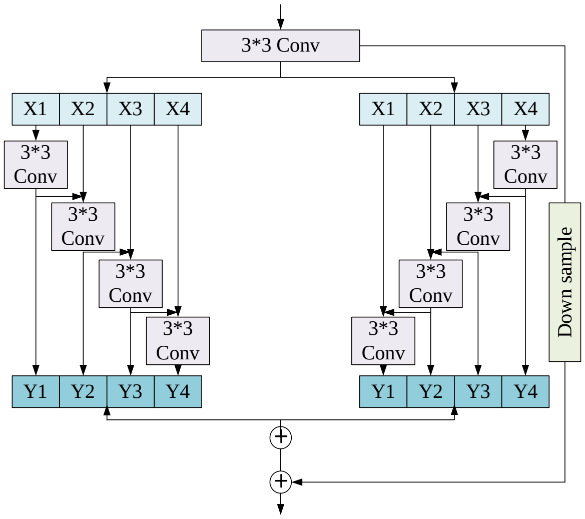


Fig. 2. Fine-grained Feature Extraction module (FFE).

Our module focuses on capturing subtle differences within the same category of facial expressions by emphasizing key facial areas. By leveraging the Res2Net-inspired architecture, the module enhances the model's ability to discern intricate details crucial for accurate facial expression recognition. This approach improves the model's capacity to distinguish between expressions with minimal variation, thereby increasing the overall accuracy and robustness of the recognition system.

The fine-grained feature extraction module is seamlessly integrated into our knowledge distillation framework, enhancing the student model's ability to learn from the teacher model. By including this module, we ensure that the lightweight student model retains the high-resolution, multi-scale features essential for precise facial expression analysis. This integration leads to a more effective and efficient facial expression recognition system, well-suited for real-world applications with limited computational resources.

In summary, the fine-grained feature extraction module effectively addresses the challenge of feature extraction in facial expression recognition by improving the system's ability to capture and analyze subtle facial cues. This innovation, coupled with our knowledge distillation approach, advances the development of more accurate and efficient facial expression recognition models.

The fine-grained feature extraction module designed in this paper utilizes a symmetrical structure to learn multi-scale features within a single basic block. This design ensures that feature subsets from both preceding and succeeding stages contain richer scale information. Specifically, after applying a 3×3 convolution to the input feature map X, the map is evenly divided along the channel axis into n subsets, denoted as X_i, where $i \in \{1, 2, \ldots, n\}$. Each subset X_i retains the same spatial dimensions as the original feature map X but with $1/n$ of the channels.

These subsets are then processed using a 3×3 convolution, denoted as $P_i^p(\cdot)$, where $p \in \{left, right\}$ indicates the position within the symmetrical structure. The output of each convolution operation, Y_i^p, can be expressed using Eq. (3). This method enables the extraction of fine-grained, multi-scale features, enhancing the model's capability to capture detailed facial expression characteristics.

$$\mathrm{Y_i^{left}} = \begin{cases} \mathrm{P_i^{left}(X_i)},\ \mathrm{i} = 1 \\ \mathrm{P_i^{left}\left(X_i + Y_{i-1}^{left}\right)},\ 1 < \mathrm{i} \leq \mathrm{n} \end{cases} \tag{1}$$

$$\mathrm{Y_i^{right}} = \begin{cases} \mathrm{P_i^{right}(X_i)},\ \mathrm{i} = 1 \\ \mathrm{P_i^{right}\left(X_i + Y_{i-1}^{right}\right)},\ 1 < \mathrm{i} \leq \mathrm{n} \end{cases} \tag{2}$$

$$\mathrm{Y_i} = \mathrm{Y_i^{left}} + \mathrm{Y_i^{right}} \tag{3}$$

According to Eq. (1), each operation of $P_i^{left}(\cdot)$ captures feature from all subsets $\{X_j, \leqslant i\}$. Conversely, Eq. (2) shows that each operation of $P_i^{right}(\cdot)$ captures feature from subsets $\{\ X_j\ , \geqslant j \geqslant i\}$. Each operation involves applying a 3×3 convolution to a split feature X_i. As a result, the output Y_i^{left} has a larger receptive field compared to $\{Y_k, < i\}$, while Y_i^{right} has a larger receptive field compared to $\{Y_k, k > i\}$. This design enables the extraction of comprehensive multi-scale features from both preceding and succeeding stages, enhancing the model’s ability to analyze fine-grained details.

Each output Y_i^p contains facial features at varying scales and quantities. To achieve a richer diversity of multi-scale features, all Y_i^p are aggregated along the channel axis, integrating the fine-grained details from each subset. While increasing n enriches the feature representation, it also raises computational costs. To strike an optimal balance between performance and efficiency, we set n = 4.

This design ensures that the fine-grained feature extraction module effectively captures multi-scale facial features crucial for accurate and robust facial expression recognition. By aggregating features from different scales, the model enhances its ability to detect subtle variations in facial expressions, thereby improving accuracy and reliability for practical applications.

Setting n = 4 allows us to balance feature richness with computational efficiency, making our method practical for real-world applications where both accuracy and speed are critical. This design choice demonstrates our commitment to optimizing performance while maintaining computational feasibility.

By utilizing this approach, we ensure that both initial and subsequent feature subsets capture detailed multi-scale information. This enhances the model's ability to detect fine-grained variations in facial expressions, resulting in more accurate and robust facial expression recognition.

Fig. 2 provides a detailed illustration of the symmetrical structure of our fine-grained feature extraction module. The figure demonstrates how the module employs symmetrical convolutions to process feature maps, enhancing the extraction of multi-scale features.

This innovative module integrates seamlessly into our facial expression recognition framework, allowing the lightweight student model to effectively learn complex, multi-scale features from the teacher model. As a result, the system achieves high performance while maintaining computational efficiency, making it well-suited for deployment in real-world applications with limited resources.

B. *The Construction of Teacher Model and Student Model*

We selected ResNet18 for its robust feature extraction and high recognition accuracy as the backbone of our teacher model. To further enhance its performance, we integrated a fine-grained feature extraction module designed to capture intricate facial details, especially from critical regions such as the eyes and mouth. This improvement enables the teacher model to provide more precise supervision and guidance to the student model. The enhanced teacher model, as illustrated in Fig. 3, processes facial images of size 224×224×3. The image first passes through standard convolutional layers, followed by batch normalization, ReLU activation, and max pooling, producing feature maps of size 112×112×64. Two residual blocks then generate feature maps of size 28×28×128. These are subsequently processed through two sequential fine-grained feature extraction modules, resulting in multi-scale, high-resolution features of size 224×224×3. Finally, these features are fed into fully connected layers for facial expression classification. This configuration allows the teacher model to effectively capture and leverage detailed facial features, ensuring that the student model benefits from enhanced guidance during training.

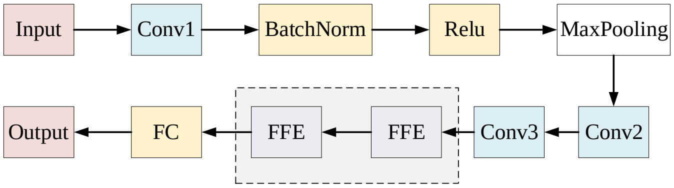


Fig. 3. Teacher model.

We selected ShuffleNetV2 as the backbone for the student model due to its computational efficiency and lightweight design. This network excels in real-time applications thanks to its use of depth wise separable convolutions, which significantly reduce model parameters and speed up computations. This makes ShuffleNetV2 particularly well-suited for deployment on edge devices for real-time facial expression recognition.

The ShuffleNetV2 architecture, depicted in Fig. 4, processes 224×224×3 facial images. It starts with a 3×3 standard convolutional layer followed by max pooling. The network then progresses through several stages, including down sampling blocks and basic blocks that utilize depth-wise separable convolutions. After these stages, the model passes through fully connected layers to produce the final facial expression classification. This design balances efficiency with performance, making it ideal for practical applications requiring real-time processing.

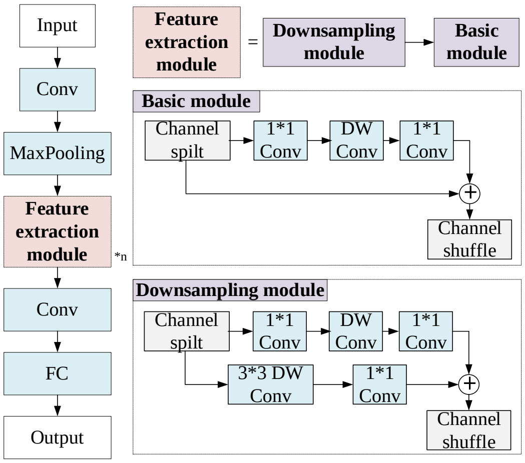


Fig. 4. Student model.

C. *Facial Expression Recognition Network Based on Knowledge Distillation*

Knowledge distillation is a widely utilized model compression technique that enhances the performance and accuracy of a lightweight student model by leveraging the expertise of a larger, more complex teacher model. This technique enables the student model to benefit from both the "hard labels" of the dataset and the "soft labels" provided by the teacher model's probabilistic outputs. The essence of knowledge distillation is to transfer the nuanced insights and detailed feature representations learned by the teacher model to the student model. In the T-SNet framework, this is achieved through a dual-branch network architecture, which includes both the teacher and student models. As illustrated in Fig. 5, the teacher model provides comprehensive guidance by generating rich, informative soft labels that the student model uses for training. This setup ensures that the student model captures not just the direct class labels but also the underlying distributions and patterns recognized by the teacher, leading to improved performance and accuracy in a more compact and efficient model.

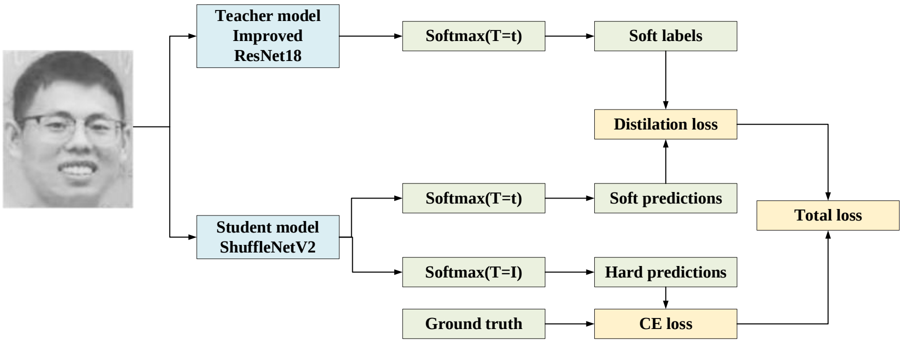


Fig. 5. Knowledge distillation model.

In the T-SNet framework, the dual-branch architecture operates as follows:

1) Teacher model branch: This branch performs global fine-grained feature extraction and classification. It leverages its complex architecture to capture detailed facial features and nuances.

2) Student model branch: This branch trains the student model using distilled features from the teacher model. It adapts these features to its more compact and efficient design for facial expression classification.

The framework employs a joint loss function to guide the training of both models. This ensures the student model retains essential knowledge from the teacher model, enhancing efficiency and performance while reducing computational requirements. This approach is especially suitable for real-time facial expression recognition where resources are constrained.

During the knowledge distillation process, facial expression images are simultaneously inputted into both the teacher and student models for feature extraction. The teacher model generates a Softmax distribution at a high temperature T, which serves as soft labels. Ultimately, the classifier outputs from both the teacher and student networks determine the expression categories. Introducing a Softmax temperature function helps smooth the probability distribution of predicted expressions, thereby providing additional class-specific feature information inherent to the teacher network.

D. Loss Function

During the knowledge distillation training process, the loss function L_{kd} for the student model comprises two key components: L_{soft} and L_{hard}, corresponding to learning from "soft labels" and "hard labels", respectively. L_{soft} uses Kullback-Leibler (KL) divergence to measure the difference between the predictions of the teacher model and the student model. Specifically, it calculates the cross-entropy loss between the Softmax output of the student model under the same temperature T and the soft labels provided by the teacher model. The L_{hard} component uses the cross-entropy loss function to compute the difference between the predictions of the student model and the true facial expression labels. The final loss for optimization is the sum of these two parts.

The L_{soft} loss function is formulated as follows:

$$\mathrm{Y_i} = \mathrm{Y_i^{left}} + \mathrm{Y_i^{right}} \tag{4}$$

where p_i^T denotes the Softmax output of the teacher model for class i at temperature T, and q_i^T denotes the Softmax output of the student model for class i at temperature T. The expressions for p_i^T and q_i^T are given by:

$$\mathrm{p_i^T} = \frac{\mathrm{e^{v_i/T}}}{\sum_\mathrm{k}^\mathrm{N} \mathrm{e^{v_k/T}}} \tag{5}$$

$$\mathrm{q_i^T} = \frac{\mathrm{e^{z_i/T}}}{\sum_\mathrm{k}^\mathrm{N} \mathrm{e^{z_k/T}}} \tag{6}$$

where v_i is the raw logits from the output layer of the teacher model, z_i is the raw logits from the output layer of the student model, and N is the total number of labels.

The L_{hard} loss function is defined as Eq. (7):

$$\mathrm{L_{hard}} = -\sum_\mathrm{i}^\mathrm{N} \mathrm{c_i} \log(\mathrm{q_i^1}) \tag{7}$$

$$\mathrm{q_i^1} = \frac{\mathrm{e^{z_i}}}{\sum_\mathrm{k}^\mathrm{N} \mathrm{e^{z_k}}} \tag{8}$$

$$\mathrm{Totleloss} = \alpha \mathrm{T}^2 \mathrm{L_{soft}} + (1-\alpha)\mathrm{L_{hard}} \tag{9}$$

where q_i^1 is the Softmax output of the student model for class i without temperature scaling, given by Eq. (8). The total loss $Totleloss$ is a combination of L_{soft} and L_{hard}.

where α is a balancing factor between the distillation loss and the cross-entropy loss.

Because the gradient magnitude of soft labels is scaled down by $1/T^2$, L_{soft} is multiplied by T^2 to ensure consistency between the true label values and the probability distribution of the teacher model. The temperature T and the coefficient α are hyperparameters that need to be empirically determined, which we explore further in subsequent sections through ablation experiments.

IV. Experiment and Verification

In this chapter, we verify the reliability and validity of the proposed method through experiments.

A. Experimental Environment

This study validated the algorithm's effectiveness using an environment comprising an 11th Gen Intel(R) Core (TM) i7-11700K @ 3.60GHz CPU with 32.0 GB of RAM, running

Python 3.6. For facial expression recognition experiments, a dataset was created from randomly captured images manually annotated. The dataset consists of 8500 images, divided into training, testing, and validation sets in a 6:2:2 ratio. The testing set comprises 1700 three-dimensional images with faces, randomly grouped into five sets of 340 images each. Each group of images was processed using the system for facial expression recognition.

B. *Evaluation Parameter*

First, the data set is aligned to the face [14], then the pre-processed face image is input to the teacher model training, the teacher model parameters are saved, and then the teacher model is trained by knowledge distillation to assist the student model [15]. Finally, the accuracy and the number of model parameters are calculated on the test set to verify the performance of the model.

We choose number of Accuracy, Precision, Accuracy, and the cords FLOPs as evaluation index, accurate calculation method is as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{10}$$

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{11}$$

Where TP represents true positives, meaning instances classified as positive that are actually positive; FN represents false negatives, where instances classified as negative but are actually positive; FP represents false positives, where instances classified as positive but are actually negative; and TN represents true negatives, where instances classified as negative that are actually negative.

AUC (Area Under the Curve) is a metric used to evaluate the performance of binary classification models, indicating the probability that the model ranks a randomly chosen positive example higher than a randomly chosen negative example. Firstly, True Positive Rate (TPR) refers to the proportion of actual positive cases correctly predicted as positive, calculated as $\frac{TP}{TP+FN}$. False Positive Rate (FPR) is the proportion of actual negative cases incorrectly predicted as positive, calculated as $\frac{FP}{TN+FP}$. The ROC (Receiver Operating Characteristic) curve plots TPR against FPR, with AUC representing the area under this curve. A higher AUC indicates better model performance in distinguishing between positive and negative examples.

Parameter count is a metric measuring model complexity and storage requirements, calculated based on the model's architecture and layer types [16]. Parameters are computed using the model.parameters function in PyTorch. Computational complexity (FLOPs) measures the total number of floating-point operations performed by the model during inference or training. Additionally, this study evaluates inference time to assess model processing speed on embedded devices.

During training, stochastic gradient descent (SGD) optimizes the loss function with the following hyperparameters: weight decay of 1e-4, momentum of 0.9, dropout rate of 0.5, and initial learning rate of 0.01. Learning rate adjustments are dynamically managed using ReduceLROnPlateau, decreasing the learning rate by a factor of 10 if the loss does not decrease after 3 epochs. The total training epochs are set to 150 with a batch size of 128. The best model is saved after 150 epochs based on the highest achieved accuracy.

C. *Test and Evaluation*

We set up ablation experiments to demonstrate the superiority of our parameter selection. Table I shows the comparison of results of different parameters. In the knowledge distillation process, the distillation temperature T and α are set to 8 and 0.4 respectively. The ablation experiment verified the best effect of this parameter setting.

TABLE I. ABLATION EXPERIMENT

T	α	**precision**
4	0.4	0.899
6	0.2	0.893
6	0.4	0.988
6	0.6	0.906
8	0.4	0.891

We tested the results of teacher-only model training, student-only model training and the results of the combined training of teacher model and student model after adding knowledge distillation on FER2013Plus and RAF-DB datassets. To ensure the fairness of experimental results, the same hyperparameters were used for both teacher model and student model. In order to evaluate the performance of knowledge distillation algorithm, experiments were conducted on FER2013Plus and RAF-DB datasets respectively to test the accuracy and number of parameters of the improved teacher model, student model and distilled student model. Table II shows the test results of teacher model and student model on FER2013Plus and self-made datasets. The faculty model is named EFF-ResNet18(Enhanced Fine- ResNet18), the student model is named ShuffleNetV2, and the distilled student model is named T-SNet.

TABLE II. TEACHER MODEL AND STUDENT MODEL TEST RESULTS

Datasets	Model	parameters	Accuracy
FER2013Plus	EFF-ResNet18	11.8	86.23
	ShuffleNetV2	1.3	88.15
	T-SNet	1.2	91.62
Our datasets	EFF-ResNet18	12.3	88.57
	ShuffleNetV2	1.3	90.45
	T-SNet	1.2	94.69

According to Table II analysis, on the FER2013Plus dataset, knowledge distillation improved the accuracy of the student model from 86.23% to 91.62%. Similarly, on our proprietary dataset, performance enhancement was observed, increasing from 88.57% to 94.69%. Across both datasets, the distilled student model surpassed the teacher model in accuracy, while reducing parameter count by 10.60 million, indicating significant improvement. Experimental results demonstrate that knowledge distillation enhances the generalization performance of the student model by transferring knowledge from the teacher model, achieving model lightweighting in the process. This highlights the effectiveness of training lightweight models with performance comparable to large-scale network models under supervision of feature extraction from large network models.

To assess the performance of T-SNet in facial expression recognition tasks, this study compared T-SNet with other mainstream algorithms including RCL-Net [17], ResNet18, TFEN [18], ECDNet, etc. Comparative results on our dataset are presented in Table III.

TABLE III. COMPARISON OF T-SNET AND SOTA METHODS

Models	Precision	Accuracy	FLOPs
TFEN	0.898	0.906	4.5
ResNet18	0.903	0.893	11.18
RCL-Net	0.901	0.891	6.4
ECDNet	0.897	0.903	3.5
Ours	0.931	0.988	1.3

Through comparison, it can be observed that the recognition accuracy of our proposed method on the data set is as high as 98.8%, which is much higher than the second method. This shows that the proposed method is superior to other SOTA methods in terms of accuracy. In addition, in terms of the number of parameters and the amount of computation, our method has the lowest parameter number of 1.3M. Our method has very good advantages in recognition rate and parameter number, which can be said to be a qualitative leap for the system with high real-time requirements. The experimental results show that the lightweight expression recognition model based on knowledge distillation proposed in this paper has strong competitiveness and can be used as one of the choices for a real-time facial expression recognition system.

TABLE IV. COMPARISON OF RECOGNITION EFFECT OF DIFFERENT EXPRESSIONS

Expression type	FER2013Plus	Our datasets
Angry	0.95	0.96
disgust	0.90	0.91
happy	0.98	0.99
sad	0.92	0.93
surprised	0.96	0.97
neutral	0.97	0.98
contempt	0.91	0.93

We respectively tested the recognition effects of different expressions on FER2013Plus and self-made data sets with our proposed method, as shown in Table IV. From the confusion matrix of the two data sets, it can be concluded that happy expression is the most easily recognized expression, mainly because happy data in the two data sets has the highest amount and more abundant features. Happiness is also highly recognizable in the real world. Neutral is easier to identify because some uncertain subtle expressions are generally marked as neutral during data set annotation. The recognition rates of surprised, angry and sad expressions declined successively, which is consistent with human visual characteristics.

In order to more intuitively verify the effectiveness of T-SNet, this paper uses GradCAM to generate visual activation diagram for visual analysis of T-SNet [19]. Among them, the data samples are from the FER2013Plus test set. The visualized result is shown in Fig. 6. T The first column is the original data sample, the second column is the ECDNet model activation map, and the third column is the T-SNet activation map. The visual activation graph can verify the importance of the network to the key areas of the image, and the brighter the color, the more important the content of the area is for the recognition of the network. Red indicates high activation and blue indicates low activation. The comparison between the second and third column shows that T-SNet after knowledge distillation can absorb the fine-grained feature extraction experience of the teacher model, and the extracted feature semantic information is richer, and the receptive field is larger than that of the baseline model. By focusing attention on important facial areas, T-SNet can more accurately identify expression categories, which verifies T-SNet's excellent expression recognition performance.

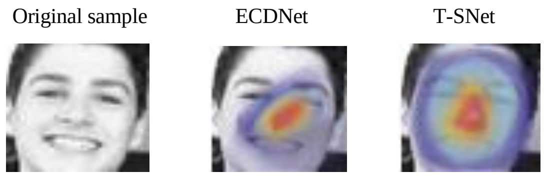


Fig. 6. Visual analysis.

V. CONCLUSION

We propose a lightweight algorithm for facial expression recognition based on neural networks, termed T-SNet. Addressing the challenge of deploying complex neural networks on smart terminals and the increasing demand for lightweight facial expression recognition models, we introduce T-SNet, a knowledge-distillation-based lightweight facial expression recognition network. We selected a refined ResNet18 with enhanced fine-grained feature extraction module as the teacher model backbone and ShuffleNetV2 as the student model backbone. By constructing distillation loss, we transfer rich classification information from the teacher model to the student model, leveraging the teacher model's experience to train the student model and improve its accuracy in facial expression recognition. The distilled student model, trained as a lightweight facial expression recognition model, is the outcome of our approach. Experimental results demonstrate that our proposed T-SNet model outperforms other mainstream facial recognition models in terms of accuracy and parameter efficiency. Despite the achievements in lightweight facial recognition, real-world scenarios still pose challenges due to factors such as environment, lighting, and occlusion.

While our method shows promising performance on current datasets, demonstrating excellent effectiveness and efficiency, further validation in practical applications is essential. Future advancements in deep learning technology and dataset enhancements will likely expand the application of facial recognition into more domains. To ensure effective deployment in diverse applications, rigorous empirical research on feasibility and efficacy is needed. This will further validate the practical potential of our approach.

Acknowledgment

This study received funding from three sources: the Campus Level Quality Engineering Project , titled " Research on the Construction and Application of Smart Classrooms Supporting the Development of Advanced Thinking in the Digital Age", sponsored by Guangzhou College of Commerce (No.: 2024JXGG47); Higher Education Scientific Research Topic Project, titled "Research on the Construction of Digital Teaching Resources in Applied Universities under the Background of School-Enterprise Collaborative Education", sponsored by Office of Guangdong Provincial Leading Group for Education Science Planning (No.: 2023GXTK581); Guangdong Province Ordinary University Characteristic Innovation Project, titled "Research on Key Technologies for Campus Network Security Situation Awareness", sponsored by scientific research division of Guangdong Provincial Department of Education(No.: 2020KTSCX169).

References

[1] J. Li, S. Soradi-Zeid, A. Yousefpour, D. Pan, “Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data,” Applied Soft Computing, 153. 2024. https://doi.org/10.1016/j.asoc.2024.111262

[2] X. Chen, “Vehicle Feature Recognition via A Convolutional Neural Network with An Improved Bird Swarm Algorithm,” *Journal of Internet Technology*, *24*(2), 421–432. 2023. https://doi.org/10.53106/160792642023032402020

[3] J. Li, J. Liu, C. Li, F. Jiang, J. Huang, S. Ji, Y. Liu, “A hyperautomative human behaviour recognition algorithm based on improved residual network,” *Enterprise Information Systems*, *17*(10). 2023. https://doi.org/10.1080/17517575.2023.2180777

[4] B. Shi, X. Chen, Z. He, H. Sun, R. Han, “Research on Gesture Recognition System Using Multiple Sensors Based on Earth’s Magnetic Field and 1D Convolution Neural Network,” *Applied Sciences (Switzerland)*, *13*(9), 2023. https://doi.org/10.3390/app13095544

[5] C. Mao, S. Liu, “A Study on Speech Recognition by a Neural Network Based on English Speech Feature Parameters,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *28*(3), 679–684, 2024. https://doi.org/10.20965/jaciii.2024.p0679

[6] H. Zhou, “Lip Print Recognition Algorithm Based on Convolutional Network,” *Journal of Applied Mathematics*, *2023*. https://doi.org/10.1155/2023/4448861

[7] K. Zheng, L. Tian, Z. Li, H. Li, J. Zhang, “Incorporating eyebrow and eye state information for facial expression recognition in mask-obscured scenes,” *Electronic Research Archive*, *32*(4), 2745–2771. 2024. https://doi.org/10.3934/ERA.2024124

[8] R. Dong, K. M. Lam, “Bi-Center Loss for Compound Facial Expression Recognition,” *IEEE Signal Processing Letters*, *31*, 641–645, 2024. https://doi.org/10.1109/LSP.2024.3364055

[9] S. Wang, H. Shuai, L. Zhu, Q. Liu, “Expression Complementary Disentanglement Network for Facial Expression Recognition,” *Chinese Journal of Electronics*, *33*(3), 742–752, 2024. https://doi.org/10.23919/cje.2022.00.351

[10] J. Yan et al., “FENP: A Database of Neonatal Facial Expression for Pain Analysis. *IEEE Transactions on Affective Computing*, *14*(1), 245–254. 2023. https://doi.org/10.1109/TAFFC.2020.3030296

[11] S. S. K. Win, P. Siritanawan, K. Kotani, “Compound facial expressions image generation for complex emotions,” *Multimedia Tools and Applications*, *82*(8), 2023. 11549–11588. https://doi.org/10.1007/s11042-022-14289-7

[12] P. Naveen, “Occlusion-aware facial expression recognition: A deep learning approach,” *Multimedia Tools and Applications*, *83*(11), 32895–32921, 2024. https://doi.org/10.1007/s11042-023-17013-1

[13] S H Gao, M M Cheng, K Zhao, et al., “Res2net: A new multi-scale backbone architecture,” IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[14] R R Adyapady, B. Annappa, “A comprehensive review of facial expression recognition techniques,” Multimedia Systems, 29(1): 73-103, 2023.

[15] Xinchen Pan, Ling Qin, Xiaojian Yang, “Facial expression recognition in complex scenes based on multi-region detection network,” Data Acquisition and Processing, 38(06): 1422-1433, 2023.

[16] Yahui Nan, Qingyi Hua, “Occluded face expression recognition deep learning Progress in law research,” Application Research of Computers, 39 (2): 321-330, 2022.

[17] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, G. He, “Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP,” *Sensors*, *23*(9), 2023. https://doi.org/10.3390/s23094204

[18] J. Teng, D. Zhang, W. Zou, M. Li, D. J. Lee, “Typical Facial Expression Network Using a Facial Feature Decoupler and Spatial-Temporal Learning,” *IEEE Transactions on Affective Computing*, *14*(2), 1125–1137, 2023. https://doi.org/10.1109/TAFFC.2021.3102245

[19] R R Selvaraju, M Cogswell, A Das, et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” IEEE international conference on computer vision. 618-626, 2017.

Enhancing Music Emotion Classification Using Multi-Feature Approach

Affreen Ara[1], Rekha V[2]
Christ University, Bengaluru, India[1]
Department of Computer Science and Engineering, Christ University, Bengaluru, India[2]

Abstract—**Emotions are a fundamental aspect of human expression, and music lyrics are a rich source of emotional content. Understanding the emotions conveyed in lyrics is crucial for a variety of applications, including music recommendation systems, emotion classification, and emotion-driven music composition. While extensive research has been conducted on emotion classification using audio or combined audio-lyrics data, relatively few studies focus exclusively on lyrics. This gap highlights the need for more focused research on lyric-based emotion classification to better understand its unique challenges and potentials. This paper introduces a novel approach for emotion classification in music lyrics, leveraging a combination of natural language processing (NLP) techniques and dimension reduction methods. Our methodology systematically extracts and represents the emotional features embedded within the lyrics, utilizing a diverse set of NLP techniques and integrating new features derived from various emotion lexicons and text analysis. Through extensive experimentation, we demonstrate the effectiveness of our approach, achieving significant improvements in accurately classifying the emotions expressed in music lyrics. This study underscores the potential of lyric-based emotion analysis and provides a robust framework for further research in this area.**

Keywords—Emotion classification; music lyrics; feature extraction; lexicon features

I. INTRODUCTION

Music can evoke strong emotions in listeners, such as happiness, sadness, anger, or excitement [1]. Analyzing the emotional content of music lyrics provides insights into why certain songs are popular and how they affect our moods. Emotion analysis employs natural language processing (NLP) and machine learning techniques to identify and extract emotional information from text. This analysis involves examining written or spoken language to understand how different songs and genres influence emotional responses. Emotion analysis of music lyrics is applicable in various fields, including marketing, social media analysis, mental health, and music therapy [2]. For instance, music therapy leverages specific songs and genres to help individuals' process emotions and improve well-being, which is also useful for diagnosing and monitoring mental states in mental health settings.

NLP, a branch of artificial intelligence, plays a pivotal role in emotion analysis by enabling computers to understand and interpret human language [3]. In emotion analysis, NLP techniques such as emotion detection are utilized to identify and classify emotions expressed in lyrics. The process typically begins with pre-processing the music lyric text, involving tokenization, normalization, and stop-word removal. This is followed by feature extraction, where relevant features such as stylistic, semantic, and lexicon-based attributes are identified and extracted from the text. Machine learning algorithms are then trained on these extracted features to identify the emotions embedded within music lyrics.

Emotion lexicons, which include vocabulary terms associated with one or more emotions, are crucial resources in NLP, sentiment analysis, and emotion detection [4]. These lexicons, such as the Affective Norm for English Words (ANEW) [5], enable emotion analysis algorithms to classify the emotional tone of text based on word associations. ANEW is grounded in the Russell Model's [6] two-dimensional circumplex model, which includes valence (pleasantness) and arousal (activation). While word-based lexicons are widely used, phrase-based and context-based lexicons offer a more nuanced understanding of emotional content. However, the subjectivity of emotions, polysemy, and contextual dependence can pose challenges in accurately assigning emotions to text. To overcome these limitations, emotion analysis often combines lexicons with other features, such as sentiment analysis, syntactic analysis, and semantic analysis. This integrated approach provides a more comprehensive and accurate understanding of the emotional content of text data.

Dimension reduction [7] is a technique used in data analysis and machine learning to reduce the number of features or variables in a dataset while preserving essential information. This technique is vital for uncovering underlying emotional patterns in lyrics, allowing for more accurate and computationally efficient machine learning models. By applying dimension reduction methods like Principal Component Analysis (PCA), researchers can effectively interpret and classify the rich emotional expressions found in music, providing valuable insights into the intersection of art and emotion.

Machine learning algorithms, such as decision trees, k-Nearest Neighbor (KNN), Naive Bayes, and Support Vector Machines (SVM), are commonly used for emotion classification in lyric text. The choice of algorithm depends on factors such as dataset size, feature representation, and computational efficiency. While most existing research in emotion analysis adopts a multimodal approach, combining both audio and lyrics to enhance emotion prediction accuracy, there is a notable gap in research focusing solely on lyrics as a textual representation of emotions. This research addresses this gap by providing a comprehensive analysis of emotions in music lyrics and designing new lexicon features that enhance emotion detection capabilities.

Our previous research [8] investigated the effectiveness of various lexicon-based features for emotion classification using music lyrics, achieving an accuracy of 53%. We extracted lexicon features from the NRC Emotion Intensity Lexicon, NRC VAD, EmoWordNet, and Synesketch lexicon. Extensive experiments were carried out using both lexicon and hybrid lexicon features for emotion classification in music lyrics; the current study extends this approach by incorporating a broader set of features, such as stylistic, lexical, and textual attributes, along with traditional NLP methods. Furthermore, we use Principal Component Analysis (PCA) for dimensionality reduction to enhance classifier performance by capturing the most informative aspects of the data.

This paper aims to enhance emotion classification accuracy in music lyrics by integrating various text features, including lexicon-based, syntactic, stylistic, and semantic attributes. Our contributions are threefold: (1) We create robust hybrid feature sets by combining lexicon-based features from multiple emotion sources to augment the emotional representation in lyrics; (2) We also use feature engineering techniques to extract syntactic, semantic, and stylistic information from the lyrics, thereby capturing emotional complexity of music lyric text; (3) We also use Principal Component Analysis (PCA) to reduce the dimensionality of the feature space in order to improve computational efficiency and model interpretability. By conducting extensive experimentation with classifiers such as Decision Trees, Random Forest, and Gradient Boosting, we demonstrate that our approach significantly enhances classification accuracy, achieving up to 98% accuracy. The remainder of this paper is structured as follows: Section II reviews the related work, Section III explains the method used, Section IV presents the experimental results, Section V discusses the findings and their implications, and lastly, Section VI concludes the paper with potential directions for future research.

II. Literature Review

This section presents literature review from papers of emotion analysis from music lyrics and text analysis.

Revathy V.R. [9] study uses knowledge from the MER dataset to train the Music4All dataset, aiming to label lyrics with emotions. The research employs a transfer learning approach and the Sentence Transformer model for emotion prediction. The LyEmoBERT model demonstrated superior performance compared to existing methods on the Music4All dataset. Sujeesha et al. [10] research develops a multimodal music mood classification system using transformers, incorporating both audio and lyrics. The study compares the system's performance with a Bi-GRU-based model, finding that the transformer-based model with transfer learning achieves higher accuracy. Priyanka et al. [11] propose a mood categorization of songs based solely on lyrics using TF-IDF feature extraction and the Random Forest algorithm. Their findings highlight the model's ability to accurately predict "happy" and "sad" emotions. Yudhik Agrawal [12] study proposes a deep neural network architecture using the XLNet Transformer model for emotion classification in music lyrics. The model employs multi-task learning through weight sharing, enhancing convergence speed and reducing over fitting. Fika Hastarita Rachman et al. [13] work introduces a method to recognize song emotions by combining lyrics and audio using the Thayer emotion model. They extract psycholinguistic and stylistic features from lyrics and audio waveforms, using various classifiers to perform emotion classification. Cong Jin et al. [14] propose a Bi-LSTM network with dilated recurrent skip connections to improve the model's ability to capture long-sequence information in lyrics. The model includes an attention mechanism to enhance the recognition of important words, improving semantic extraction performance. Study presents MoodNet[15], a deep convolution neural network (CNN)-based architecture designed to determine emotions from audio and lyrics. The model is evaluated using the MIREX Multimodal Dataset and the Million Song Dataset. Shahrzad Naseni et al. [16] research explores the connection between song lyrics and mood using two transformer-based approaches: natural language inference and next sentence prediction. The study focuses on lyric classification tasks to understand how lyrics and acoustics contribute to song mood. Leroto Parisi [17] study examines various feature vector representations like BERT, ELMo, and fastText embeddings combined with deep learning mechanisms to predict emotions conveyed in song lyrics. Yingjin Song and Daniel Beck [18] work introduces a two-stage BERTLex-State Space Model framework for sequence-labeling emotion intensity recognition tasks. The framework is aimed at predicting emotion dynamics in song lyrics without requiring supervision at the song level.

The literature review discusses various approaches to emotion analysis in music lyrics and text analysis. It highlights the use of datasets like MER and Music4All for training models to label lyrics with emotions, employing techniques such as transfer learning and transformer-based models. Notable works include the LyEmoBERT [9] model, which outperformed existing methods, and a multimodal music mood classification system that integrates audio and lyrics. Other studies, such as those by Yudhik Agrawal et al. [12], explore the use of deep neural networks like XLNet for emotion classification, utilizing multi-task learning for improved model performance. Fika Hastarita Rachman et al. [13] work focuses on combining psycholinguistic and stylistic features from lyrics with audio features using various classifiers for emotion recognition. Cong Jin et al.[14] introduce a Bi-LSTM network with dilated skip connections and attention mechanisms to better capture long-sequence dependencies in lyrics. A Bhattacharya, et al. propose the MoodNet architecture, a CNN-based model that analyzes both audio and lyrics for emotion detection. Additionally, Shahrzad Naseni [16] and colleagues use transformer-based models for lyric classification tasks, demonstrating the effectiveness of advanced NLP techniques. The review also emphasizes the gap in research focusing solely on lyrics for emotion classification, advocating for more targeted studies to address unique challenges in this area. It covers the calculation of emotional metrics, machine learning algorithms, feature reduction techniques like PCA, and the importance of lexicons in analyzing emotional content in lyrics.

III. Method for Music Emotion Classification for Lyric Text

The system takes music lyrics as input. These lyrics undergo a cleaning process to remove any unwanted elements and are then processed and converted into individual tokens. Feature extraction is performed on the lyrics, extracting various relevant features from the textual content, including leveraging emotion lexicons. To reduce the dimensionality of the feature space and remove irrelevant features, dimension reduction techniques are applied. Classification models are then constructed using the extracted features, both with and without dimension reduction, to accurately identify and classify the emotions embedded within the lyrics (Fig. 1).

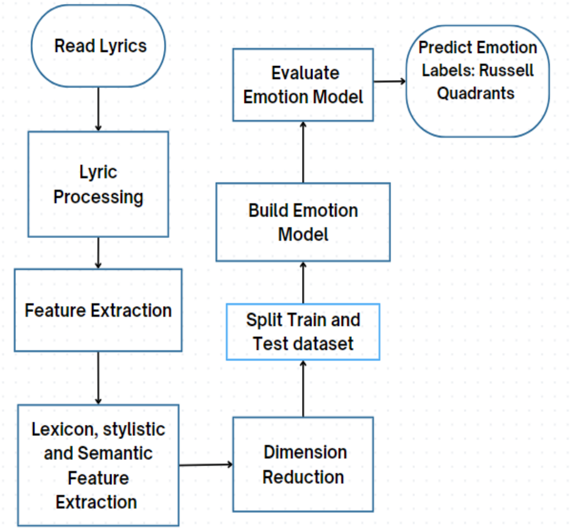

Read Lyrics
Lyric Processing
Feature Extraction
Lexicon, stylistic and Semantic Feature Extraction
Dimension Reduction
Split Train and Test dataset
Build Emotion Model
Evaluate Emotion Model
Predict Emotion Labels: Russell Quadrants

Fig. 1. Method for music emotion classification for lyrics text.

A. Datasets

In this research we have used MER and Mood Lyrics dataset. The MER Dataset [4] contains 771 song lyrics extracted from the AllMusic platform. All songs are equally distributed among the four emotion quadrants of the 2-D Russell's circumflex emotion model. The dataset is generated by linking mood tags from AllMusic to words in the ANEW dictionary. The values of A and V are assigned to each word from the ANEW dictionary. Each song is categorized within a particular quadrant if all the corresponding AllMusic tags fall within that Russell quadrant. We have also used 1935 song from Mood Lyrics [19] dataset containing English songs. The Russell Model is used to annotate each song with its four Quadrants with output classes (sad, relaxed, angry, and happy). The dataset is constructed by merging three lexicons, Word Net, WorldNet-Affect, and the ANEW dictionary, to assign Valence and Arousal scores to every word in the lyrics. Emotion annotation is performed by calculating the combined Valence and Arousal values for each lyric. Based on the VA (Valence-Arousal) values, each lyric is assigned to a specific quadrant in Russell's 2-dimensional model.

B. Feature Extraction

Text feature extraction involves transforming textual data into a numerical format suitable for use in various natural language processing (NLP) tasks. The goal is to capture the essential information and characteristics of the text in a way that machine learning algorithms can process and understand. This process includes both lexical and syntactic features, as well as stylistic elements. Techniques like Term Frequency-Inverse Document Frequency *(TF-IDF)* [20] and Bag of Words (BOW) are commonly used to capture the lexical and syntactic properties of text. The Bag of Words *(BOW)* approach represents a document by breaking it down into individual words and creating a feature vector based on the frequency of each word, ignoring word order. TF-IDF is a statistical method used in natural language processing to determine the importance of a word within a document relative to a collection of documents. It's a versatile technique that can be applied in various NLP tasks, such as text classification, information retrieval, and sentiment analysis. TF-IDF calculates a score for each word in a document based on its frequency within the document and its rarity across the entire corpus. This enables it to identify the most significant terms that contribute to the document's meaning and differentiate it from others.

In addition to statistical measures, stylistic features such as the use of slang, part-of-speech *(POS)* distributions, and verb usage help capture the linguistic style and emotional tone of the text. Style features capture the linguistic characteristics of text data, such as slang words, part-of-speech distributions, and verb usage. They are related to length of sentence, length of paragraph. We implemented length *(L)* features (word count, character count, sentence count, average word length, average sentence count). Word count, character count, sentence count, average word length and average sentence length are extracted from song lyrics. Slang words are informal and non-standard language often used in music lyrics. Counting the frequency of slang using online dictionaries can provide insight into a song's lyrical style and cultural references. We used slang count *(SL)* using Online Slang Dictionary. An example of a slang word from an online urban and slang dictionary is "lit". "Lit" is an adjective that means exciting, excellent, or highly enjoyable. It is used to describe something fun, energetic, or impressive. Slang words can vary in popularity and usage over time.

Part of speech tags *(POS)* represents the frequency of various word types (e.g., noun, verb, adjective). It can help capture the syntactic structure of a sentence. Each sentence is converted to form – list of words, list of tuples. Each tuple is represented in the form *(word, tag)*). The tag is part-of-speech, which signifies whether the word is a noun, adjective, verb. For example, some common POS tags include proper noun (NNP), noun (NN), verb (VB) and adjective (JJ), coordinating conjunction (CC), and personal pronoun (PRP). We use 21 features of POS tags using the NLTK dictionary in this work—an example of POS Tagging. Example - (Dirty, NNP), (old, JJ), and (river, NN).

The frequency of BE verbs *(BE)* in positive, negative, and interrogative sentences can provide insight into the emotion tone and sentiment conveyed by the lyrics. "Be" verbs indicate a state of being. Verbs must match subjects. Present Sentence, Negative Sentence, and Interrogative Sentence are three forms of BE Verbs. BE Verbs frequency is considered in this work, along with negative words such as "not" and question mark count. Present Sentence are sentences that start with “I am”, “You are” , “He is”, “She is” , “It is”, “We are” , “You are” and “They are”. Negative Sentence have not included with Present sentences. Examples are “I am not”, “You are not”, “and He is not” and so on. Interrogative Sentences start with “Am I?”, “Are You?”, “Is he?”, etc. By combining these features, including lexical, syntactic, and stylistic elements, classification models can gain more nuanced insights into the text. This comprehensive feature set can improve the accuracy of text classification tasks by leveraging the strength of different approaches.

C. Lexicons Feature Extraction

There are six lexicon dictionaries used in this work. The Norm of Valence, Arousal, and Dominance lemmas [21] has affective words for valence, arousal, and dominance with values ranging from 0 to 1. The NRC Affect Intensity Lexicon [22] contains words with intensity scores for Ekman's basic emotions (Anger, Fear, Anticipation, Trust, Surprise, Sadness, Joy, and Disgust). The intensity score varies between 0 and 1. A score of 1 indicates the intensity of emotion is high, whereas a score of 0 indicates the intensity of the word is low. The EmoWordNet [23] lexicon comprises of words associated with eight emotions (fear, anger, joy, sadness, and surprise) with scores ranging from 0 to 1. The Synesketch lexicons [24] contain English words annotated manually with emotion weights. It uses Ekman's six basic emotions (anger, joy, surprise, sadness, disgust, and fear).The Dictionary of Affect [25] in Language comprises 8743 words annotated in 3 dimensions: pleasantness, activation, and imagery. Pleasantness is similar to valence and activation to arousal. Imagery is a word that creates a mental picture. A score of 1 indicates the difficulty in forming a mental and score of 3 indicates it is easy to form a mental picture. The DepecheMood [26] is a high coverage and high precision emotion lexicon using distributional semantics, with numerical scores associated with more than one emotion (Afraid, Amused, Annoyed, Don’t care, Happy, Inspired and Sad), obtained from crowd sourcing, it has scores ranging from 0 to 1.

1) Lexicon features for intensity emotion weight: To extract lexicon features from lyrics, the process begins with preprocessing. This involves cleaning the text to remove any non-alphabetic characters, punctuation, and extra spaces. Next, the cleaned text is tokenized, meaning it is split into individual words (tokens). Each token is then matched with its corresponding emotion dimension/intensity score from the relevant lexicons. Finally, aggregate metrics such as the mean, maximum, minimum, and standard deviation are computed based on the scores of the mapped tokens for each lyric text. It's important to note that one word can be associated with more than one emotion, depending on the type of lexicon used. This systematic approach ensures the accurate extraction of lexicon features, which can then be used for further analysis.

In previous research, we introduced emotion lexicon features using intensity/emotion weight [8] Eq. (1) to Eq. (11). The work is expanded to include Dictionary of Affect and DepecheMood Lexicon, and hybrid feature combinations. This section details the lexicon features extracted from NRC Emotion Intensity, Synesketch, EmoWordNet, and DepecheMood; these are categorical lexicons. The categorical lexicon comprises words associated with specific emotions, each of which is linked with an intensity score.

a) Average emotion intensity/weight: Definition: It is the average emotion intensity/weight of lyrics l for emotion E. The formula for average emotion intensity is shown in Eq. (1):

$$\text{Average Emotion Intensity} = \frac{\sum \text{I(W,E)}}{\text{N}} \tag{1}$$

Where I (W, E) is the intensity of the word W for emotion E and N is total number of word in the lyrics. This metric provides a general sense of how strongly an emotion is expressed throughout the lyrics. It helps in understanding the overall emotional tone.

b) Maximum emotion intensity/weight: Definition: It is the maximum intensity/emotion weight of lyrics l for emotion E. This is defined in Eq. (2):

$$\text{Maximum Emotion Intensity} = \text{Max}\big(\text{I(W, E)}\big) \tag{2}$$

Where I (W, E) is the intensity of word W for emotion E and N is total number of word in the lyrics. This metric identifies the peak emotional intensity in the lyrics, showing the highest level of emotional expression.

c) Minimum emotion intensity/weight: Definition: It is the minimum intensity/emotion weight of lyrics l for emotion E, as defined in Eq. (3):

$$\text{Minimum Emotion Intensity} = \text{Min}(\text{I(W, E)}) \tag{3}$$

Where *I (W, E)* is the intensity of word W for emotion E and N is total number of word in the lyrics. This metric highlights the least intense emotional expression in the lyrics.

d) Threshold emotion word count(TEWC): Definition: It is a metric that quantifies the number of words in the lyrics, denoted as 'W, that have an intensity score greater than a predetermined threshold limit, 'TH', for a specific emotion, ‘E’. This is shown in Eq. (4):

$$\text{TEWC (E, I, TH)} = \Sigma\ (\text{W} \in \text{l})\ [\text{I(W, E)} > \text{TH}] \tag{4}$$

Where *TEWC (E, l, TH),* represents the count of words per emotion E with a threshold for a given emotion 'E' in the lyrics 'l'. I (W, E) represents the intensity of word 'W' for the specific emotion 'E' and W ∈ I. From Eq. 4, [I(W, E) > TH] is a function that evaluates to 1 if the intensity of the word 'W' for emotion 'E' is greater than the threshold 'TH'. [I (W, E) > TH] evaluates to 0 if condition is not met. This formula calculates the total count of words in the lyrics that have an intensity surpassing the specified threshold TH for the given emotion ‘E’. Emotion Intensity score is the strength of emotional expression in the lyrics. Words surpassing the threshold are likely to have a substantial impact on the overall emotional tone. By counting words that exceed the threshold, TEWC helps identify which parts of the lyrics contribute significantly to a particular emotion. TEWC can enhance music recommendation systems by considering emotional relevance. Researchers can compare TEWC across different songs, genres, or artists to understand variations in emotional content.

e) Emotion proportion within class: Definition: It measures the proportion of a particular emotion within a specific class relative to its overall occurrence in all classes of lyrics l for emotion E, as shown in Eq. (5):

EPC = (Count of emotion within class) / (Total count of emotion across all classes)

$$\text{EPC} \ = (\text{CE}) / |\text{E}| \qquad (5)$$

Where CE denotes the count of a specific emotion within a particular class and |E| signifies the total count of the specific emotion across all classes. This formula quantifies the relative proportion of a specific emotion within a class by dividing the count of that emotion in the class by its total count across all classes.

TABLE I. SAMPLE CALCULATION FOR LEXICON FEATURES DERIVED FROM NRC EMOTION INTENSITY LEXICON WITH JOY, SADNESS AND TRUST

Word	Joy	Sadness	Trust
joyous	0.9	0	0.23
elated	0.8	0	0
cheerful	0.95	0	0
TEWC (E, I, 0.6)	3	0	0

Let's consider the Table I provided for emotions "Joy," "Sadness," and "Trust" and their corresponding intensity score.

- Using equation (1), the average Intensity for the emotion “Joy” is calculated as (09 + 0.8 + 0.95) =0.883.
- The minimum intensity score for the emotion "Joy”, using Eq. (3) is 0.8 and the maximum intensity score for the emotion "Joy" calculated using Eq. (2) is 0.95.
- For Threshold Emotion Word Count (TEWC) calculation we apply Eq. (4). TEWC (Joy, 0.9, 0.6) calculates the number of words in the lyrics with intensity greater than the threshold value (0.6) for the emotion "Joy". Using Eq. (4), TEWC (Joy, 0.9, 0.6) gives result of three and TEWC (Trust, 0.23, 0.6) gives result of zero.

2) Lexicon features for dimension emotions: This section describes lexicon features extracted from Norm of Valence, Arousal, and Dominance lemma and Dictionary of Affect. The Russell three dimension model contains dimensions of valence, arousal and dominance.

a) Average valence: It is the sum of all Valence scores divided by the total no of words (N)) in lyrics. This is shown in Eq. (6):

$$\text{Average Valence} = \frac{\Sigma \text{V}}{\text{N}} \qquad (6)$$

Where Valence V is non zero and N is total number of word in the lyrics. The average valence score represents the average emotional positivity or negativity expressed in the lyrics. This metric provides a general sense of how strongly valence is expressed throughout the lyrics.

b) Average arousal: Definition*:* It is the sum of all arousal values (A) divided by the total no of words (N) in lyrics. This is defined in Eq. (7):

$$\text{Average Arousal} = \frac{\Sigma \text{A}}{\text{N}} \qquad (7)$$

Where Arousal A is non zero and N is total number of word in the lyrics. The average arousal score represents the average calmness or excitement expressed in the lyrics. This metric provides a general sense of how strongly an arousal is expressed throughout the lyrics.

c) Average dominance: Definition: It is the sum of all dominance values (D) divided by the number of words (N) in lyrics. This is defined in Eq. (8):

$$\text{Average Dominance} = \frac{\Sigma \text{D}}{\text{N}} \qquad (8)$$

Where Dominance D is non zero and N is total number of word in the lyrics. The mean dominance score represents the average level of control or power expressed in the lyrics. This metric provides a general sense of how strongly dominance is expressed throughout the lyrics.

d) Standard deviation: The standard deviation provides a measure of the dispersion or variability of the scores around the mean, indicating how consistent the emotional expression is across the lyrics. It is calculated for Valence V, Arousal A and Dominance D values of all tokens extracted from lyrics where VAD values are non-zero. Given set of dimension scores X={x1, x2, x3….an} for lexicon L. Standard deviation for score x is shown in Eq. (9):

$$\sigma = \frac{1}{N}\sum_{i=1}^{n}(x_i \ - x_{mean} \ )^2 \qquad (9)$$

Where σ is Standard deviation, if we apply this to valence, arousal and dominance score from lyrics. For given set of score for S= {s$_1$, s2…..Sn} for valence, arousal and dominance. This is defined in Eq. (10) and (11):

$$\sigma = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{n}}(\text{s}_\text{i} \ - \ \text{s}_{\text{mean}} \ )^2 \qquad (10)$$

$$\text{s}_{\text{mean}} \ = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{n}}(\text{s}_\text{i}) \qquad (11)$$

Where N is the total number of score and i_s is the score of (valence, arousal, or dominance). $\text{S}_{\text{mean}}$is the mean (average) of all the scores.

e) Average imagery: Definition: It is sum of all imagery scores divided by total number of words (N) in lyrics. This is defined in Eq. (12):

$$\text{Average Imagery} = \frac{\Sigma \text{Im}}{\text{N}} \qquad (12)$$

Where Imagery Im is non zero and N is total number of word in the lyrics. The mean imagery score represents the average level of mental picture evoked by the lyrics. Higher mean imagery indicates that, on average, the lyrics are more capable of creating clear mental pictures for the listener. The average scores for Pleasantness and Activation are determined in a similar manner. Additionally, the minimum and maximum scores for Pleasantness, Activation, and Imagery are also calculated. We use a consistent method to ascertain the mean, minimum and maximum values for arousal, dominance, activation, imagery, and pleasantness in lyrics. These metrics are crucial for understanding the range and variation of emotional and sensory expressions. Minimum values indicate the least intense expressions, reflecting calmness, weakness, passivity, lack of vividness, and negative affect. In contrast, maximum values capture the most intense peaks of power, energy, vividness, and positive affect. For example, minimum arousal corresponds to low emotional intensity, while maximum arousal signifies high intensity. Similarly, minimum valence reflects negative emotions, whereas maximum valence indicates positivity. This method offers a nuanced view of lyrical dynamics for deeper interpretation.

For emotion classification, we use notations as Lyrics Length Features: *L*, Slang Count: *SL,* POS tags: *POS*, BE Verbs: *BE,* Norm of Valence, Arousal and Dominance features: *AF1*, Depeche Mood features: *BF1*, EmoWordNet features: *BF2*, Synesketch features: *BF3*, Emotion Intensity features: *BF4*, Dictionary of Affect features: *BF5*.

D. Dimension Reduction

The main goal of dimension reduction [7] is to simplify the dataset without losing critical information. PCA achieves this by identifying the principal components of the data, which are linear combinations of the original features that capture the most variance. The variance values associated with each principal component reflect the amount of information that component carries; higher variance indicates more information. These variance values are calculated using Eigen values derived from the covariance matrix of the data. The first principal component accounts for the greatest variance, with each subsequent component capturing progressively less. By selecting only the most significant principal components, we can effectively reduce the dimensionality of our dataset while retaining essential information.

E. Classifiers

Random Forest [27] and Gradient Boosting are popular machine learning algorithms used for classification and regression tasks. Random forest is an ensemble method that combines multiple decision trees to make predictions. It works by constructing multiple decision trees using subsets of the training data and random feature subsets. The predictions from each tree are then combined to make a final prediction. Random Forest is known for its accuracy and ability to handle high-dimensional datasets. Gradient boosting is another ensemble method that works by iteratively adding weak learners, typically decision trees, to a model to improve its performance. Gradient boosting uses a loss function to determine the error of the current model and update the model by fitting a new tree to the residuals of the previous model. Gradient boosting is known for its high accuracy and ability to handle complex datasets but can be computationally expensive and prone to over fitting. The choice of algorithm depends on the specific problem and dataset at hand.

F. Performance Metric

Classification accuracy is a metric used to evaluate the performance of a machine learning model in classification tasks. It measures the proportion of correctly predicted instances (both true positives and true negatives) to the total number of instances in a dataset. In simple terms, classification accuracy represents how often the model makes correct predictions.

IV. Experiments and Results

We conducted extensive experiments for classification, both with and without dimension reduction, using the Mood Lyrics dataset (D1) and the MER dataset (D2). We employed Random Forest and Gradient Boost classifiers for the classification task. Due to the large number of experiments, it is not feasible to present all the results here. The features notation used for classification is given in the Lexicon sub section. The Emotion output classes are four Russell Quadrants (sad, relaxed, angry, and happy). Classification experiments were conducted using datasets D1 and D2, employing Random Forest and Gradient Boost classifiers. These experiments utilized stylistic, lexical, and length features derived from lyric text, with the feature acronyms detailed in Method subsection C.

A. Emotion Classification Accuracy by Russell Quadrants

For Table II, the combination of (Bag Of Words, Norm of Valence, Arousal and Dominance features, BE Verbs, Slang Count, Dictionary of Affect features and POS tags) and (Bag Of Words, POS tags, Norm of Valence, Arousal and Dominance, Dictionary of Affect, Synesketch features, Emotion Intensity, Slang Count, BE Verbs) features gives an accuracy of 93.16% for Gradient Boost, using Dataset D1.

TABLE II. EMOTION CLASSIFICATION BY QUADRANTS FOR DATASET D1, USING BAG OF WORDS (BOW), STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
BOW + AF1 + BE + SL + BF5 + POS	64.5	93.16
BOW + AF1 + BF4 + BE + SL + BF5	61.49	80.68
BOW + AF1 + SL + BF5	57.76	91.92
BOW + AF1 + BF3 + B5 + BF4	50.93	50.93
BOW + AF1 + BF4 + SL + BF5	50.31	53.41
BOW + POS + AF1 + BF5 + SL	52.17	50.31
BOW + POS + AF1 + BF5 + BF3 + BF4 + SL + BE	54.03	93.16
BOW + POS + A1 + B4 + B5 + B3	54.03	90.68
BOW + POS + A1 + B4 + B5 + B3 + SL	57.76	63.35

In the Table III, the combination of (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity features, Slang Count , Dictionary of Affect features) and (TF-IDF, Norm of Valence, Arousal and Dominance features, Slang Count: and Dictionary of Affect features)for Gradient Boost classifier gives an accuracy of 91.92%, using Dataset D1.

TABLE III. EMOTION CLASSIFICATION BY QUADRANTS FOR THE DATASET D1, USING TF-IDF, STYLE, AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
TF-IDF + AF1 + BF4 + BE + SL + BF5 + POS	60.24	91.30
TF-IDF + AF1 + BF4 + BE + SL + BF5	57.76	91.92
TF-IDF + AF1 + SL + BF5	57.96	91.92
TF-IDF + AF1 + BF3 + BF5 + EF4	55.90	54.65
TF-IDF + AF1 + BF4 + SL + BF5	56.52	47.20
TF-IDF + POS + AF1 + BF5 + SL	54.65	53.41
TF-IDF + POS + AF1 + BF5 + BF3 + BF4 + SL + BE	55.27	88.81
TF-IDF + POS + AF1 + BF4 + BF5 + BF3	57.14	90.68
TF-IDF + POS + AF1 + BF4 + BF5 + BF3 + SL	54.03	55.27

For Table IV, the combination of (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity, BE Verbs, Slang count, Lyrics Length and Synesketch) features give 78.49 % accuracy for Gradient Boost. The combination of (TF-IDF, POS tags , Norm of Valence, Arousal and Dominance features, Dictionary of Affect features, Synesketch features and Emotion Intensity features) and (TF-IDF, Norm of Valence, Arousal and Dominance features, Emotion Intensity features, BE Verbs, Slang Count , Lyrics Length Features and Synesketch features) give accuracy of 77.42% for Random Forest with Dataset D2.

TABLE IV. EMOTION CLASSIFICATION BY QUADRANTS FOR DATASET D2, USING TF-IDF, STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
TF-IDF + POS + A1 + DAL + BF3 + BF4	65.59	69.89
TF-IDF + AF1 + BF4 + SL + BF3 + LF	63.44	65.59
TF-IDF + AF1 + BF4 + SL + BF3 + LF + DAL	64.51	66.67
TF-IDF + POS + AF1 + BF4 + SL + BF3	74.19	75.26
TF-IDF + POS + AF1 + DAL + BF3 + BF4	77.42	76.34
TF-IDF + POS + AF1 + DAL + BF3 + LF	68.81	75.26
TF-IDF + POS + AF1 + DAL + SL + BF4	67.74	73.11
TF-IDF + POS + AF1 + DAL + BF3 + BF4	65.59	69.89
TF-IDF + AF1 + POS + BF3 + BF2 + SL	77.42	75.26
TF-IDF + AF1 + POS + BF3 + BF2 + SL + L	72.04	68.81
TF-IDF + AF1 + BF4 + BE + SL + L + BF3	72.04	78.494

For Table V, the combination of (Bag of Words, POS Tags, Norm of Valence, Arousal and Dominance features, Dictionary of Affect features, Slang Count and Emotion Intensity) gives accuracy of 75.26% for Gradient Boost and 72.04% for Random Forest, using Dataset D2.

B. Principal Component Analysis (PCA)

PCA transforms the original data into a new set of variables, known as principal components, which are uncorrelated and ordered so that the first few retain most of the variation present in all the original variables. The variance helps determine how much information each principal component holds from the original dataset. Explained Variance shows how much of the data's variation a principal component captures. Cumulative Variance sums up the variances of components to show the total variation explained.

TABLE V. EMOTION CLASSIFICATION BY QUADRANT FOR THE DATASET D2, USING BOW, STYLE AND LEXICON FEATURES

Feature Sets Combinations	Random Forest	Gradient Boost
BOW + POS + AF1 + DAL + BF3 +BF4	72.04	69.89
BOW + AF1 + BF4 + SL + BF3 + L	68.81	62.36
BOW + AF1 + BF4 + SL + BF3 + L + DAL	64.5	64.5
BOW + POS + AF1 + BF4 + SL + BF3	70.96	67.74
BOW + POS + AF1 + DAL + BF3 + BF4	72.04	69.89
BOW + POS + AF1 + DAL + BF3 + LF	62.36	64.5
BOW + POS + AF1 + DAL + SL + F4	72.04	*75.26*
BOW + POS + AF1 + DAL + BF3 + BF4	43.08	49.46
BOW + POS + AF1 + DAL + BF3 + BF4	72.04	69.89
BOW + AF1 + POS + BF3 + BF2 + SL + L	58.06	62.36
BOW + AF1 + BF4 + BE + SL + LF + BF3	66.66	62.365

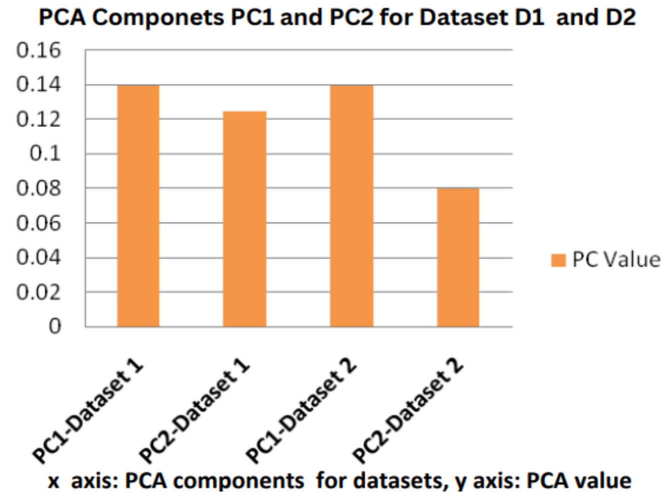


Fig. 2. PC1 and PC2 for dataset D1 and D2.

The bar chart in Fig. 2 shows the values of two main components, PC1 and PC2, for two datasets. The x axis shows PCA components for (D1 and D2); and y axis shows PCA value. In Dataset 1, PC1 explains a moderate amount of variance, showing it captures a significant part of the data's variability. PC2 has less variance than PC1, meaning the remaining variability isn't as focused in one direction. Dataset 2 has a slightly higher variance for PC1 than Dataset 1, indicating more of its variability is due to the first principal component. PC2's variance in Dataset 2 is even lower than in Dataset 1. Result indicates that Dataset 2's variability is more concentrated in the first component. While both datasets have similar PC1 values, Dataset 2 relies more on PC1, showing a more unidirectional variance. In contrast, Dataset 1 has a higher variance in PC2, indicating its variability is spread across more dimensions. This means Dataset 1 has a more balanced variance distribution between the components, while Dataset 2 leans more on PC1 for its variance explanation.

The PCA analysis of music lyrics for emotion classification highlights key features influencing the first two principal components across two datasets. In Dataset 1 (D1), PC1 is primarily influenced by character count, word count, sentence count, singular nouns (NN), and personal pronouns (PRP), pointing to detailed and intense emotional expression. PC2 is shaped by personal pronouns (PRP), adverbs (RB), present tense verbs (VBP), and general verbs (VB), indicating immediacy in emotional experiences. For Dataset 2 (D2), PC1 is similarly influenced by character count, word count, singular nouns (NN), personal pronouns (PRP), and sentence count, suggesting a focus on emotional intensity. However, PC2 is defined by personal pronouns (PRP), proper nouns (NNP), singular verbs (VBP), conjunctions (CC), and past tense verbs (VBD), which highlight introspection and narrative elements. By using 40 PCA components, the analysis effectively reduces dimensionality while capturing the nuanced features of emotional expression in lyrics, enhancing emotion classification accuracy.

Table VI shows result for emotion classification by quadrants for D1 using PCA, achieving 90% explained variance with 40 components. The Gradient Boost classifier performed the best with an accuracy of 98.13%. Table VII shows that the emotion classification by quadrants for Dataset 2 using PCA, achieving 85% explained variance with 40 components. The Gradient Boost classifier performed the best with an accuracy of 65.59%.

TABLE VI. EMOTION CLASSIFICATION BY QUADRANTS FOR D1 USING PCA

PCA Explained Variance	PCA Components	Classifiers	Accuracy
90%	40	Random Forest	96.89%
90%	40	Gradient Boost	98.13%
90%	40	Decision Tree	96.89%

TABLE VII. EMOTION CLASSIFICATION BY QUADRANTS FOR D1 USING PCA

PCA Explained Variance	PCA Components	Classifiers	Accuracy
85%	40	Random Forest	59.13%
85%	40	Gradient Boost	65.59%
85%	40	Decision Tree	47.31%

V. DISCUSSION

The classification experiments for emotion detection in lyrics using various feature sets and classifiers demonstrated notable differences in performance across datasets, classifiers, and feature combinations. The emotion detection experiments in song lyrics showed notable differences in performance across datasets, classifiers, and feature combinations. For Dataset D1, the Gradient Boost classifier consistently outperformed Random Forest, achieving the highest accuracy of 93.16% with a Bag of Words approach combined with stylistic, lexical, and length features, including Norm of Valence, Arousal, and Dominance, BE verbs, slang count, dictionary of affect features, and POS tags. This indicates that a comprehensive feature set is effective for capturing the nuances of emotion in lyrics. Using TF-IDF instead of BOW yielded similar results, with a slight decrease to 91.92% accuracy for Gradient Boost, showing that text representation (BOW or TF-IDF) has a marginal impact when combined with these features. In contrast, Random Forest did not perform well on D1, it achieved highest accuracy of 64.5%.

For Dataset D2, the results were less consistent. Gradient Boost achieved the best accuracy of 78.49% using TF-IDF, AF1, BE verbs, slang count, lyrics length, and Synesketch features, highlighting the importance of emotional content and syntactic complexity. Random Forest also performed relatively well on D2, with an accuracy of 77.42% using a different feature set. The result suggests shows that, it might be more effective for this dataset due to differences in data distribution or emotional expression. Principal Component Analysis (PCA) for dimensionality reduction further clarified the effectiveness of feature sets. For D1, using 40 PCA components, Gradient Boost achieved an accuracy of 98.13%, showing that PCA effectiveness in capturing the most informative aspects of the data. However, for D2, the highest accuracy with PCA was 65.59%, indicating that PCA might not capture the emotional nuances as effectively as in D1. Overall Dataset D1 generally achieved higher classification accuracy than D2. The difference in performance is likely due to differences in dataset characteristics. The results emphasize the need for comprehensive feature sets and suitable algorithms to capture emotional complexity in lyrics. PCA is effective for reducing dimensionality and enhancing classifier performance, especially for Dataset D1, though its impact varies with different datasets and feature representations.

The use of PCA highlights its value in reducing dimensionality and enhancing classifier performance when sufficient variance is retained, as seen with Dataset D1. However, the mixed results for Dataset D2 indicate that while PCA can improve efficiency, it may also overlook some nuances if the components do not adequately represent the data's emotional characteristics. In music emotion recognition from lyrics, our method has shown superior effectiveness compared to both transformer-based and traditional approaches. For Dataset D1, the Gradient Boost classifier achieved an accuracy of 93.16%, which improved to 98.13% with PCA.

This study builds upon our previous work [8] by incorporating a more diverse set of features and applying Principal Component Analysis (PCA) for dimensionality reduction. While the prior study primarily used lexicon-based features, the current work integrates a wide range of NLP, textual, stylistic, and lexical features. This expanded feature set includes POS tags, DAL, slang count, BE verbs, length-based attributes, and affective dictionaries. The use of the expanded feature set allows for a more comprehensive analysis of emotional nuances in song lyrics. The application of PCA for dimensionality reduction is another novel aspect, resulting in improved accuracy and model interpretability. The results of this approach are significant, with an accuracy of 98.13% achieved on Dataset D1 using PCA, compared to 93.16% without PCA. This performance surpasses previous methods in the field. For instance, Transformer-based models like XLNet [12] achieved 83% accuracy on the Mood Lyrics dataset, while traditional Bag of Words and TF-IDF approaches reported accuracies [29] between 65.49% and 67.98%. Malheiro et al. attained an F1 score of 73.6% for MER Lyrics [4] using the SVM classifier. This study achieved a remarkable accuracy of 98.13% on Dataset D1, surpassing previous research that employed multi-modal deep learning approaches by Pyrovolakis et al. (2022) [28] and Transformer-model [12]. Jiddy Abdullah [29] achieved classification accuracies of 76% and 83% using Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers, respectively, on the Mood Lyrics dataset. Malheiro et al. (2016) [4] utilized emotionally relevant features; this study's uses broader integration of stylistic, lexical, and content features which led to superior performance. Accuracy gains are crucial for fine-grained emotion detection applications, but they come with increased computational cost and complexity. While this may limit scalability, the benefits often outweigh the added resources, making the improvement worthwhile in many practical scenarios.

Despite these improvements, the study has limitations. These include potential issues with generalizability across different datasets and a lack of exploration into the impact of various hyper parameters for the Gradient Boost classifier and other dimension reduction methods. Additionally, the focus on accuracy may have overlooked other important metrics. This study shows that using various features with methods like PCA to reduce complexity effectively captures the emotions in song lyrics, improving upon the accuracy and representation issues of past approaches.

VI. Conclusion

In conclusion, this research significantly advances the field of emotion analysis in music lyrics by filling a critical gap with a focused study on the textual representation of emotions. The development of an NLP-based framework and the introduction of hybrid lexicon features have proven to be instrumental in enhancing emotion detection capabilities. The NLP-based framework and novel lexicon features significantly enhance emotion detection. Advanced classification techniques like Gradient Boost and Decision Trees, along with innovative feature extraction and dimension reduction, have achieved high accuracy in emotion classification. The superior performance of Gradient Boost classifiers highlights the importance of choosing appropriate algorithms for handling high-dimensional data. This research provides valuable insights into understanding emotional expressions in music. However, it also faces limitations, such as reliance on specific features and the resource-intensive nature of initial computations for large datasets. Additionally, model generalizability across diverse datasets remains a challenge. Future research directions could include exploring other emotion datasets, feature engineering and integrating deep learning approaches.

References

[1] A. Kawakami, K. Furukawa, and K. Okanoya, “Music evokes vicarious emotions in listeners,” Front. Psychol., vol. 5, p. 431, May 2014, doi: 10.3389/fpsyg.2014.00431.

[2] M. de Witte, A. da Silva Pinho, G.-J. Stams, X. Moonen, A. E. R. Bos, and S. van Hooren, “Music therapy for stress reduction: a systematic review and meta-analysis,” Health Psychol. Rev., vol. 16, no. 1, pp. 134–159, 2022, doi: 10.1080/17437199.2020.1846580.

[3] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, “Emotion detection of textual data: An interdisciplinary survey,” in 2021 IEEE World AI IoT Congress (AIIoT), May 2021, pp. 255–261, doi: 10.1109/AIIoT52608.2021.9454192.

[4] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Emotionally-relevant features for classification and regression of music lyrics,” IEEE Trans. Affective Comput., vol. 9, no. 2, pp. 240–254, 2016, doi: 10.1109/TAFFC.2016.2598569.

[5] M. M. Bradley and P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, vol. 30, no. 1, Univ. Florida, Tech. Rep. C-1, 1999.

[6] J. A. Russell, “Core affect and the psychological construction of emotion,” Psychol. Rev., vol. 110, no. 1, pp. 145–172, 2003, doi: 10.1037/0033-295X.110.1.145.

[7] K. N. K. Singh, D. Devi, H. M. Devi, and A. K. Mahanta, “A novel approach for dimension reduction using word embedding: An enhanced text classification approach,” Int. J. Inf. Manage. Data Insights, vol. 2, no. 1, p. 100061, 2022, doi: 10.1016/j.jjimei.2022.10006.

[8] A. Ara and V. Rekha, “Analyzing the impact of lexicon-based features for emotion classification,” Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 4, pp. 677–687, 2024.

[9] V. R. Revathy, A. S. Pillai, and F. Daneshfar, “LyEmoBERT: Classification of lyrics’ emotion and recommendation using a pre-trained model,” Procedia Comput. Sci., vol. 218, pp. 1196–1208, 2023.

[10] S. A. Suresh Kumar and R. Rajan, “ Transformer-based automatic music mood classification using multi-modal framework,” J. Comput. Sci. Technol., vol. 23, 2023.

[11] P. Padmane, K. Agrahari, R. Kesharwani, K. Mohitkar, S. Khan, and N. P. Kamale, “Prediction of song mood through lyrics,” Open Access Repository, 9(6), 137-141,2022

[12] Y. Agrawal, R. G. R. Shanker, and V. Alluri, “Transformer-based approach towards music emotion recognition from lyrics,” in European Conference on Information Retrieval, Mar. 2021, pp. 167–175, doi: 10.48550/arXiv.2101.02051.

[13] F. H. Rachman, R. Sarno, and C. Fatichah, “Hybrid approach of structural lyric and audio segments for detecting song emotion,” Int. J. Intell. Eng. Syst., vol. 13, no. 1, Feb. 2020, doi: 10.22266/ijies2020.0229.09.

[14] C. Jin, Z. Song, J. Xu, and H. Gao, “Attention-based Bi-DLSTM for sentiment analysis of Beijing Opera lyrics,” Wireless Commun. Mobile Comput., vol. 2022, p. 1167462, 2022.

[15] Bhattacharya and K. V. Kadambari, “A multimodal approach towards emotion recognition of music using audio and lyrical content,” arXiv preprint, vol. 1811, p. 05760, 2018.

[16] Naseri, S., Reddy, S., Correia, J., Karlgren, J., & Jones, R. (2022). The Contribution of Lyrics and Acoustics to Collaborative Understanding of Mood. arXiv:2207.05680, https://doi.org/10.48550/arXiv.2207.0568

[17] L. Parisi, S. Francia, S. Olivastri, and M. S. Tavella, "Exploiting synchronized lyrics and vocal features for music emotion detection," arXiv preprint arXiv:1901.04831, 2019.

[18] Y. Song and D. Beck, "Modeling Emotion Dynamics in Song Lyrics with State Space Models," arXiv preprint arXiv:2210.09434, 2022.

[19] E. Cano and M. Maurizio, "MoodyLyrics: A Sentiment Annotated Lyrics Dataset," in Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, Hong Kong, 2017, pp. 118–124.

[20] F. Karabiber , TF-IDF-Term Frequency- Inverse Document Frequencey , https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/

[21] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," Behavior Research Methods, vol. 45, pp. 1191-1207, 2013, doi: 10.3758/s13428-012-0314-x.

[22] S. M. Mohammad, "Word Affect Intensities," in Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), Miyazaki, Japan, May 2018, doi: 10.48550/arXiv.1704.08798.

[23] G. Badaro, H. Jundi, H. Hajj, and W. El-Hajj, "EmoWordNet: Automatic expansion of emotion lexicon using English WordNet," in Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 2018, pp. 86–93, doi: 10.18653/v1/S18-2009.

[24] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An open source library for sentence-based emotion recognition," IEEE Transactions on Affective Computing, vol. 4, no. 3, pp. 312–325, 2013, doi: 10.1109/T-AFFC.2013.18.

[25] C. Whissell, M. Fourner, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: IV. Reliability, validity, and applications," Perceptual and Motor Skills, vol. 62, no. 3, pp. 875-888, 1986.

[26] J. Staiano and M. Guerini, "Depechemood: A lexicon for emotion analysis from crowd-annotated news," arXiv preprint arXiv:1405.1605, 2014.

[27] Logistic Regression Vs. Random Forest Classifier https://www.geeksforgeeks.org/logistic-regression-vs-random-forest-classifier/

[28] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Multi-modal song mood detection with deep learning," Sensors, vol. 22, no. 3, p. 1065, 2022.

[29] J. Abdillah, I. Asror, Y. F. A. Wibowo, et al., "Emotion classification of song lyrics using bidirectional LSTM method with GloVe word representation weighting," Journal RESTI (Rekayasa Sistem Dan Teknologi Informasi), vol. 4, no. 4, pp. 723–729, 2020.

Art Image Color Sequence Data Processing Method Based on Artificial Intelligence Technology

Xujing Zhao, Xiwen Chen, Jianfei Shen*
School of Visual Arts, Hunan Mass Media Vocational and Technical College, Changsha 410000, China

***Abstract*—With the traditional quality enhancement methods cannot control the best field density range resulting in too large threshold value of colour difference in art works. Therefore, a research on art works quality enhancement based on image processing technology is proposed. The CIE L* a* b* color space model is established to divide the color magnitude and then transform the color space by RGB space conversion model. On this basis, the quality of art works is enhanced according to the process of the quality enhancement of art works. As considering that the actual density is not within the control range, the image processing technology is used to separate targets to solve this problem. In the experiment, Adobe Illustrator CS6 software was used to make the experimental color target and six test samples were selected to test whether the distribution results of the two methods in different degree of color difference perception met the quality enhancement requirements. The experimental results show that the quality enhancement effect of the proposed method is better and more in line with the design requirements.**

Keywords—Artworks; quality enhancement; image processing technology; color space model; target separation; experimental color target

I. Introduction

Graphic artworks mainly come from original works and reproductions. The materials of the original works are various. The most common ones are canvas, paper, wood board, stele, and murals [1]. The mid-term steles and murals are immovable, so the digitalization method is special. Scan with a camera or contactless. Photography or non-contact scanning are inseparable from the intervention of optical lenses, so certain physical deformation is inevitable [2]. Of course, modern optical lenses have developed to a high level of technology, so generally, only professional lens components can be used to avoid obvious deformation problems. Modern printing has completely adopted the digital technology process to handle each production link, and the rough division is input and output. The first is the digital input link. The reason for the different colors of the same set of prints is that different digital materials are used. Although the same technology and materials are used in printing, the image files used are not produced by a unified image digital standard [3]. So it is difficult to adjust to the same color effect in post-processing. For art professionals, the expressiveness of the color of the work is sometimes even more important than the accuracy of the graphics [4]. The sharpest criticism, I have heard in the digitization of art works is the overall tone of many digitized images and the original work. They are all reversed. For example, the overall tone of the original work is warmer, and the result obtained after scanning is cooler. Moreover, such pictures are obtained by professionals after repeatedly adjusting the color curve of the scanner [5]. The problem of color deviation between the digitized art image and the original is a problem that exists widely in the book digitization project but has not been significantly improved [6].

Today, with the rapid development of computer graphics and image processing technology, image processing technology is used for image beautification processing and applied to the field of art to improve the imaging quality and visual sense of the image [7]. The beautification processing of art images mainly includes image noise reduction, image information fusion, image enhancement and image white balance deviation compensation, etc. In the process of landscape image acquisition, due to the influence of the light intensity and the physical characteristics of the acquisition device itself, the original images usually collected need to be processed in post-processing [8]. Image beautification technology is the key to post-art image processing. Image understanding is a science that automatically extracts information through computer programs [9]. It is the process of recognizing image content and interpreting and expressing it in natural language. Specifically, image description can be understood as a special translation process from image to text, in which multi-party knowledge in the fields of CV, NLP and machine learning is used [10]. For humans, the process of "seeing pictures and talking" from images to texts is a very basic ability, but for machines, it is to build associations between completely different modeling systems. The process of automatic image understanding firstly obtains the cognition of the image content through the model, and secondly, the binary cognition needs to be converted into the form of natural language for output [11].

The traditional method of enhancing the quality of works of art is achieved by calculating the dot gain value of works of art. Although the color reproducibility can be enhanced, the optimal solid density cannot be determined, resulting in ghosting and deformation of works of art [12]. Ghosting and deformation are caused by failing to control the difference in density within a certain range. In order to solve this problem, this paper proposes a method for enhancing the quality of artworks based on image processing technology [13]. In the study of domestic and foreign cases, it is found that the use of color defect detection, noise processing, color division and other methods can effectively enhance the quality of various art works. In order to overcome the instability of the subjective visual inspection method used in the detection, some foreign scholars, based on the image processing technology, combined with the concept of machine vision, carried out real-time detection with the help of detectors, and realized the online control of the quality of art works [14]. On this basis, adjust the acquired information, and use multiple color line scan cameras to capture the surface information of the artwork, compare it with the standard artwork, and analyze and detect defects point by point. However, this method still lacks certain practicality. In this design, this method is used as a reference to complete the design. In terms of expression, the model has further requirements for readability [15]. The generated text description should not only reflect the information of the image, but also take into account the fluency of the language [16]. After years of development, current image understanding technologies can usually be classified into three methods: template-based methods, retrieval-based methods, and end-to-end automatic understanding methods based on deep learning [17].

The quality enhancement method of artworks based on image processing technology is a collection of various techniques that can improve the visual effect of artworks. First establish the CIE *L* a* b** color space model, divide the color magnitude, and then apply a simple algorithm to give a three-dimensional color histogram [18]. After feature extraction, detect color defects, classify images, and enhance the quality of artwork. In the research process, it is found that the application of defect identification algorithm will affect the effect of enhancing the quality of artworks [19]. For this reason, dynamic thresholds are used to set pixel points to achieve inspection-free and shorten the inspection time [20]. In order to ensure the feasibility of the established method, the method of numerical comparison is used in the experiment to verify whether the quality enhancement method of art works based on image processing technology meets the needs of use.

Section I of the study elaborated on the background description of various technologies that can enhance the visual effects of artworks. Section II analyzed the construction of machine learning in the context of artificial intelligence by various scholars. Compared to re-creating works of a specific style through extensive training, correctly establishing the artistic understanding ability of machines is more in line with the long-term needs of artificial intelligence cognitive intelligence. Section III describes the artistic image processing scheme. Section IV conducted simulation experiments, analyzed experimental parameter settings, established an RGB space conversion model, and then used raw data to model color targets according to experimental requirements. Section V summarizes the entire text. The enhancement effects of the two methods were compared numerically. The experimental results indicate that the method meets the design requirements.

II. Related Work

The research on the processing method of artistic image color sequence data based on artificial intelligence technology is highly relevant for improving the digital efficiency of artistic creation, accurately controlling color transition and expression, and promoting the deep integration of the art field and technological innovation. This study aims to analyze and optimize the color sequence of artistic images through intelligent algorithms, providing powerful color processing tools for artists and designers, and promoting the personalized and diversified development of artistic creation. Assisting the artistic process with artificial intelligence technology is an emerging topic in recent years, but it has a vitality that cannot be ignored. The study in [21] organizes the computational aesthetics task of combining art and artificial intelligence from the perspective of human cognition. The particularity of art makes researchers pay special attention to what kind of performance machines can have in this field, how to make machines have the same understanding and aesthetic power as humans, and whether existing technologies can endow machines with intelligence in this area. In the understanding of artificial intelligence art, there are two schools of thought. One school believes that artificial intelligence replaces only repetitive labor and techniques, and cannot replace human free will to create. The other faction recognizes the autonomous creativity of artificial intelligence and believes that artificial intelligence has the potential to replace human designers to achieve independent creation. Repetitive work is just replaced with a higher priority; and the process of being replaced takes longer for links that require more creativity. The researchers conducted a variety of research explorations against these two viewpoints.

The advantages of deep learning in tasks such as feature selection and image matching provide new options for image restoration. Currently, it has great auxiliary significance for tasks such as restoration of historical relics and video special effects rendering. Bertalmio et al. used the idea of partial differentiation [22] to complete and repair the image, and use the information outside the entire area to repair the contour inward. Pathak et al. Combining the codec architecture and generative adversarial network technology [23], the image information is complemented by the judgment of the prediction map, and the image can also be filled in the case of large-area information defects, and the completed image can be semantically. It matches the original image. Iizuka et al. added the judgment of local information on the basis of the former, and the output repair map takes into account global and local information, which is semantically consistent with the overall semantics, and also strengthens the performance of details locally [24]. Image restoration guided by artificial intelligence technology is especially suitable for the restoration of stained and damaged paintings.

When expanding the research boundaries of artificial intelligence (AI) systems' emotional response capabilities, it innovatively shifts the focus to AI's ability to understand and respond to emotional tones in artistic image color sequence data [25]. The research aims to explore in depth how the emotional tendencies contained in color sequences in artistic images affect the judgment and decision-making process of AI systems, particularly through an AI based method for processing color sequence data in artistic images. In order to construct this research scenario, some scholars carefully selected a series of art promotional images of vacation rentals as experimental materials [26]. These images not only showcase different vacation environments, but also convey diverse emotional atmospheres through changes in color sequences. Using advanced image analysis techniques, extract color sequence data from images and design algorithms to identify and analyze the emotional tones (such as positive, negative, or neutral) contained in these color sequences. There are still many controversies about the intelligence and originality of works obtained by intelligent machines through style imitation.

Compared with re-creating works of a specific style through a lot of training, the correct establishment of the machine's artistic understanding ability is more in line with the long-term needs of artificial intelligence cognitive intelligence. Although Gatys et al.'s art transfer model and subsequent Prisma applications have achieved significant results in art style imitation, there is still controversy over the intelligence and originality of intelligent machines obtaining works through style imitation. This reflects the limitations of artificial intelligence in the field of artistic creation - that is, current technology is more focused on imitation rather than creation. Considering that errors are prone to occur in color difference space conversion, the RGB space conversion model is used to improve the conversion accuracy. On this basis, according to the process of improving the quality of artistic works, enhance the quality of artistic works. Finally, the enhancement effects of the two methods were compared numerically.

III. Art Picture Processing Scheme

A. Color Scale Division

In order to avoid the difference in color evaluation, the color difference evaluation standard is divided into the demand level, as shown in Table I.

TABLE I. COLOR DIFFERENCE EVALUATION STANDARD

Color Difference Value	Color Difference Evaluation	Visual Perception Cognition
0-1.5	Slight chromatic aberration	Very little difference
1.5-2.5	Small color difference	Slight difference
2.5-3.5	Small color difference	The difference is obvious
3.5-4.5	Large color difference	The difference is very obvious
4.5 以上	Large color difference	Strong difference

Select a device-dependent color space according to Table I. In information collection, images in RGB color mode should be used to enhance the connectivity of components. The value in the RGB color space is 0~256, which is divided into 257 levels. When the value of R, G, B is 0, the color space will appear black. If the value of R, G, B is 255, R, G, B will appear cyan in turn. If the values of R, G, and B is 255, 0, and 255 in sequence, the color space presents magenta. In terms of grayscale information, the values of R, G, and B are the same, the color components will increase, and the image will gradually change from black to white. The color range of the RGB color space is small, and color defects will appear when the magnitude is divided. For this reason, the CIE *L** *a** *b** color space model is established, and the expression is:

$$\begin{cases} L^* = 106\left(Y/Y_n\right)^{1/3} \\ a^* = 500\left[f\left(X/X_n\right) - f\left(Y/Y_n\right)\right] \\ b^* = 200\left[f\left(X/X_n\right) - f\left(Z/Z_n\right)\right] \end{cases} \quad (1)$$

Where *L** represents the brightness of the artwork, *a** represents the red-green axis, *b** represents the yellow-blue axis. It can be seen from Formula (1) that in the RGB color space, *L** represents the brightness of the luminance axis, and when the color space is 0, it appears black. When the color space is 50, it is white. When the color space is between 0 and 50, it is gray. X, Y, Z represent the tristimulus values under the color of the light source. Color defects can be detected by applying Formula (1), and color levels can be accurately divided.

B. Color Space Conversion

After dividing the color level, in order to realize the nonlinear conversion of the color space, it is necessary to use image processing technology to measure the source color patch data of the RGB space model. The specific conversion steps are as follows:

1) Use image processing technology to build two spatial transformation models, and select standard sample colors. The RGB space conversion model is shown in Fig. 1.

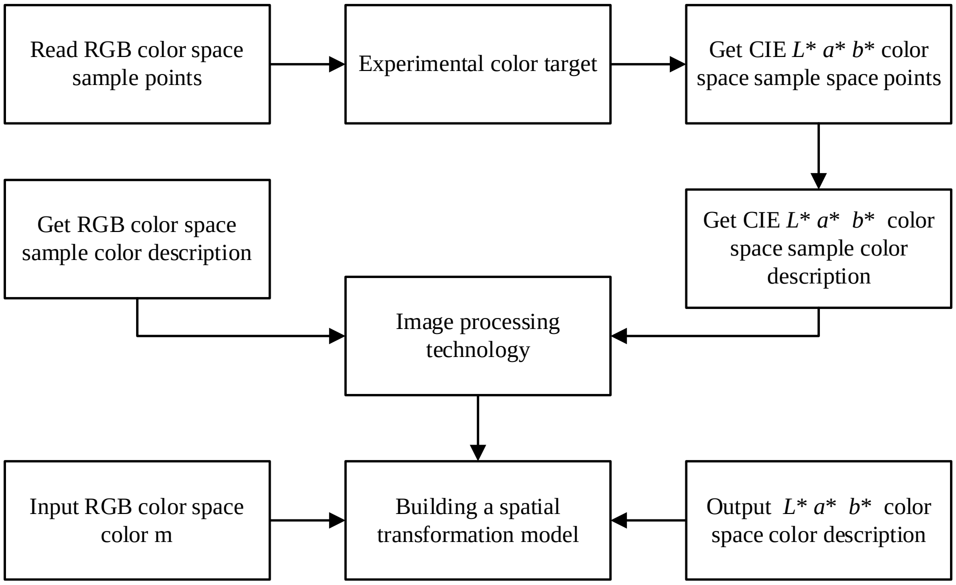

Read RGB color space sample points
Experimental color target
Get CIE L* a* b* color space sample space points
Get RGB color space sample color description
Get CIE L* a* b* color space sample color description
Image processing technology
Input RGB color space color m
Building a spatial transformation model
Output L* a* b* color space color description

Fig. 1. RGB space conversion model.

Applying the RGB space conversion model, the L*, a*, b* expressions are:

$$a^* = \sum a_1 R^i G^j B^k \tag{2}$$

$$b^* = \sum a_2 R^i G^j B^k \tag{3}$$

Where *i*, *j*, *k* represents the order of R, G, B, respectively. And a_0, a_1, a_2 represent the color description data of R, G, B, and its value is not determined according to the color that needs to be adjusted.

Apply Formula (2) to Formula (4) to complete the conversion of color space. Without the choice of sample point requirements when solving for polynomial coefficients, it is impossible to ensure the accuracy of color conversion for all regions. In order to reduce the occurrence of such phenomena, a polynomial method is used to select an original point and add it to the RGB space conversion model to enhance the accuracy of the model transformation. The specific formula is:

$$W_{RCB} = a_0 + a_1 R + a_2 G + a_3 B \tag{4}$$

Where W_{RGB} represents the balanced color after model transformation; a3 represents the color description data of *R*, *G*, *B*, the same as a0~a2.

2) According to the determined model, calculate the optimal parameters.

3) After completing the conversion of the chromaticity information of the original color space, the data of each cube vertex is measured by using a three-dimensional look-up table, and a cube is formed to ensure that the overall color space has high precision. Then divide the RGB space into N^3 cubes, and measure the *L**, *a**, *b** values of the vertices of each cube. Calculate the surrounding vertices according to the RGB space conversion table, and use the vertices on the three-dimensional geometric solid to calculate the converted color data.

4) Color space conversion using neural network. The schematic diagram of the neural network structure is shown in Fig. 2. The conversion of color space is completed by transmitting information through the input layer, hidden layer, and output layer.

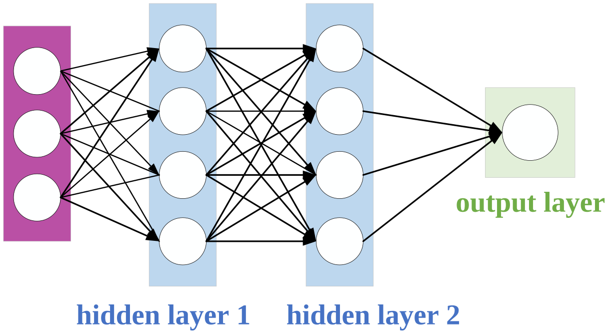


Fig. 2. Schematic diagram of neural network structure.

C. Methods of Enhancing the Quality of Works of Art

First divide the color scale of the artwork, and then convert the color space. In order to enhance the quality of art works, use color defect detection to determine the color difference threshold.

The specific operation steps are as follows:

Step 1: Determine the color defect threshold, and measure the *L**, *a**, *b** values according to user requirements.

Step 2: Set the values of luminance information ΔL, Δa and Δb according to the threshold requirement of the total color difference ΔE_{ab}. If one item does not meet the requirements, it means that the quality enhancement effect of the artwork is unqualified. If the threshold point of the total color difference ΔE_{ab} exceeds 60% of the original image, it is an overall color cast or a local color cast.

Step 3: Step3: Judging whether it is in line with the quality enhancement effect of art works, the judgment is based on:

1) When $\Delta L > 5.05$, it means that the artwork to be inspected is brighter than the standard artwork, and the quality enhancement effect is ideal.

2) When $\Delta L < -5.05$, it means that the artwork to be inspected is darker than the standard artwork, and the quality enhancement effect is not good.

3) When $\Delta a > 4.65$, it means that the artwork to be inspected is redder than the standard artwork, and the quality enhancement effect is better.

4) When $\Delta a < -4.65$, it means that the artwork to be inspected is greener than the standard artwork, and the quality enhancement effect is better.

5) When $\Delta b > 3.21$, it means that the artwork to be inspected is more yellow than the standard artwork, and the quality enhancement effect is good.

6) When $\Delta b < -3.21$, it means that the artwork to be inspected is bluer than the standard artwork, and the quality enhancement effect is poor.

Step 4: After screening, select qualified samples, and use image processing technology to enhance the visual effect of art works to meet quality standards.

Taking into account the problem of image difference, image processing technology is used to separate the target from the back. Let the artwork after the difference shadow be $f(i,j)$, the binarization threshold is Th, and after binarization, the artwork is $g(i,j)$, and the calculation formula is:

$$f(i,j) = \begin{cases} 0, & f(i,j) < \text{Th} \\ 255, & f(i,j)..\text{Th} \end{cases} \tag{5}$$

Set the Th threshold to 56, and apply Formula (6) to eliminate the shadow. Combining the above steps, the design of a method for enhancing the quality of art works based on image processing technology is completed.

D. Image Beautification Processing Implementation

This paper analyzes the application of computer graphics and image processing technology in art, and on the basis of image noise reduction preprocessing, the beautification of night scene images is carried out. This paper proposes a night scene image beautification processing technology based on illumination multiple chromatic aberration compensation, adopts the illumination adaptive equalization technology to optimize the white balance of the night scene image, and searches for the threshold value according to the edge contour information of the image. The neighborhood of the feature point *i* in the image noise distribution is defined by N_{i} as:

$$N_i = \left\{ i' \in S \middle| \left[dist\left(i, i'\right) \right]^2 ,, \; r, i \neq i' \right\} \tag{6}$$

In the white balance processing of night scene images, the illumination adaptive equalization technology is used to optimize the white balance of night scene images. Use $dist\left(i, i'\right)$ to describe the distance of pixels in the neighborhood of the night scene image to be beautified, *r* is a constant, and get the multiple color difference kernels of the night scene image to be beautified.

$$R_i = \frac{1}{\gamma_{ije}} \sum_j d\left(\Box i - j \Box_2\right) l\left(g_i - g_{j1}\right) \tag{7}$$

The morphological segmentation method is used to deconvolute the image completely blindly. In the sparse prior regularization distribution space, the adaptive equalization constraint function of image color difference is obtained according to the prior knowledge of the blur kernel as follows:

$$\min kp = \lambda g_i + \beta g_j \tag{8}$$

where, λ and β are regularization parameters. Based on the mathematical expression of the blur kernel, the white balance optimization result of the night scene image is obtained as:

$$\mathrm{LRT} = \min kp \left| R_i \cdot N_i \right|^s \tag{9}$$

Where *s* is the Total Variation (TV) term of the fuzzy kernel.

The illumination chromatic aberration of night scene adopts the method of shadow area and brightness area segmentation to compensate for contour shadow deviation. Through image multi-threshold segmentation, the segmentation curve of shadow area and brightness area is obtained and described as:

$$G_{\mathrm{mar}}\left(\overrightarrow{x_i}\right) = \frac{\sum_{j=1}^{p} G_j\left(\vec{x}_i\right) / G_j^{\max}}{p} \tag{10}$$

where, $i \in \{1,2,\ldots,\mathrm{N}\}$ is the sequence value. In the measurement of the whole image, the time cost of the regular term of the blurred image is measured, and the blur kernel of the high-frequency image *y* is obtained as:

$$\Omega = \left\{ \vec{x} \in \mathrm{s} \middle| \; g_j(\vec{x}),, \, 0, \, j = 1, 2, \cdots, l; h_j(\vec{x}) = 0, \, j = l+1, l+2, \cdots, p \right\} \tag{11}$$

The fuzzy kernel is updated using the unconstrained iterative reweighted least squares (IRLS) algorithm, which is expressed as:

$$\min \vec{y} = \vec{f}(\vec{x}) = \left(f_1(\vec{x}), f_2(\vec{x}), \cdots, f_m(\vec{x}) \right) \tag{12}$$

Where $\vec{x} = \left(x_1, x_2, \cdots, x_n\right) \in X \subset \Re^n$ is the initial value vector of each feature point of the beautified night scene image, *X* is the decision space for optimal evolution, *Y* is the target space of optimal evolution. The image contour shadow deviation compensation realizes the superposition of the low-frequency information of the image, and the objective function of the image contour shadow deviation compensation is:

$$G(\vec{x}, \vec{y}) = \min \vec{y} \sum_3^4 G_{\mathrm{nor}}(\vec{x}) + \Omega^3 \tag{13}$$

IV. Simulation Test

A. Experimental Parameter Settings

First establish the RGB space conversion model, then use the original data, model the color target according to the experimental requirements, and use professional equipment to detect the quality enhancement effect of art works. According to the data requirements of the experiment, the RGB source color space is divided into nine levels, which are 0, 20, 34, 56, 68, 124, 234, 245, and 276 respectively. Use Adobe Illustrator CS6 software to make experimental color targets. In order to ensure the validity of the built RGB space conversion model, it is necessary to select six works of art, and the color difference information of the works is shown in Fig. 3.

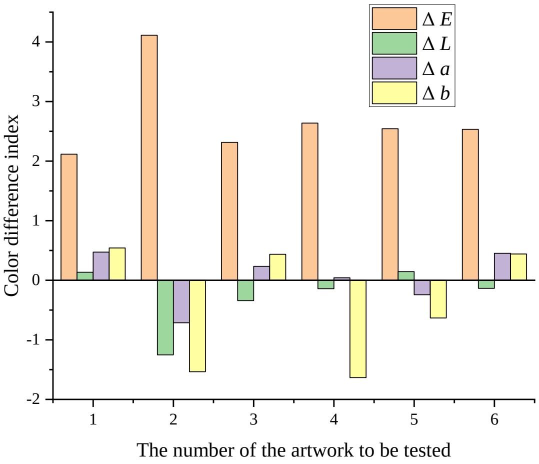


Fig. 3. The color difference information of the works.

Taking the quality enhancement method of artworks in this study as the test object of the experimental group, and the traditional enhancement method as the test object of the control group, the quality enhancement processing of six works of art was carried out by using the two methods, and the processed works were different in color. The color difference thresholds under different brightness and brightness are detected, and the results are shown in Fig. 4 and Fig. 5.

It is known that the human eye has the strongest ability to recognize colors in the mid-lightness range. According to the test results in Fig. 4, it can be seen that the color difference threshold of the control group increases gradually with the increase of the brightness L. However, the chromatic aberration threshold of the images in the experimental group did not increase significantly due to the increase of the lightness L, but stabilized within a fixed range. There is a chromatic aberration deviation. Looking at Fig. 5 again, the control group has a larger color difference threshold due to the increase of color C, and the color difference of the visible image is not obvious. The chromatic aberration threshold of the experimental group is also within a small range. It can be seen that the two works of the experimental group did not show extreme differences in chroma after adding color. Based on the above test results, it can be seen that after the method proposed this time enhances the artwork, whether it is adjusting the brightness or enhancing the color of the artwork, the chromatic aberration threshold of the artwork has always been in a stable range. It can be seen that the proposed enhancement method is better.

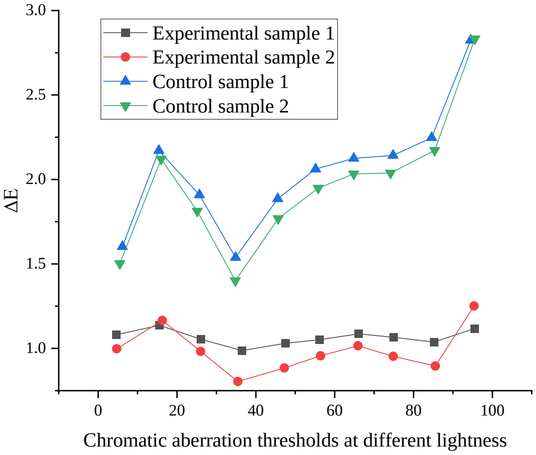


Fig. 4. Image quality enhancement effect experiment 1.

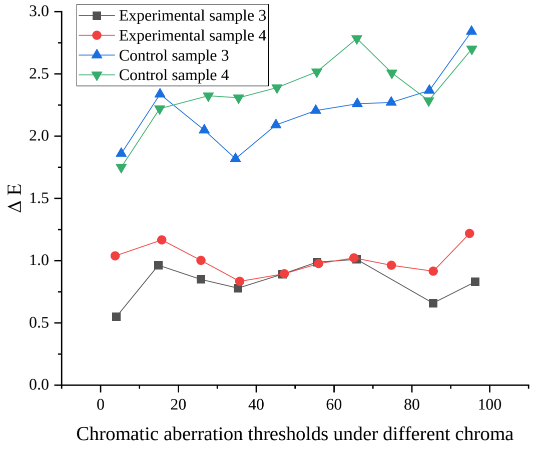


Fig. 5. Image quality enhancement effect experiment 2.

B. Analysis of Simulation Results

In order to explore the practicability of the proposed method, the method proposed in this paper, study [4] and study [5] are used to simulate and compare the digital painting images of R, G and B colors. The experiment adopts objective evaluation, and uses some mathematical variables that can objectively present the essential characteristics of the image, such as mean, variance or histogram characteristics as evaluation criteria, and analyzes the pros and cons of the method. Firstly, the influence of the number of iterations on the performance of the method is analyzed, and the image angle error before and after color correction is used as the measurement standard. The simulation results are shown in Fig. 6.

It can be seen from Fig. 6 that with the increase of the number of iterations, the angle errors of the three methods will gradually decrease. When the number of iterations is greater than or equal to 8, the image angle errors remain stable. The angle error of the method in this paper is always lower than the two literature methods, showing better algorithm performance. The reason for this phenomenon is that the method in this paper uses the optical flow-oriented feature method to effectively divide the image color distortion area, which reduces the probability of color correction angle errors to a certain extent.

The following is the comparison of the average values of R, G, and B colors between the three methods and the original image after the image color correction. After the double analysis of horizontal and vertical, it is evaluated whether the method in this paper has application advantages. The comparison results are shown in Fig. 7 to Fig. 9.

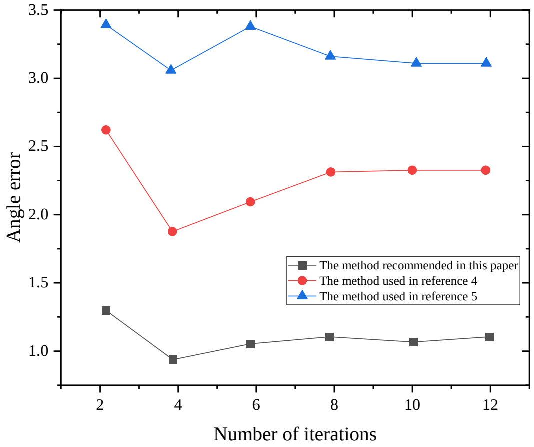


Fig. 6. The effect of the number of iterations on the performance of the method.

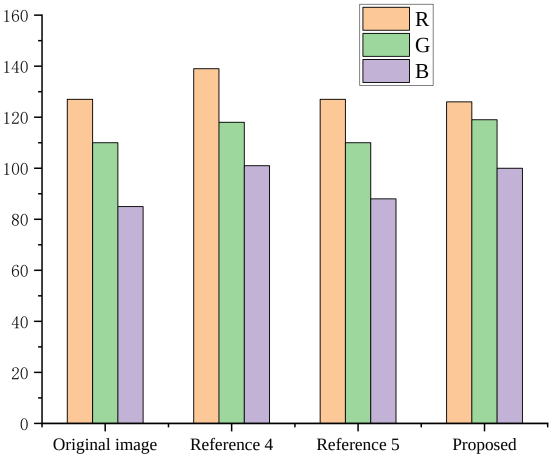


Fig. 7. Color correction comparison of reddish images.

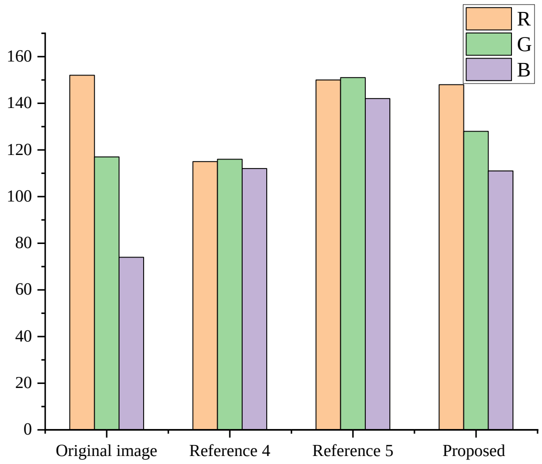


Fig. 8. Color correction comparison of greenish image.

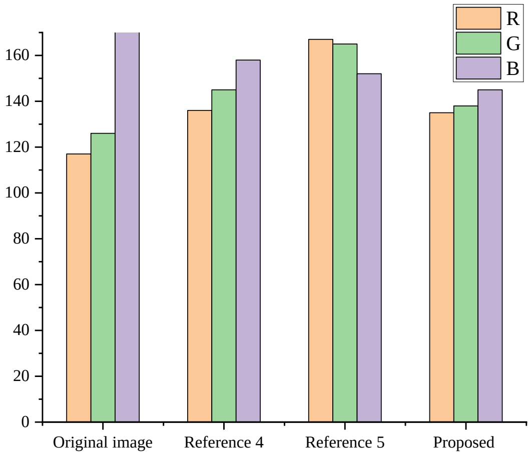


Fig. 9. Color correction comparison for bluish images.

From the experimental data in the three figures, it can be seen that the method in this paper has a good calibration effect for different color cast images. After the correction, the color distribution of the three color channels with uneven distribution of the initial color cast image is more even, and the difference of the mean value of the three color channels and the degree of color cast are reduced. The two methods in the literature are not universal, and may result in failure for some color cast images, and the correction accuracy needs to be improved. Although this method effectively divides color distortion areas using optical flow oriented feature methods, in some complex or boundary blurred images, this division may still not be accurate enough, resulting in poor local color correction effects. Although the experimental results show that the angle error tends to stabilize after increasing the number of iterations to a certain extent, excessive iterations will increase computational costs and reduce processing efficiency. Therefore, determining the optimal number of iterations is a problem that requires further optimization. The method proposed in this article performed well in experiments, but in practical applications, there may be more types of color cast images. How to ensure that the method is equally effective for these images is a problem that needs further verification. By improving the segmentation algorithm and iterative optimization strategy of color distortion areas, the angle error after color correction can be further reduced and the correction accuracy can be improved. Adjust and optimize algorithm parameters for different types of color cast images, improve algorithm adaptability and robustness, and ensure effectiveness and stability in various complex scenarios. On the premise of ensuring calibration accuracy, optimize algorithm implementation, reduce unnecessary computational overhead, improve processing efficiency, and meet the needs of real-time processing or large-scale dataset processing.

V. Conclusion

Aiming at the problems existing in traditional art quality enhancement methods, this paper proposes a new method, which uses image processing technology to enhance the visual effect of the quality of art works. Considering that errors are prone to occur in the conversion of color difference space, the RGB space conversion model is used to enhance the conversion accuracy. On this basis, the quality of art works is enhanced according to the quality enhancement process of art works. Finally, the enhancement effect of the two methods is compared by numerical comparison. The experimental results show that the proposed method meets the design requirements. However, it is found in the experiment that the quality enhancement method of art works based on image processing technology has higher requirements on the surrounding environmental factors, and further research is needed to improve it. When introducing the method proposed in this article, its theoretical basis can be elaborated in more detail. How to specifically apply the feature method for optical flow to the division of color distortion areas, and why this method can effectively reduce the angle error of color correction. In addition to using mean, variance, and histogram features, it is also possible to consider introducing more evaluation metrics such as Structural Similarity Index (SSIM), Peak Signal to Noise Ratio (PSNR), etc. to more comprehensively evaluate image quality. The paper mentions that art quality enhancement methods based on image processing technology have higher requirements for surrounding environmental factors. It is necessary to specify which environmental factors include (such as lighting, temperature, humidity, etc.), and explore how to reduce the impact of these factors in practical applications.

Fund Projects

1) Hunan Provincial Social Science Achievement Evaluation Committee in 2023:“Research on the Design of Suitable Aging Environment for Old Residential Areas in Huxiang Region,” project number: XSP2023YSC081.

2) Hunan Provincial Social Science Achievement Evaluation Committee in 2023:“Research on the epidermal texture of tourism villages in Meishan area under the background of rural revitalization,” project number: XSP2023YSC062.

3) The Natural Science Foundation of Hunan Province in 2021:“Research on Tactile Graphic Design and Application for Visually Impaired Population in the Era of Graphic-Based Reading,”project number: 2021JJ60022.

REFERENCES

[1] Anantrasirichai N, Bull D. Artificial intelligence in the creative industries: a review[J]. Artificial Intelligence Review, 2021: 1-68. DOI: 10.1007/s10462-021-10039-7.

[2] Partel V, Kakarla S C, Ampatzidis Y. Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence[J]. Computers and electronics in agriculture, 2019, 157: 339-350. DOI: 10.1016/j.compag.2018.12.048.

[3] Adegun A, Viriri S. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art[J]. Artificial Intelligence Review, 2021, 54(2): 811-841. DOI: 10.1007/s10462-020-09865-y.

[4] Patrício D I, Rieder R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review[J]. Computers and electronics in agriculture, 2018, 153: 69-81. DOI: 10.1016/j.compag.2018.08.001.

[5] Murtaza G, Shuib L, Abdul Wahab A W, et al. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges[J]. Artificial Intelligence Review, 2020, 53(3): 1655-1720. DOI: 10.1007/s10462-019-09716-5/.

[6] Yang J, Wang C, Jiang B, et al. Visual perception enabled industry intelligence: state of the art, challenges and prospects[J]. IEEE Transactions on Industrial Informatics, 2020, 17(3): 2204-2219. DOI: 10.1109/TII.2020.2998818.

[7] Yas Q M, Zadain A A, Zaidan B B, et al. Towards on develop a framework for the evaluation and benchmarking of skin detectors based on artificial intelligent models using multi-criteria decision-making techniques[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2017, 31(03): 1759002. DOI: 10.1142/S0218001417590029.

[8] Shi W, Zhang M, Zhang R, et al. Change detection based on artificial intelligence: State-of-the-art and challenges[J]. Remote Sensing, 2020, 12(10): 1688. DOI: 10.3390/rs12101688.

[9] Brahimi M, Boukhalfa K, Moussaoui A. Deep learning for tomato diseases: classification and symptoms visualization[J]. Applied Artificial Intelligence, 2017, 31(4): 299-315. DOI: 10.1080/08839514.2017.1315516.

[10] Bini S A. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?[J]. The Journal of arthroplasty, 2018, 33(8): 2358-2361. DOI: 10.1016/j.arth.2018.02.067.

[11] Wäldchen J, Mäder P. Machine learning for image based species identification[J]. Methods in Ecology and Evolution, 2018, 9(11): 2216-2225. DOI: 10.1111/2041-210X.13075.

[12] Gao Z, Zhang H, Dong S, et al. Salient object detection in the distributed cloud-edge intelligent network[J]. IEEE Network, 2020, 34(2): 216-224. DOI: 10.1109/MNET.001.1900260.

[13] Liu D, Liu F, Xie X, et al. Accurate prediction of responses to transarterial chemoembolization for patients with hepatocellular carcinoma by using artificial intelligence in contrast-enhanced ultrasound[J]. European radiology, 2020, 30(4): 2365-2376. DOI: 10.1007/s00330-019-06553-6.

[14] Willemink M J, Noël P B. The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence[J]. European radiology, 2019, 29(5): 2185-2195. DOI: 10.1007/s00330-018-5810-7.

[15] Kanagasingam Y, Xiao D, Vignarajan J, et al. Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care[J]. JAMA network open, 2018, 1(5): e182665-e182665. DOI: 10.1001/jamanetworkopen.2018.2665.

[16] Górriz J M, Ramírez J, Ortíz A, et al. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications[J]. Neurocomputing, 2020, 410: 237-270. DOI: 10.1016/j.neucom.2020.05.078.

[17] Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19[J]. IEEE reviews in biomedical engineering, 2020, 14: 4-15. DOI: 10.1109/RBME.2020.2987975.

[18] Darko A, Chan A P C, Adabre M A, et al. Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities[J]. Automation in Construction, 2020, 112: 103081. DOI: 10.1016/j.autcon.2020.103081.

[19] Aggarwal P, Papay F A. Artificial intelligence image recognition of melanoma and basal cell carcinoma in racially diverse populations[J]. Journal of Dermatological Treatment, 2022, 33(4): 2257-2262. DOI: 10.1080/09546634.2021.1944970.

[20] Ke Q, An S, Bennamoun M, et al. Skeletonnet: Mining deep part features for 3-d action recognition[J]. IEEE signal processing letters, 2017, 24(6): 731-735. DOI: 10.1109/LSP.2017.2690339.

[21] Sollini M, Antunovic L, Chiti A, et al. Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics[J]. European journal of nuclear medicine and molecular imaging, 2019, 46(13): 2656-2672. DOI: 10.1007/s00259-019-04372-x.

[22] Suri J S, Agarwal S, Gupta S K, et al. Systematic review of artificial intelligence in acute respiratory distress syndrome for COVID-19 lung patients: a biomedical imaging perspective[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(11): 4128-4139. DOI: 10.1109/JBHI.2021.3103839.

[23] Albahri O S, Zaidan A A, Albahri A S, et al. Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects[J]. Journal of infection and public health, 2020, 13(10): 1381-1396. DOI: 10.1016/j.jiph.2020.06.028.

[24] Chen X, Xu C, Yang X, et al. Gated-gan: Adversarial gated networks for multi-collection style transfer[J]. IEEE Transactions on Image Processing, 2018, 28(2): 546-560. DOI: 10.1109/TIP.2018.2869695.

[25] Mubashshir Bin Mahbub, Afia Ayman. Utilising Artificial Intelligence-Prospects and Obstacles For Modern Businesses. Malaysian E Commerce Journal. 2024; 8 (1): 23-28. http//doi.org/10.26480/mecj.01.2024.23.28.

[26] Kelvin Leong, Anna Sung (2023). An Exploratory Study of How Emotion Tone Presented in A Message Influences Artificial Intelligence (AI) Powered Recommendation System. Journal of Technology & Innovation, 3(2): 80-84.

AI-Driven Prioritization Techniques of Requirements in Agile Methodologies: A Systematic Literature Review

Aya M. Radwan, Manal A. Abdel-Fattah, Wael Mohamed
Information System Department, Faculty of Computers and Artificial Intelligence, Cairo, Egypt

***Abstract*—Software requirements are the foundation of a successful software development project, outlining the customer's expectations for the software's functionality. Conventional techniques of requirement prioritization present several challenges, such as scalability, customer satisfaction, efficiency, and dependency management. These challenges make the process difficult to manage effectively. Prioritizing requirements by setting criteria in order of importance is essential to addressing these issues and ensuring the efficient use of resources, especially as software becomes more complex. Artificial intelligence (AI) offers promising solutions to these challenges through algorithms like Machine Learning, Fuzzy Logic, Optimization, and Natural Language Processing. Despite the availability of reviews on conventional prioritization techniques, there is a notable gap in comprehensive reviews of AI-based methods. This paper offers a systematic literature review (SLR) of AI-driven requirements prioritization techniques within Agile methodologies, covering 32 papers published between 2010 and 2024. We conducted a parametric analysis of these techniques, identifying key parameters related to both the prioritization process and specific AI methods. Our findings clarify the application domains of various AI-based techniques, offering crucial insights for researchers, requirement analysts, and stakeholders to choose the most effective prioritization methods. These insights consider dependencies and emphasize the importance of collaboration between stakeholders and the development team for optimal results.**

***Keywords*—Requirement analysis; requirement prioritization; agile; fuzzy logic; machine learning; optimization**

I. INTRODUCTION

Requirement Engineering (RE) is a crucial component of software engineering. It involves identifying and understanding client needs, the contexts in which the system will be developed, modeling, analyzing, prioritizing, and documenting stakeholder requirements. RE ensures that these documented requirements align with agreed-upon specifications and manages the evolution of changing requirements [1], [2], [3].

Requirement Engineering (RE) faces numerous challenges, particularly in the areas of human communication and collaboration, and understanding and clarifying requirements. Effective communication and collaboration between project teams and customers are critical, yet often fraught with difficulties, including conflicts among stakeholders and the need for active involvement from all parties. Understanding and clarifying requirements present additional challenges, as ensuring high-quality requirements and well-defined user stories can be complex and time-consuming [4], [5].

Agile Software Development (ASD) introduces specific challenges for requirements prioritization (RP) techniques, such as stakeholder conflicts, changes in priority lists leading to rework, and factors influencing requirement selection during the RP process. Among the most significant challenges are managing and coordinating distributed teams, prioritizing requirements, maintaining proper documentation, adapting to changing and over-scoping requirements, and effectively organizing processes while monitoring progress and incorporating feedback. Addressing these challenges requires robust strategies and tools to enhance communication, clarify requirements, and streamline prioritization and management processes [6].

A key aspect of Requirement Engineering (RE) is requirements prioritization (RP). As the name implies, RP involves identifying the most critical requirements for implementing a successful system. It is an iterative process involving complex decision-making activities that support the development of a high-quality system within defined constraints. RP ensures the correct ordering of requirements based on stakeholder perceptions, by rearranging them according to various criteria such as importance, cost, penalty, and risk. Active stakeholder involvement is crucial for achieving accurate prioritization results [7].

While addressing the prioritization challenge, these techniques have been applied without considering the hierarchical dependencies among requirements, such as stakeholder needs and their derived requirements. Derived requirements are the detailed requirements extracted from stakeholder needs, often in the form of use cases or non-functional requirements [6] [8].

Most existing requirement prioritization techniques lack scalability, dependability, continuous prioritization, rank updating, feedback handling, and comprehensive implementation of methods or algorithms. In Agile development, where requirements change rapidly, a continuous requirement prioritization process is essential [8] [9].

The objective of this paper is to systematically review and analyze AI-driven techniques, such as Fuzzy Logic, Machine Learning, NLP, and Optimization, for requirements prioritization in Agile methodologies. It aims to compare these techniques based on their strengths, weaknesses, and effectiveness in addressing challenges like scalability, stakeholder collaboration, and requirement dependencies. Additionally, the paper seeks to evaluate the impact of AI on improving the prioritization process and provide recommendations for future research.

This section presents the idea of the requirement prioritization in Agile methodology and discusses the common challenges. In the second section the background of the used techniques to prioritize the requirements in Agile methodology is presented. In the third section the research method is described, and the requirements prioritization analysis is illustrated at the taxonomy, the research questions are defined and research strategy. The fourth section provides the surveyed techniques and reviews the 32 papers that were sorted in subsection by the type of technique. In section five the Evaluation criteria are defined by clearing the strengths, weakness and limitations of the research. Finally, a comprehensive of the evaluation criteria is presented in the tables for each technique.

II. Background

Requirement prioritization in Agile methodology can be improved using advanced techniques such as Fuzzy Logic, Machine Learning, Unsupervised Learning, Optimization, and Natural Language Processing (NLP). In this section, an illustration of the four techniques is defined.

Fuzzy Logic has the unique ability to process numeric input and linguistic information simultaneously. It applies a nonlinear transformation to the input feature vector, resulting in a single numeric output, thus transforming numeric values into other numeric values [10]. Defining requirements into precise numerical values is a significant challenge. Fuzzy Logic allows for the accurate definition of concepts such as low cost, high quality, and high progress, making it a powerful tool for managing uncertainty. It does this by using human language and allowing for interpretation of assertions that are not entirely precise or incorrect [11].

Machine Learning (ML) provides a range of algorithms and strategies to imitate the way of human think by acquiring knowledge from data, it can be used to address software engineering challenges using supervised, unsupervised, and reinforcement learning approaches. Machine Learning facilitates software development by utilizing predictive models to make informed decisions on algorithms and features. The process of applying Machine Learning requires understanding the problem at hand, collecting relevant data, performing preprocessing tasks, and conducting comprehensive evaluations [12]. Software requirements engineering, and Machine Learning are major areas of research, where Machine Learning is applied in many software engineering procedures. Machine Learning approaches, such as text feature extraction and algorithms, are employed to categorize and prioritize software needs, utilizing the large amount of data and domain expertise gathered throughout the development process [12]. Natural Language Processing is interconnected to Machine Learning (ML) as NLP heavily relies on Machine Learning approaches, utilizing algorithms and models derived from ML. The objective of Natural Language Processing (NLP) or computational linguistics is to devise algorithms and methodologies that construct computational models with the ability to analyze natural languages. These models are specifically built to carry out important functions, such as enabling the exchange of information between humans and machines, improving communication among individuals, or simply analyzing and interpreting text or speech [13].

Optimization is a strategy to select the most efficient solution from a set of options, considering constraints and intended results. By considering aspects such as minimizing costs or maximizing efficiency. Optimization is a commonly employed strategy in diverse fields such as mathematics, computer science, engineering, economics, and operations research. It improves decision-making processes and successfully solves complex challenges [14].

In Agile methodologies, various AI techniques offer unique solutions to key challenges in requirements prioritization. Fuzzy Logic stands out in managing uncertainty and imprecise requirements by allowing for flexible decision-making under ambiguous conditions, thus improving the overall accuracy of decisions [11]. Machine Learning (ML) techniques are particularly effective in handling large datasets and automating the prioritization process, which addresses scalability concerns and significantly reduces manual effort [12]. Natural Language Processing (NLP) plays a crucial role in automating the interpretation of user stories and feedback, enhancing communication and resolving issues related to unclear or incomplete requirements. Finally, Optimization algorithms, such as AHP and PSO, provide a structured approach to balancing conflicting priorities, making them highly effective in resolving stakeholder conflicts and managing competing requirements efficiently [15][16]. By applying these techniques, the prioritization of requirements becomes more efficient, accurate, and adaptable to the dynamic nature of Agile software development.

III. Research Method

A. *Planning the Review (Preparation Stage)*

This Systematic Literature Review aims to investigate, analyze, and summarize Agile software requirements prioritization techniques, including Fuzzy Logic, Machine Learning, Natural Language Processing, and Optimization within the framework of parametric benchmarks. The review aims to define the key strengths and weaknesses and define how artificial intelligence techniques can affect requirement prioritization in Agile methodology, also explore the limitations of applying these techniques in the Agile prioritization process. Additionally, the study seeks to understand the significance of these techniques in the software development process, the role of stakeholders, and the challenges or limitations in Agile practices, while providing future directions for further research. The main objective of this Systematic Literature Review (SLR) is to identify the various techniques employed to prioritize requirements in Agile methodology.

The key of the SLR is illustrated through three phases at the taxonomy depicted in Fig. 1, to illustrate the core ideas of the paper. First, defining with requirements prioritization techniques as Fuzzy Logic, Machine Learning, Natural Language Processing and Optimization. Second, comparative analysis presented in a table to show the strength, weakness and limitations in requirements prioritization for the 32 papers. Third, the evaluation criteria to determine the effectiveness of each requirements prioritization technique.

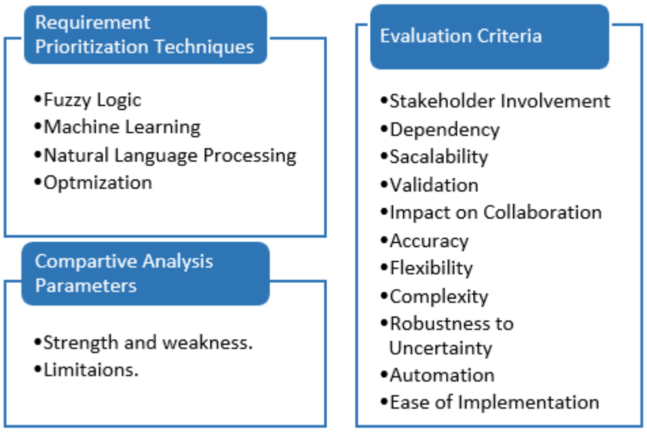


Fig. 1. Requirements prioritization taxonomy.

B. Research Question

The following research questions have been defined:

RQ1: What requirement prioritization techniques used in Agile Software Development in the research literature?

RQ2: How are the proposed techniques evaluated?

RQ3: What evaluation criteria can be used in requirement prioritization in Agile Software Development?

RQ4: What are the limitations in the current Agile Software Development technique?

RQ5: Does dependency play a role in prioritizing requirements in Agile methodology?

RQ6: Is collaboration with stakeholders considered when prioritizing requirements in Agile methodology?

RQ7: Is scalability considered when prioritizing requirements in Agile methodology?

C. Search Strategy

A comprehensive search strategy is developed to collect all research articles that are relevant to the domain of our study from a variety of online resources. The most renowned online databases and combined key terms was selected to produce search strings. Subsequently, the search was implemented to identify all relevant articles. 900 prospective studies were identified during the initial search phase. 250 studies were determined to be relevant to requirement engineering in general after the titles and abstracts were reviewed. The selection criteria focused on 32 papers specifically applying AI techniques, such as Machine Learning, Fuzzy Logic, NLP, and Optimization, to requirements prioritization in Agile software development. Studies published between 2010 and 2024 were included to ensure the review reflected the most current research. Papers were chosen based on their application of AI techniques within Agile methodologies like Scrum, Kanban, XP, and DevOps. After a thorough quality assessment, these 32 papers were selected to address the research questions, with no additional recent studies identified.

IV. Surveyed Techniques

In this section, all the surveyed techniques were reviewed. Software organizations are addressing the limitations of conventional techniques by implementing AI-based methods for software requirements prioritization in Agile. These techniques are divided into four categories: Optimization-based, Fuzzy Logic-based, Machine Learning-based, and Natural Language Processing-based ones. A detailed review is provided of each technique within these groups individually.

A. Fuzzy Logic

Borhan in [15] introduces an innovative approach to prioritizing user stories by collaborating with stakeholders in the Agile-Scrum prioritization process. The approach implements Fuzzy Logic operations to prioritize stakeholders according to predetermined criteria, thereby facilitating a more comprehensive comprehension of their contributions and concerns. This methodology is applied to an ATM system's requirements and their related user stories. A group of software experts who are experienced in Agile methodologies give feedback within their organizations for this approach. The findings suggest that including stakeholder analysis has a positive influence on the prioritization process, demonstrating the effectiveness of this method in achieving a balance between functional and non-functional user stories. By carefully considering stakeholder perspectives, this approach enhances the accuracy and efficacy of user story prioritization in Agile-Scrum projects. It guarantees a more equitable assessment of the system's requirements and the diverse demands of users.

Abusaeed in [16] suggest a quantitative framework that effectively prioritizes the identified cost factors based on the following four categories: people, initiatives, processes, and products. They conduct a systematic literature review and empirical study within the ASD context to identify and validate the cost overhead factors. Similarly, the validated factors are classified and prioritized in the present study by employing a multi-criterion decision-making Fuzzy-Analytic Hierarchy Process technique. This method effectively mitigates the subjectivity and ambiguity of the identified factors. The implementation results provide a prioritized list of cost overhead factors that would be beneficial to Agile practitioners in the context of Agile Software Development.

Ottoli in [17] introduces a novel approach to requirement prioritization that uses expert opinions presented by fuzzy linguistic labels on multiple decision criteria. Fuzzy linguistic labels are used to allocate weights to each expert and criterion, and these opinions are aggregated using a majority-guided linguistic IOWA (Induced Ordered Weighted Averaging) operator. The method contrasts requirements by comparing the aggregated expert opinions and their weighted importance. The algorithm demands users to submit opinions. However, additional empirical validation with actual users is required. Allowing selective expert opinions, incorporating consensus metrics, and investigating other (IOWA) Induced Ordered Weighted Averaging operators and T-norms are all potential future enhancements.

B. Machine Learning

Anand in [18] used the apriori algorithm to prioritize requirements. They addressed the limitations of the traditional methods used such as MoSCoW, Validated Learning, Walking Skeleton, and Business Value which frequently fail to effectively resolve stakeholder conflicts. The apriori algorithm is implemented to identify the most frequently requested requirements, to reduce conflicts between stakeholders by prioritizing requirements based on stakeholder input by continually performing join and prune functions to identify frequent items within a database of transactions.

Hudaib in [19] employs Self-organizing Maps (SOMs), a type of Unsupervised Neural Network as a method to prioritize requirements within Agile methodologies. This method is designed to classify and prioritize requirements automatically by applying patterns that are derived from historical data and stakeholder feedback. The SOM technique prioritizes requirements by grouping similar requirements based on their historical significance and characteristics. This enables project teams to identify which requirements should be addressed first. They provide a visualization to understand the priority of requirements.

Varsha in [20] improved the decision-making process of the requirements by grouping stakeholders according to their interests and perspectives. They aimed to improve the accuracy and efficiency of requirements prioritization in complex project environments. When requirements values are gathered from many individual stakeholders, it is necessary to organize stakeholders into groups to implement requirements prioritization. They involve the use of a hierarchical clustering analysis technique to create such groups from stakeholder ratings. The group weights were determined using AHP.

Belsis in [21] introduces "PBURC", a Patterns-Based, Unsupervised Requirements Clustering Framework. It utilizes Machine Learning to effectively manage data inconsistencies and validate requirements, ensuring that requirements are clustered properly using K-means. This method defines the optimal number of sprints by collaborating with stakeholders. The framework seeks to simplify the development process by prioritizing client requirements and protecting business value, despite the complex nature of distributed development environments.

Achimugu in [22] focuses on addressing the problem of prioritizing many software requirements using the k-means clustering technique. K-means is utilized to classify requirements based on the weights of their attributes, provided by project stakeholders. The effectiveness of this method was validated with the RALIC dataset, revealing that different stakeholder weights influence the clustering outcome. The approach was further refined by employing a synthetic method with scrambled centroids, proving effective in prioritizing requirements. The results indicated that this technique enhances the scalability and reliability of requirement prioritization, successfully addressing previous limitations such as rank reversals and disparity in weighted rankings.

Kumar in [23] designed an approach to enhance Agile methodology by clustering user stories using the k-means algorithm and cosine similarity. The process includes preprocessing steps like tokenization, stop-word removal, stemming, and lemmatization, followed by clustering based on similarity measures, and validating cluster cohesion with silhouette coefficient values. Experimental results show that cluster quality is improved as the number of clusters (k) increases and this reflects effectively reducing the time required for requirement implementation.

C. Natural Language Processing

Sachdeva in [24] aims to propose an approach that effectively balances both business value and process flow in the prioritization process. This approach helps the Product Owner in deciding which requirements to prioritize and how to organize them in the Product Backlog. This approach models requirements iteratively through user stories to align with the rapidly changing business needs and with the continuous change of requirements. The Product Owner (PO) typically prioritizes based on Business Value without considering dependencies of user stories. They used UML Activity Diagrams to visualize process flows, then they converted user requirements into these diagrams using NLP. This visualization helps prioritize stories based on their process flow.

Shafiq in [25] introduces the NLP4IP approach, a recommendation system to prioritize issues such as stories, bugs, or tasks. It is a semi-automatic approach that uses Natural Language Processing (NLP) to address prioritization challenges to create a recommendation model. The rank of newly added or modified issues is dynamically predicted by this model. The JIRA issue tracking software was employed to evaluate the approach across 19 projects from 6 repositories, resulting in a total of 29,698 issues. Furthermore, they implemented a JIRA plug-in that illustrates the predictions generated by the new Machine Learning model.

Sami in [26] introduced a tool that integrates OpenAI, Flask, and React to automatically generate and rank user stories based on core requirements to improve the project management workflows. The tool allows users to input the requirements then it's used to generate and prioritize user stories and epics. The tool converts responses into a user-friendly JSON format and supports CSV downloads then the tool integrates with task management platforms such as JIRA and Trello.

Sharma in [27], a proposed approach that employs Natural Language Processing (NLP) algorithms to categorize user experiences that are similar and organize them into project releases. The goal of this approach is to enhance the release planning process for intricate and substantial software initiatives that can be simplified. The method entailed the initial development of a word corpus for each project release, followed by the conversion of user stories into vector representations using Java utilities. Lastly, the RV coefficient NLP algorithm is employed to organize them into distinct software releases. Furthermore, these algorithms can be integrated into commercial tools such as JIRA and Rally to facilitate enhanced release planning in Agile environments.

Kifetew in [28] proposes ReFeed approach. It aims to provide a more accurate and automated way to prioritize requirements based on user feedback. The approach employs

Natural Language Processing (NLP) to prioritize requirements. The approach extracts and propagates quantifiable properties from related user feedback. They used domain knowledge to bridge the vocabulary gap between users and developers. This approach bridges the gap between end-users' words and developers' words to formulate requirements.

D. Optimization

Asghar in [15] presents a methodology that integrates traditional prioritization factors with contemporary metrics and techniques, such as AHP and MOSCOW, as well as ISO standards. The objective of this method is to enhance the quality of both the process and the product by offering a more comprehensive and consistent framework for prioritizing requirements. Additionally, the methodology aims to improve the quality of requirements that are selected and prioritized during SCRUM sessions. The proposed model was validated through the use of comprehensive simulations with the iThink software.

Kumar in [16] identifies eight critical attributes—such as leadership support, human resources, and information technology—through literature review and expert validation. To rank these attributes effectively, the study employs an integrated technique called the Analytic Hierarchy Process (AHP) Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). This combined approach allows for a structured evaluation of the attributes based on their relative importance and their distance from an ideal solution.

Muhammad in [17] applied Multi-Criteria Decision Making (MCDM) techniques to rank non-functional requirements (NFRs). The study integrates Fuzzy Logic to manage the imprecision in NFR evaluations and employs pairwise comparison to establish the relative importance of various NFRs. The effectiveness of the proposed model was validated through a case study involving Agile projects, which demonstrated its capability to improve the prioritization process by accurately addressing and ranking NFRs.

Agrawal in [18] proposes an Agile-based Risk Rank (AR-Rank) method to prioritize risk factors in Agile methodology. This method is applied by using Particle Swarm Optimization (PSO) for iterative Optimization. The method provides precedence ranking of risks to minimize their impact and ensure timely delivery of risk-free software.

The AR-Rank approach is validated against other methods and tested on ten real-life projects.

Prakash in [19] proposed ARP–GWO method, this method uses grey wolf Optimization and the k-means clustering algorithm. They aim to prioritize risks in Agile Software Development and to enhance quality, reduce costs, and improve delivery times. The Experimental demonstrates five industrial projects that demonstrate ARP–GWO's effectiveness and high satisfaction among developers and users.

Chaves in [20] presents a Mult-objective swarm intelligence presents a multi-objective swarm intelligence metaheuristic (MOABC) to solve The Next Release Problem (NRP). They present superior performance compared to other methods on real-world datasets. They are seeking to improve the Optimization process by incorporating hybrid approaches, larger datasets, and additional constraints.

Brezočnik in [21] aim to enhance iteration planning in Agile Software Development. By introducing STAPSO, a novel algorithm. This algorithm combines Scrum task allocation and Particle Swarm Optimization. STAPSO is tested on a real-world dataset. It showed promising results in task allocation and applies to various estimation techniques.

Brezočnik in [22] provides an overview of swarm intelligence algorithms. The paper systematically classifies swarm intelligence algorithms based on Agile Software Development tasks like the next release problem, risk, and cost estimation, and discusses their promising results.

Sagrado in [23] address multi-objective Optimization problems in the Next Release Problem (NRP). They proposed approach includes the Ant Colony System (ACS) to multi-objective Optimization, where ants build solutions probabilistically, considering pheromone levels and heuristic information. And Comparative Analysis compared with Greedy Randomized Adaptive Search Procedure (GRASP) and Non-dominated Sorting Genetic Algorithm (NSGA-II) to validate its effectiveness in generating high-quality solutions for requirement prioritization.

Somohano in [24] aims to reduce computational complexity and enhance the precision of requirements prioritization. They present an approach to enhancing the Analytic Hierarchy Process (AHP) by integrating evolutionary computing techniques. Additionally, the research suggests integration of multiple criteria and developing a software tool to assist project managers in the prioritization process more efficiently.

E. Other Approaches

AbdElazim in [9] introduces a comprehensive framework for prioritizing requirements in Agile Software Development, focusing on continuous and scalable prioritization. It effectively manages rapidly changing requirements and their dependencies by integrating early into the Agile process with epics and user stories. This fully integrated approach ensures that the framework can handle requirement changes at any stage of the development cycle, making the Agile development process more adaptable and responsive to evolving project needs but no tool was provided.

Govil in [39] presents a comparative analysis of various requirement prioritization techniques such as AHP, Pair wise comparison, MoSCoW, planning poker, ping pong balls, bubble sort, and others detailing their strengths and weaknesses. The paper highlights discrete parameters that influence the success of software projects and discusses the difficulty of prioritizing requirements in Agile methodologies, where changes can occur late in development. The goal is to provide product owners with insights to select the most suitable prioritization technique based on the project's specific needs and constraints.

Kamal in [40] explores the factors that contribute to the success of Agile Requirements Change Management (ARCM) in the context of Global Software Development (GSD). It follows a two-step approach: first, identifying success factors through a Systematic Mapping Study (SMS) and validating them with industry practitioners via a questionnaire survey. And the second step is to prioritize these factors using the Analytical Hierarchy Process (AHP). This study shows twenty-one critical success factors for ARCM in GSD, with top priorities being resource allocation at overseas sites, communication, coordination, control (3Cs), process improvement expertise, a geographically distributed change control board (CCB), and continuous top management support. The findings aim to assist practitioners in effectively implementing ARCM activities in GSD settings. The research further needs to develop a comprehensive readiness model for ARCM. This model should identify negative impact factors and collect best practices through multiple case studies.

Kifetew in [41] discusses the use of automated decision-making techniques to help engineers in the process of selecting and prioritizing requirements. Effective involvement of the development team and stakeholders is crucial in the decision-making process enabled by these tools. The paper used Analytic Hierarchy Process (AHP) and Genetic Algorithms to introduce a tool-supported to perform collaborative requirements prioritization process. The tool enables an iterative prioritization process, therefore enabling stakeholders to actively participate in decision-making throughout the process.

Perkusich in [42] presents an approach that leverages intelligent software engineering techniques to enhance requirement prioritization in Agile environments. The approach utilizes a combination of automated tools and Machine Learning algorithms to analyze user feedback, historical project data, and other relevant metrics to prioritize requirements dynamically. This method allows for continuous re-evaluation of priorities based on new data and evolving project needs.

AL-Ta’ani in [43] proposes a conceptual framework for continuous requirements prioritization in Agile development, addressing the challenge of selecting key user requirements for implementation across iterations. The framework is informed by a thorough review of related literature and content analysis of the data. It delineates the critical factors impacting the requirements prioritization process and their effect on the final product, categorizing them into three primary dimensions: environment, process, and product. The environment includes stakeholders’ characteristics, constraints relevant to the project, and Requirement Nature; the process outlines the particular processes involved in prioritization, and the Product describes the outcomes. The study highlights the importance of systematic prioritization to prevent costly development errors and potential project failures, suggesting areas for future research to refine prioritization methods and techniques.

Alkandari in [44] discusses three models for requirements prioritization in Agile development. Model 1 focuses on estimating business value using work breakdown structure and knowledge factors but lacks clarity on selection criteria for iteration. Model 2 consists of initial project backlog, prioritized project backlog, sprint backlog, and implemented requirements, emphasizing client-driven prioritization. Model 3 is an improved version of Model 2 based on case studies, incorporating business value, negative value, and risk for more accurate prioritization. The models were evaluated based on factors like cost, importance, risk, and dependencies to propose a comprehensive prioritization approach.

Saeed in [45] conducted a case study across five organizations. They used the grounded theory to assign numerical values to qualitative data, resulting in a more efficient and effective prioritization process. The study shows that organizations often deviate from traditional steps, using unique methods to handle requirements prioritization, focusing on Business Value, Risk Factors, and Priority Criteria. The proposed mode allows for enhancement by integrating other techniques. Agile development's evolving nature means prioritization methods will keep changing. To improve this process, the study suggests more case studies to propose better models.

Many researchers have noted gaps in the literature regarding AI-based techniques for requirements prioritization in Agile methodologies. The current literature identifies several key issues, including a lack of real-world empirical validation of these techniques. Scalability problems still exist, particularly when handling multiple requirements or large-scale projects. Furthermore, managing requirement dependencies is insufficient in current methodologies. Stakeholder collaboration is rarely fully integrated into prioritization models. This study addresses these gaps by evaluating the effectiveness of AI techniques in real-world Agile settings, focusing on scalability, dependency management, and stakeholder collaboration, while providing recommendations for future research and practical implementation.

V. Evaluation Criteria

Prioritization of requirements is the systematic procedure of arranging requirements according to specific inputs, processing techniques, and intended outputs. For this goal, the following primary categories of AI-based methodologies—Fuzzy Logic, Optimization algorithms, and Machine Learning—have been utilized. It is essential to have evaluation criteria that assess each type separately. This is necessary because these approaches differ significantly in their characteristics and operational methods. Therefore, in addition to general criteria for evaluating prioritization techniques, unique standards have been developed to evaluate Machine Learning approaches, Optimization algorithms, and Fuzzy Logic are developed. All the surveyed strategies are evaluated using ten metrics that are derived from Fuzzy Logic, Optimization, and Machine Learning, including Stakeholder Involvement, Dependency, Scalability, Validation, Impact on Collaboration, Accuracy, Flexibility, Complexity, Robustness to Uncertainty, Automation, Ease of Implementation. These criteria are evaluated using different values: ‘Yes’, ‘No’, and ‘Moderate’.

VI. Results Analysis and Comparison

A. Demographics of the SLR

Table I presents a comprehensive summary of 32 research studies categorised by year, with a detailed analysis of the publication types, specifically journals. The analysis of 32 research studies from 2011 to 2024 reveals that 21 were published in journals and 11 in conferences. The publishers include Springer with 9 papers, IEEE with 10, and Elsevier with 6. The studies cover a range of topics: 3 on Fuzzy Logic, 6 on

Machine Learning, 5 on NLP, 10 on Optimization techniques, and 8 on other approaches like frameworks and gamification. Springer leads in publishing both journals and conferences, while IEEE has a strong presence in both areas, especially in Machine Learning and Natural Language Processing. Elsevier's contributions are consistent across various topics, all in journals.

In Fig. 2, the frequency of different approaches that used in requirements prioritization in Agile methodology. In Fig. 3, The year chart provides an insightful look into the trend of publications focused on using AI to prioritize requirements in Agile methodology from 2011 to 2024 that's illustrated in Fig. 3.

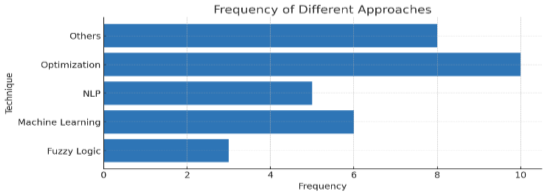


Fig. 2. Frequency of different approaches.

TABLE I. OVERVIEW OF SELECTED STUDIES (PUBLICATION TYPE JOURNAL)

No	Reference	Year	Technique Used	Related Category	Type	Publisher
15	Borhan, et al. [15]	2024	Fuzzy Logic Operations	Fuzzy Logic	Journal	JOAASR
16	Abusaeed,et al.[16]	2023	Fuzzy AHP (Analytic Hierarchy Process)	Fuzzy Logic	Journal	Elsevier
17	Rottoli, et al.[17]	2021	Fuzzy Linguistic Labels	Fuzzy Logic	Conference	CEUR-WS
18	Anand, et al. [18]	2017	Apriori Technique	Machine Learning	Journal	Elsevier
19	Hudaib, et al. [19]	2019	Self-Organizing Maps	Machine Learning	Journal	IEEE
20	Veerappa, et al. [20]	2011	Clustering	Machine Learning	Conference	Springer
21	Belsis, et al. [21]	2014	Unsupervised Requirements Clustering	Machine Learning	Journal	Springer
22	Achimugu, et al. [22]	2014	Clustering	Machine Learning	Conference	Springer
23	Kumar, et al. [23]	2022	K-Means Algorithm	Machine Learning	Conference	IEEE
24	Sachdeva, et al. [24]	2018	User Requirements Prioritization	Natural Language Processing (NLP)	Conference	IEEE
25	Shafiq,et al. [25]	2021	NLP-based Recommendation Approach	Natural Language Processing (NLP)	Journal	IEEE
26	Sami, et al. [26]	2024	Large Language Models	Natural Language Processing (NLP)	Conference	Springer
27	Sharma, et al. [27]	2019	NLP Algorithm	Natural Language Processing (NLP)	Journal	IEEE
28	Kifetew, et al. [28]	2021	User-Feedback Driven Prioritization	Natural Language Processing (NLP)	Journals	Elsevier
29	Asghar, et al. [29]	2016	Requirements Elicitation & Prioritization MoSCoW, Interviews, Workshops	Optimization	Journal	IJACSA
30	Kumar, et al. [30]	2020	AHP and TOPSIS	Optimization	Journal	Emerald
31	Muhammad, et al.[31]	2023	Multi-Criteria Decision Making Analysis	Optimization	Journal	IEEE
32	Agrawal, et al. [32]	2016	Risk Prioritization and Optimization	Optimization	Journal	IEEE
33	Prakash, et al. [33]	2021	Grey Wolf Optimizer (GWO)	Optimization	Journal	Springer
34	Chaves, et al. [34]	2015	Multiobjective Swarm Intelligence Evolutionary Algorithm	Optimization	Journal	Elsevier
35	Brezočnik, et al. [35]	2018	Particle Swarm Optimization	Optimization	Conference	Springer
36	Brezočnik, et al. [36]	2020	Swarm Intelligence Algorithms	Optimization	Conference	Springer
37	Del Sagrado, et al. [37]	2015	Ant Colony Optimization	Optimization	Conference	Springer
38	Somohano [38]	2021	Evolutionary Computing for AHP	Optimization	Conference	Springer
39	AbdElazim, et al. [9]	2020	Framework-Based Approach	Other Approaches	Journal	IOP
40	Govil, et al. [39]	2021	Information Extraction Techniques	Other Approaches	Journal	IEEE
41	Kamal, et al. [40]	2020	Prioritization Techniques	Other Approaches	Journal	IEEE
42	Kifetew, et al.[41]	2017	Gamification, Collaborative Techniques	Other Approaches	Conference	IEEE
43	Perkusich, et al.[42]	2020	Intelligent Software Engineering	Other Approaches	journal	Elsevier
44	AL-Ta'ani, et al. [43]	2013	Conceptual Framework	Other Approaches	Journal	Elsevier
45	Alkandari, et al. [44]	2017	Enhancement Techniques	Other Approaches	Journal	JSW
46	Saeed, et al. [45]	2023	Requirements Prioritization Techniques	Other Approaches	Journal	Technical Journal of UET

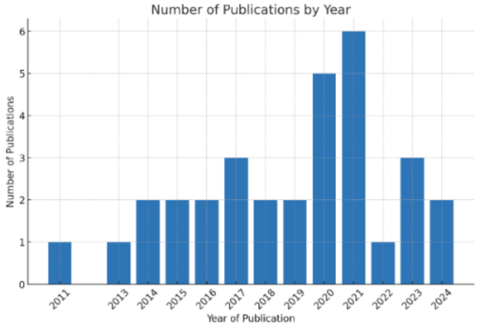


Fig. 3. Number of publications per year.

B. Strengths and Weaknesses of the Surveyed Techniques

The evaluation of the 32 papers reveals a diverse range of strengths, weaknesses, and limitations in applying various methodologies to Agile Software Development is presented in Table II. Many papers showcase innovative approaches, such as the combination of Fuzzy Logic with AHP for enhanced decision-making, the use of NLP for automating prioritization, and the application of advanced algorithms for task allocation and Optimization. These strengths are often balanced by significant weaknesses, including complexity in implementation, the need for expert knowledge, and high computational or data requirements. Moreover, limitations such as scalability, data dependency, and subjective judgment frequently emerge, indicating that while these methods provide valuable solutions, they may not be universally applicable without careful consideration of context and resources. Overall, the analysis underscores the need for a tailored approach when applying these techniques to Agile environments, ensuring that their strengths are fully leveraged while mitigating their inherent weaknesses and limitation.

TABLE II. STRENGTH, WEAKNESS, AND LIMITATIONS OF REQUIREMENTS PRIORITIZATIONS TECHNIQUES

Reference	Strengths	Weaknesses	Limitations
Borhan, et al. [15]	Improved decision making under uncertainty	Complexity in fuzzy rule definitions	Limited scalability
Abusaeed,et al.[16]	Combines advantages of AHP and fuzzy logic	Requires expert knowledge	Complexity in AHP hierarchy formulation
Rottoli, et al.[17]	Handles linguistic uncertainty effectively	Potential bias	Subjective interpretation of linguistic labels
Anand, et al. [18]	Resolves conflicts through data mining	Complexity in implementation	Data dependency
Hudaib, et al. [19]	Visualizes data patterns effectively	Requires expertise in SOM	Complexity
Veerappa, et al. [20]	Groups similar stakeholders effectively	May not capture all stakeholder nuances	Data dependency
Belsis, et al. [21]	Unsupervised approach reduces bias	Requires large dataset	Complexity
Achimugu, et al. [22]	Efficiently handles large scale data	May overlook individual nuances	Scalability
Kumar, et al. [23]	Simple and fast clustering method	Sensitive to initial conditions	Scalability
Sachdeva, et al. [24]	Focuses on user requirements	May overlook technical requirements	Subjective judgement
Shafiq,et al. [25]	Automates prioritization using NLP	Requires large dataset	Data dependency
Sami, et al. [26]	Leverages advanced language models	Computationally intensive	Data dependency
Sharma, et al. [27]	Automates release planning using NLP	Requires expertise in NLP	Data dependency
Kifetew, et al. [28]	Incorporates user feedback in prioritization	Potential bias in feedback	Data dependency
Asghar, et al. [29]	Enhances quality by thorough elicitation	Time-consuming	High dependency on stakeholder input
Kumar, et al. [30]	Effective prioritization by combining AHP and TOPSIS	Data-intensive	Complexity in combining methods
Muhammad, et al.[31]	Comprehensive evaluation of multiple criteria	Resource-intensive	High complexity
Agrawal, et al. [32]	Focuses on risk management and optimization	Complexity in risk assessment	High complexity
Prakash, et al. [33]	Efficient optimization algorithm	Requires parameter tuning	Complexity
Chaves, et al. [34]	Handles multiple objectives simultaneously	Computationally intensive	High complexity
Brezočnik, et al. [35]	Efficient task allocation algorithm	Requires parameter tuning	Complexity
Brezočnik, et al. [36]	Effective in solving complex problems	Requires expertise in swarm intelligence	Complexity
Del Sagrado, et al. [37]	Effective multi-objective optimization	Computationally intensive	High complexity
Somohano [38]	Enhances AHP with evolutionary computing	Complexity in implementation	Requires expert knowledge
AbdElazim, et al. [9]	Provides structured approach for prioritization	May not fit specific project needs	Subjective judgement
Govil, et al. [39]	Automates data extraction process	Requires comprehensive datasets	Data extraction accuracy
Kamal, et al. [40]	Focuses on global software development challenges	Difficult to generalize	High variability in global context
Kifetew, et al.[41]	Engages stakeholders through gamification	Potential for bias in game dynamics	Engagement dependency
Perkusich, et al.[42]	Leverages AI for software engineering	Requires extensive training data	Data dependency
AL-Ta'ani, et al. [43]	Provides a structured approach	May not fit specific project needs	Subjective judgement
Alkandari, et al. [44]	Improves existing processes	Requires adaptation to specific projects	Subjective judgement
Saeed, et al. [45]	Enhances quality of agile practices	Requires adaptation to specific contexts	Subjective judgement

C. Comparison of AI-based Requirements Prioritization Techniques

A comprehensive study is presented among Fuzzy Logic based, Machine Learning based, Optimization and Natural Language Processing-based techniques. The effectiveness of AI-driven prioritization in Agile methodologies is influenced by key parameters such as scalability, dependency management, stakeholder collaboration, automation, and accuracy. AI techniques like machine learning and optimization improve scalability by handling large datasets and managing dependencies effectively. NLP enhances stakeholder collaboration by automating feedback interpretation, while automation in AI reduces manual effort and allows dynamic prioritization. Fuzzy logic supports decision-making under uncertainty but may struggle with scalability. Overall, these parameters ensure that AI techniques provide adaptable, efficient, and accurate solutions to the challenges of requirements prioritization in Agile development. This comprehensive is defined based on evaluation criteria with answer of ‘Y’ for Yes, ‘N’ for No, and ‘M’ for ‘Moderate’.

Regarding the Fuzzy Logic-based technique, evaluating the implementation of Fuzzy Logic in Agile Software Development requires examining both its theoretical advantages and practical challenges. Fuzzy Logic offers a flexible and nuanced approach to handling uncertainty and imprecision in decision-making, particularly in complex environments like Agile projects where requirements are often ambiguous and evolving. The evaluation, as shown in Table III, reveals that while these Fuzzy Logic-based approaches are strong in handling uncertainty and, in some cases, fostering stakeholder collaboration, they generally struggle with scalability and managing dependencies. The high complexity and limited usability in some methods could restrict their practical implementation in real-world Agile projects. This highlights a need for more balanced solutions that are both theoretically robust and practically applicable.

TABLE III. FUZZY LOGIC TECHNIQUE APPLICATIONS

Reference	Stakeholder Collaboration	Dependency	Scalability	Complexity and Usability	Flexibility	Robustness to Uncertainty
Borhan, et al. [15]	Y	N	N	N	Y	Y
Abusaeed,et al.[16]	N	N	N	N	N	Y
Rottoli, et al.[17]	Y	N	N	Y	Y	Y

Regarding Machine Learning technique, the evaluated research offers innovative clustering and prioritization techniques for Agile Software Development most to techniques aims to categorize related requirements into clusters. The cluster are pushed to be the next sprints based on their ranking. Some evaluation criteria as scalability, accuracy, and stakeholder management shows good indicator as it’s shown in Table IV. However, they generally face challenges in practical implementation due to complexity, lack of empirical validation, that would show limited effectiveness in handling stakeholder conflicts. While these approaches are valuable, their applicability may be constrained by the need for specialized knowledge and tools.

TABLE IV. MACHINE LEARNING TECHNIQUE APPLICATIONS

Reference	Stakeholder Collaboration	Dependency	Scalability	Validation	Impact on Collaboration	Accuracy	Flexibility
Anand, et al. [18]	Y	N	N	N	Y	Y	N
Hudaib, et al. [19]	Y	Y	Y	N	N	Y	Y
Veerappa, et al. [20]	Y	Y	N	N	Y	Y	Y
Belsis, et al. [21]	Y	N	Y	N	Y	Y	Y
Achimugu, et al. [22]	Y	N	Y	N	Y	Y	Y
Kumar, et al. [23]	N	N	Y	Y	Y	Y	Y

Regarding Natural Language Processing technique, most of the paper used NLP techniques reduce direct stakeholder involvement, which is crucial for aligning requirements with business needs. In Table IV, we can realize many of the papers moderate to high scalability, ease of implementation, suggesting a potential gap in practical applicability. The validation factor generally moderate to high. Most of the papers didn't consider dependency directly (Table V).

TABLE V. NATURAL LANGUAGE PROCESSING TECHNIQUE APPLICATION

Reference	Stakeholder Collaboration	Dependency	Scalability	Automation	Validation
Sachdeva, et al. [24]	M	M	Y	Y	M
Shafiq,et al. [25]	Y	Y	M	M	Y
Sami, et al. [26]	M	M	Y	Y	Y
Sharma, et al. [27]	Y	M	Y	Y	Y
Kifetew, et al. [28]	M	M	Y	Y	M

Regarding Optimization technique, the evaluation of these papers indicates a week-long focus on addressing requirements dependencies and facilitating scalability across most approaches. As shown in Table VI, traditional decision-making techniques, such as AHP and TOPSIS, are well-integrated with collaboration with stakeholders. And less collaboration with swarm intelligence or evolutionary algorithms. Continuous improvement is a common feature across all methods, reflecting the iterative nature of Agile methodologies. Overall, while most approaches are not considered dependencies directly and scalability, there is variability in how well they facilitate stakeholder collaboration, which underscores the potential of improvement in algorithm-based methods.

Regarding other frameworks and techniques, as it’s shown in Table VII, the collection of papers on requirements prioritization in Agile Software Development offers a range of theoretical frameworks and innovative tools, most approaches address well result in collaboration, dependencies, and continuous improvement effectively. But there is challenges in such terms as ease of implementation and scalability for certain innovative methods like DMGame. This suggests that while these approaches offer significant potential, they may require additional refinement or adaptation to be fully effective across different Agile contexts

TABLE VI. OPTIMIZATION TECHNIQUE APPLICATION

Reference	Stakeholder Collaboration	Dependency	Scalability	Ease of Implementation	Validation
Asghar, et al. [29]	N	N	N	Y	Y
Kumar, et al. [30]	Y	N	N	Y	Y
Muhammad, et al.[31]	Y	Y	Y	Y	Y
Agrawal, et al. [32]	Y	N	N	Y	Y
Prakash, et al. [33]	N	N	N	Y	Y
Chaves, et al. [34]	N	N	N	N	Y
Brezočnik, et al. [35]	N	Y	N	Y	Y
Brezočnik, et al. [36]	N	N	N	Y	Y
Del Sagrado, et al. [37]	N	N	N	Y	Y
Somohano, et al, [38]	Y	Y	Y	Y	Y

TABLE VII. OTHER TECHNIQUES APPLICATION

Reference	Stakeholder Collaboration	Dependency	Scalability	Ease of Implementation	Validation	Flexibility in Handling Change
AbdElazim, et al. [9]	Y	Y	Y	Y	Y	Y
Govil, et al. [39]	Y	Y	Y	Y	Y	Y
Kamal, et al. [40]	Y	Y	Y	Y	Y	Y
Kifetew, et al.[41]	Y	N	N	N	Y	Y
Perkusich, et al.[42]	N	Y	Y	N	Y	Y
AL-Ta'ani, et al. [43]	Y	Y	Y	Y	Y	Y
Alkandari, et al. [44]	Y	Y	Y	Y	Y	Y
Saeed, et al. [45]	Y	Y	Y	Y	Y	Y

Results show that Machine Learning techniques, such as clustering algorithms, effectively address the scalability challenges in requirement prioritization for large Agile projects by grouping similar requirements, reducing manual effort, and improving accuracy. Fuzzy Logic also enhances stakeholder collaboration, resolving conflicts and improving decision-making. These findings suggest that AI techniques not only offer technical advantages but also enhance team communication and responsiveness in Agile workflows, though real-world validation is still needed.

VII. LIMITATIONS

Despite the promising results of AI-based techniques for requirements prioritization, this study has several limitations. A key limitation is the lack of empirical validation, which affects the practical applicability of many AI methods in Agile environments. While these techniques show potential in addressing scalability and improving stakeholder collaboration, their complexity and the need for specialized knowledge pose challenges for widespread adoption and for handling changes in requirements effectively. Additionally, dependency management is not adequately addressed, leaving a crucial aspect of Agile prioritization insufficiently explored.

VIII. CONCLUSION

This paper has conducted a comprehensive systematic literature review of AI-driven techniques for requirements prioritization within Agile methodologies. This paper addresses a significant gap by analysing 32 key studies spanning 2010 to 2024. The SLR finds the strengths and weaknesses of Fuzzy Logic, Machine Learning, Optimization, and Natural Language Processing (NLP) techniques. Our findings reveal that each method has its distinct limitations. These limitations are particularly in terms of scalability, accuracy, and simplicity of implementation. These AI-based approaches offer promising solutions to the challenges of requirements prioritization in Agile environments.

In Future work, the analysis underscores the necessity of developing hybrid AI-based techniques that integrate the strengths of Fuzzy Logic, Machine Learning, Natural Language Processing, and Optimization to create more scalable, and efficient prioritization methods. The integrated approach could better handle the complexities of modern software development in Agile methodology. That approach should handle the rapidly changing of the Agile environments where continuous prioritization and stakeholder collaboration are critical. The approach should focus on empirical validation of AI-based requirements prioritization techniques through case studies and real-world applications to ensure their practical relevance. Additionally, there is a need to align this approach to be more closely with stakeholder expectations and to meet the business need, and to ensure that the prioritization process in Agile technically and contextually appropriate. Such efforts should prioritize creating scalable, efficient, and adaptive methods that consider dependencies in requirements and align closely with stakeholder needs and Agile practices and to align with the continuous improvement. By addressing these challenges, it will be possible to develop more effective and adaptable requirements prioritization techniques that can be broadly implemented in diverse, real-world software development environments.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisors, for their invaluable guidance. I also extend my thanks to the Software Engineering Department whose unwavering support during my qualification year played a pivotal role in the successful completion of this research.

REFERENCES

[1] T. Ambreen, N. Ikram, M. Usman, and M. Niazi, "Empirical research in requirements engineering: trends and opportunities," Requir Eng, vol. 23, pp. 63–95, 2018.

[2] W. Jiang, H. Ruan, L. Zhang, P. Lew, and J. Jiang, "For user-driven software evolution: Requirements elicitation derived from mining online reviews," in Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II 18, Springer, 2014, pp. 584–595.

[3] P. A. Laplante and M. Kassab, Requirements engineering for software and systems. Auerbach Publications, 2022.

[4] A. Radwan, A. Abdo, and S. A. Gaber, "An Approach for Requirements Engineering Analysis using Conceptual Mapping in Healthcare Domain," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, 2021.

[5] A. Gupta, G. Poels, and P. Bera, “Using conceptual models in agile software development: a possible solution to requirements engineering challenges in agile projects,” IEEE Access, vol. 10, pp. 119745–119766, 2022.

[6] N. H. Borhan, H. Zulzalil, and N. M. A. Sa’adah Hassan, “Requirements prioritization techniques focusing on agile software development: a systematic literature,” International Journal of Scientific and Technology Research, vol. 8, no. 11, pp. 2118–2125, 2019.

[7] W. R. Fitriani, P. Rahayu, and D. I. Sensuse, “Challenges in agile software development: A systematic literature review,” in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2016, pp. 155–164.

[8] Alawneh, L, "Requirements prioritization using hierarchical dependencies." Information technology-new generations: 14th International Conference on Information Technology. Springer International Publishing, 2018.

[9] K. AbdElazim, R. Moawad, and E. Elfakharany, “A framework for requirements prioritization process in agile software development,” in Journal of Physics: Conference Series, IOP Publishing, 2020, p. 012001.

[10] J. M. Mendel, “Fuzzy logic systems for engineering: a tutorial,” Proceedings of the IEEE, vol. 83, no. 3, pp. 345–377, 1995.

[11] R. Anwar and M. B. Bashir, “A Systematic Literature Review of AI-based Software Requirements Prioritization Technique,” IEEE Access, 2023.

[12] M. D. Nagpal, K. Malik, and A. Kalia, “A comprehensive analysis of requirement engineering utilizing machine learning techniques,” Design Engineering, pp. 2662–2678, 2021.

[13] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, “A survey on the state-of-the-art machine learning models in the context of NLP,” Kuwait journal of Science, vol. 43, no. 4, 2016.

[14] K.-L. Du and M. N. S. Swamy, Search and optimization by metaheuristics, vol. 1. Springer, 2016.

[15] N. H. Borhan, H. Zulzalil, N. M. Ali, A. B. M. Sultan, and R. Bahsoon, “Stakeholder Analysis using Fuzzy Logic Operations for Integrated User Story Prioritisation Approach in Agile-Scrum Method,” Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 47, no. 2, pp. 76–93, 2024.

[16] S. Abusaeed, S. U. R. Khan, and A. Mashkoor, “A Fuzzy AHP-based approach for prioritization of cost overhead factors in agile software development,” Appl Soft Comput, vol. 133, p. 109977, 2023.

[17] G. D. Rottoli and C. Casanova, “Multi-criteria group requirement prioritization in software engineering using fuzzy linguistic labels.,” in ICAI Workshops, 2021, pp. 16–28.

[18] R. V. Anand and M. Dinakaran, “Handling stakeholder conflict by agile requirement prioritization using Apriori technique,” Computers & Electrical Engineering, vol. 61, pp. 126–136, 2017.

[19] A. Hudaib and F. Alhaj, “Self-Organizing Maps for Agile Requirements Prioritization,” in 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE, 2019, pp. 1–5.

[20] V. Veerappa and E. Letier, “Clustering stakeholders for requirements decision making,” in Requirements Engineering: Foundation for Software Quality: 17th International Working Conference, REFSQ 2011, Essen, Germany, March 28-30, 2011. Proceedings 17, Springer, 2011, pp. 202–208.

[21] P. Belsis, A. Koutoumanos, and C. Sgouropoulou, “PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development,” Requir Eng, vol. 19, pp. 213–225, 2014.

[22] P. Achimugu, A. Selamat, and R. Ibrahim, “A Clustering Based Technique for Large Scale Prioritization during Requirements Elicitation. 2014,” Cham: Springer International Publishing.

[23] B. Kumar, U. K. Tiwari, D. C. Dobhal, and H. S. Negi, “User Story Clustering using K-Means Algorithm in Agile Requirement Engineering,” in 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), IEEE, 2022, pp. 1–5.

[24] S. Sachdeva, A. Arya, P. Paygude, S. Chaudhary, and S. Idate, “Prioritizing user requirements for agile software development,” in 2018 International Conference On Advances in Communication and Computing Technology (ICACCT), IEEE, 2018, pp. 495–498.

[25] S. Shafiq, A. Mashkoor, C. Mayr-Dorn, and A. Egyed, “NLP4IP: Natural language processing-based recommendation approach for issues prioritization,” in 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2021, pp. 99–108.

[26] M. A. Sami, Z. Rasheed, M. Waseem, Z. Zhang, T. Herda, and P. Abrahamsson, “Prioritizing Software Requirements Using Large Language Models,” arXiv preprint arXiv:2405.01564, 2024.

[27] S. Sharma and D. Kumar, “Agile release planning using natural language processing algorithm,” in 2019 Amity International Conference on Artificial Intelligence (AICAI), IEEE, 2019, pp. 934–938.

[28] F. M. Kifetew, A. Perini, A. Susi, A. Siena, D. Muñante, and I. Morales-Ramirez, “Automating user-feedback driven requirements prioritization,” Inf Softw Technol, vol. 138, p. 106635, 2021.

[29] A. R. Asghar, S. N. Bhatti, A. Tabassum, Z. Sultan, and R. Abbas, “Role of requirements elicitation & prioritization to optimize quality in scrum agile development,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 12, 2016.

[30] R. Kumar, K. Singh, and S. K. Jain, “A combined AHP and TOPSIS approach for prioritizing the attributes for successful implementation of agile manufacturing,” International Journal of Productivity and Performance Management, vol. 69, no. 7, pp. 1395–1417, 2020.

[31] A. Muhammad, A. Siddique, M. Mubasher, A. Aldweesh, and Q. N. Naveed, “Prioritizing non-functional requirements in agile process using multi criteria decision making analysis,” IEEE Access, vol. 11, pp. 24631–24654, 2023.

[32] R. Agrawal, D. Singh, and A. Sharma, “Prioritizing and optimizing risk factors in agile software development,” in 2016 ninth international conference on contemporary computing (IC3), IEEE, 2016, pp. 1–7.

[33] B. Prakash and V. Viswanathan, “ARP–GWO: an efficient approach for prioritization of risks in agile software development,” Soft comput, vol. 25, no. 7, pp. 5587–5605, 2021.

[34] J. M. Chaves-Gonzalez, M. A. Perez-Toledano, and A. Navasa, “Software requirement optimization using a multiobjective swarm intelligence evolutionary algorithm,” Knowl Based Syst, vol. 83, pp. 105–115, 2015.

[35] L. Brezočnik, I. Fister Jr, and V. Podgorelec, “Scrum task allocation based on particle swarm optimization,” in International Conference on Bioinspired Methods and Their Applications, Springer, 2018, pp. 38–49.

[36] L. Brezočnik, I. Fister, and V. Podgorelec, “Solving agile software development problems with swarm intelligence algorithms,” in New Technologies, Development and Application II 5, Springer, 2020, pp. 298–309.

[37] J. Del Sagrado, I. M. Del Águila, and F. J. Orellana, “Multi-objective ant colony optimization for requirements selection,” Empir Softw Eng, vol. 20, pp. 577–610, 2015.

[38] J. C. B. Somohano-Murrieta, J. O. Ocharán-Hernández, Á. J. Sánchez-García, X. Limón, and M. de los Ángeles Arenas-Valdés, “Improving the analytic hierarchy process for requirements prioritization using evolutionary computing,” Programming and Computer Software, vol. 47, pp. 746–756, 2021.

[39] N. Govil and A. Sharma, “Information extraction on requirement prioritization approaches in agile software development processes,” in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2021, pp. 1097–1100.

[40] T. Kamal, Q. Zhang, M. A. Akbar, M. Shafiq, A. Gumaei, and A. Alsanad, “Identification and prioritization of agile requirements change management success factors in the domain of global software development,” IEEE Access, vol. 8, pp. 44714–44726, 2020.

[41] F. Kifetew, D. Munante, A. Perini, A. Susi, A. Siena, and P. Busetta, “DMGame: a gamified collaborative requirements prioritisation tool,” in 2017 IEEE 25th International Requirements Engineering Conference (RE), IEEE, 2017, pp. 468–469.

[42] M. Perkusich et al., “Intelligent software engineering in the context of agile software development: A systematic literature review,” Inf Softw Technol, vol. 119, p. 106241, 2020.

[43] R. H. AL-Ta’ani and R. Razali, “Prioritizing requirements in agile development: A conceptual framework,” Procedia Technology, vol. 11, pp. 733–739, 2013.

[44] M. A. Alkandari and A. Al-Shammeri, “Enhancing the Process of Requirements Prioritization in Agile Software Development-A Proposed Model.,” J. Softw., vol. 12, no. 6, pp. 439–453, 2017.

[45] N. Saeed, E. Yasmin, A. R. Riaz, N. Khalid, Y. Hafeez, and I. Rubab, “Enabling the Requirements Prioritization Techniques to Improve the Quality of Agile Practices.,” Technical Journal of University of Engineering & Technology Taxila, vol. 28, no. 4, 2023.

Intelligent Detection and Search Model for Communication Signals Based on Deep-Re-Hash Retrieval Technology

Hui Liu[1]*, Xupeng Liu[2]

Department of Electronic Engineering, Zhengzhou Railway Vocational and Technical College, Zhengzhou 451460 China[1]
Tianshi Academy, Zhengzhou Railway Vocational and Technical College, Zhengzhou 451460, China[2]

Abstract—**With the explosive growth of image data, traditional image retrieval methods face challenges of low efficiency and insufficient accuracy. In view of this, the study first analyzed the traditional Deep-Re-Hash detection technology and constructed a general hash detection model. Secondly, Cauchy functions and Hadamard matrices were introduced to optimize the generation of hash centers, and an improved Deep-Re-Hash detection model was proposed. The experimental results showed that the highest precision of the improved Deep-Re-Hash was 97%, and the highest MAP value was 90%. In simulation testing, the lowest detection similarity of the improved Deep-Re-Hash detection model was 64.8%, and the detection speed at this time was 7.6s. The hash codes generated by this model were highly aggregated, with very clear edges. In the indicator rating, the highest storage occupancy rating was close to 45 points, the highest detection satisfaction rating was close to 50 points, and the highest detection time rating was close to 30 points. Based on the above data, the proposed improved Deep-Re-Hash detection model shows great potential in processing large-scale image data. It successfully improves the intelligent detection and search efficiency of communication image signals, providing useful reference and inspiration for researchers in related fields.**

***Keywords**—Deep-Re-Hash retrieval; communication signals; image data; cauchy function; hadamard matrix*

I. Introduction

In modern communication systems, the transmission and storage of image signals is an important field. With the explosive growth of image data, it is important to efficiently detect and search for image signals [1]. Traditional image retrieval methods face challenges of low efficiency and insufficient accuracy. To address these challenges, deep learning technology has emerged in recent years. Its outstanding feature learning ability and highly non-linear processing ability have achieved significant success in image signal processing [2]. However, when faced with large-scale image signal datasets, traditional methods still face enormous pressure on storage and computing resources. To overcome these challenges, Deep-Re-Hash (DRH) retrieval technology has emerged [3]. DRH maps images into compact binary encoding, enabling efficient image retrieval and search. This method has broad application potential in Communication Image Signal (CIS) processing. By learning the hash code of images, it is possible to compare images in low-dimensional binary space and perform high-speed similarity search [4]. To further optimize the practical application performance of this technology, the algorithm structure was innovatively decomposed. Cauchy functions and Hadamard matrices were introduced for hash center generation optimization, aiming to achieve more efficient hash retrieval and search capabilities. The subject of the research is to optimize the intelligent detection and search efficiency of communication image signals using improved DRH search techniques and to solve the efficiency and accuracy problems of existing methods in large-scale data processing. The research is motivated by the rapidly growing demand for image signal processing in modern communication systems, especially in the task of intelligent detection and search of large-scale, multi-labelled communication image signals, where traditional methods are difficult to meet the requirements of real-time and accuracy. The contribution of the research is to propose an improved DRH that combines the Cauchy function and Hadamard matrix, which effectively enhances the detection and retrieval efficiency of communication image signals. The method is capable of generating more compact and discriminative hash codes, which are suitable for large-scale image data processing. This research is divided into six sections, Section II systematically analyzes and summarizes the existing related work. Section III proposes a communication image signal detection and search model based on a deep hash retrieval technique constructed by the research. Section IV demonstrates the performance of the improved model through experimental tests. Discussion is given in Section V. Section VI summarizes the research results and proposes directions for future research.

II. Related Work

With the developing communication technology, the transmission and processing of communication signals become increasingly important in various fields. CIS covers a wide range from traditional image communication to modern multimedia communication, characterized by large data volume and high diversity. Therefore, intelligent detection and search of CIS becomes a highly focused research direction. To improve the image communication performance of existing optical cameras, Nguyen D T et al. proposed a novel method for extracting image signal features of interest by combining low-density parity check codes. This method had more advantages in teaching traditional image communication processing techniques [5]. To improve the signal detection accuracy of the fifth generation wireless communication system, You Y H et al. proposed a new detection method by combining complex secondary synchronization signals. This method was highly effective and greatly reduced complexity compared to similar methods [6]. Li K et al. found that traditional image signal quality detection methods were no longer able to meet the requirements of existing medical information detection. Therefore, this team introduced the concepts of fully known signals and statically known signals using deep neural networks for optimization, and finally proposed a new detection method. This new method ensured the preservation of task-related information while denoising, which had strong usability [7]. Wen X et al. believed that autoencoders using convolutional neural networks still had insufficient capabilities in extracting spatiotemporal information from the network. Given this, the team proposed an image signal detection model using a pseudo 3D encoder and multi-level storage mechanism. The detection performance of this model was superior to other similar methods on three datasets [8]. Orsuti D et al. In order to improve the retrieval of communication image signals, the researchers finally proposed a communication image signal recovery method by combining full convolutional neural networks for non-iterative and robust phase recovery. The experimental results show that the performance of this method is compared with that of a conventional KK receiver using 4x digital upsampling, which can achieve 7% hard verdict forward error correction and 1.8 dB increase in receiver sensitivity [9].Kumar S et al. found that content-based image signal retrieval is susceptible to degradation of retrieval accuracy with the adoption of cloud storage. For this reason, the researchers proposed a new retrieval model after incorporating a nonlinear stacking algorithm. Experimental results show that the accuracy of this new model is significantly improved and more robust compared to the existing methods [10].

Hash retrieval technology is a technique used for quickly searching and retrieving large-scale datasets. It converts data into a set of fixed length hash values and uses these hash values to organize and index the data. Qin et al. found that most existing hash techniques only focused on the semantic similarity between image pairs, ignoring the ranking information of retrieval results. Therefore, this team proposed a new type of DRH by combining the deep top similarity theory. This method was more effective than traditional methods on several multi-label datasets [11]. To expand the application of hashing technology in image representation and approximate queries, Shuai C et al. proposed a principal component analysis hashing technique by combining longer binary codes and mapping vector rotations. This method achieved more types of partitions and significantly improved performance [12]. To further improve the speed and accuracy of hash technology for cross-modal information retrieval, Shen X et al. proposed a clustering driven deep adversarial hash model by combining end-to-end thinking. This model accurately maximized the distance calculation between images and texts with different labels, which had better performance [13]. Zhang D et al. believed that when the hash length changed, the hash model needed to be retrained. This consumed additional computing power and reduced its application scalability. Therefore, this team proposed a multi-hash model for cross-media retrieval. This model did not require training when the hash length changed, which was superior to some competitive shallow and deep hash methods [14]. Zhu L et al. In order to solve the problem of high storage cost and slow retrieval efficiency during multimedia retrieval, the researchers proposed a retrieval enhancement method after combining hash retrieval techniques. The experimental results show that under the reinforcing effect of the method makes the retrieval of Hamming distance calculation significantly accelerated, and the cost of retrieval time is effectively reduced [15].Sabahi F et al. found that the existing image retrieval reordering technique has the problem of high computational complexity, for this reason, the researchers proposed a novel and computationally efficient image retrieval reordering method after combining with hash retrieval technology. Experimental results show that this new method generally outperforms other traditional methods in terms of computational efficiency and retrieval performance [16].

In summary, many researchers have explored the detection of communication image information and proposed some remarkable research results. Meanwhile, the continuous development and evolution of hash retrieval technology makes its application wider. Therefore, the study attempts to combine these two and innovatively optimize hash technology to improve its performance in various modules and promote the development of CIS detection.

III. Construction of an Intelligent Detection and Search Model for Communication Signals based on Deep-Re-Hash Retrieval Technology

This study first analyzes traditional DRH and constructs an image signal detection model under this algorithm. Secondly, the shortcomings of DRH are elaborated, and the Cauchy concept function and Hamada matrix are introduced for hash center generation optimization. Finally, a new detection and search model is proposed.

A. *Communication Image Signal Processing Model Based on Deep-Re-Hash*

DRH can map high-dimensional data to low-dimensional binary encoding [17]. By learning the deep features of data, DRH can convert the original image signal into binary encoding, thereby achieving efficient processing and transmission of image signals. To generate binary hash codes, DRH typically adds n+1 fully connected layers for optimization [18]. This layer guides the generation of the required number of neurons and the length of the hash code. Fig. 1 shows the structure of DRH.

In Fig. 1, after inputting the initial image signal data, DRH uses a deep neural network to extract image features, and then converts these features into binary hash codes in Hamming space through a hash function. This process typically involves converting data, such as images, text, etc., into a compact binary representation, with key steps including data representation, feature extraction functions, binarization process, and hash code generation. The feature extraction function is represented by Eq. (1).

$$Z = f(X, \theta) \tag{1}$$

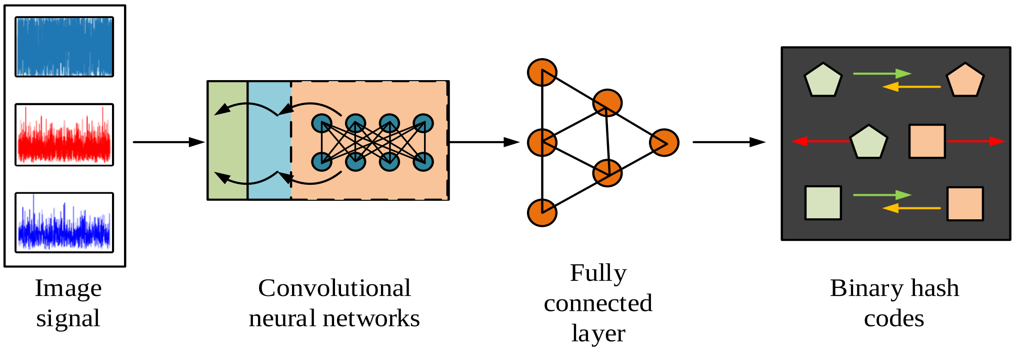


Fig. 1. Structure diagram of DRH.

In Eq. (1), Z represents the feature set of the input data, which is a real number vector. θ is a model parameter. X refers to the initial image input. $f(X,\theta)$ represents a deep neural network function. The binarization process is represented by Eq. (2).

$$H(X)=sign(Z), si\ gn(z_i)=\begin{cases}1, if z_i \ge 0\\ 0, otherwise\end{cases} \tag{2}$$

In Eq. (2), z_i represents the i th element in the feature set. If z_i is greater than or equal to 0, the corresponding hash code point is set to 1. Otherwise, it is set to 0. $sign$ represents a binarization function. The final hash code is a vector composed of 0 and 1, representing the hash code of the input data X. The hash function, as a key point that can directly affect the stability of the conversion result, should be consistent in the same class of image signals, but inconsistent in different classes of image signals. The hash function is represented by Eq. (3).

$$Hash(X)=sign(f(X,\theta)) \tag{3}$$

In Eq. (3), $sign$ maps each element in the feature vector to {− 1,1}, thereby generating a hash code. In similar images, DRH usually has two retrieval strategies, namely Hamming sorting and Hamming retrieval. Fig. 2 is a schematic diagram of two retrieval strategies.

In Fig. 2, both Hamming sorting and Hamming retrieval involve using Hamming distance to process data. Hamming sorting sorts data items based on their Hamming distance from the query item, while Hamming retrieval searches for data items that have a specified Hamming distance from the query item [19]. The Hamming distance is used to measure the difference between two equally long strings, that is, the number of different characters at the same position. In the context of hash retrieval, the Hamming distance is represented by Eq. (4).

$$D_H(HM(X),HM(Y))=\sum_{i=1}^{L}[HM(X)_i \neq HM(Y)_i] \tag{4}$$

In Eq. (4), L represents the length of the hash code. $HM(X)$ refers to inputting the hash code of X. $HM(Y)$ refers to inputting the hash code of $HM(Y)$. $HM(X)_i \neq HM(Y)_i$ represents an indicator function, which takes a value of 1 when these two are not equal, and 0 otherwise. On this basis, Hamming sorting is to calculate the Hamming distance between a given query hash code $HM(Q)$ and each hash code $HM(X)$ in the database, and then sort it according to this distance. The Hamming search is represented by Eq. (5).

$$Search(HM(Q),\tau)=\{X\,|D_H(HM(Q),HM(X)\le\tau)\} \tag{5}$$

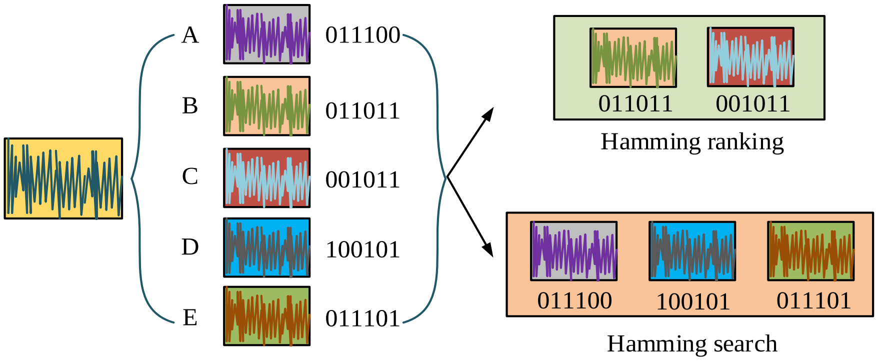


Fig. 2. Schematic diagram of hamming sorting and hamming retrieval.

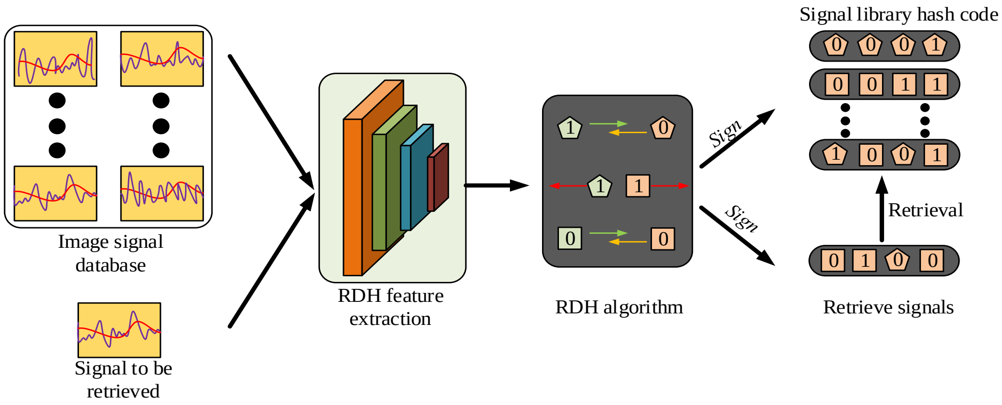


Fig. 3. DRH image signal detection model structure.

In Eq. (5), τ represents a predetermined threshold. All hash codes $HM(X)$ in the database were retrieved so that their Hamming distance from the query hash code $HM(Q)$ does not exceed τ. In practical applications, Hamming sorting and Hamming retrieval are commonly used for fast approximate retrieval, especially in large-scale image or document databases. In summary, this study proposes a CIS detection model in combination with DRH. Fig. 3 shows the model structure.

In Fig. 3, first, the signal database and the information to be queried are inputted. Then, these are extracted by the feature extraction module of DRH. After completion, these features are fed into a deep neural network to learn more complex feature representations. The network output is binarized using an adjusted $sign$ to generate a binary hash code. Finally, these hash codes are used for fast image retrieval, allowing the system to effectively match and search for image data.

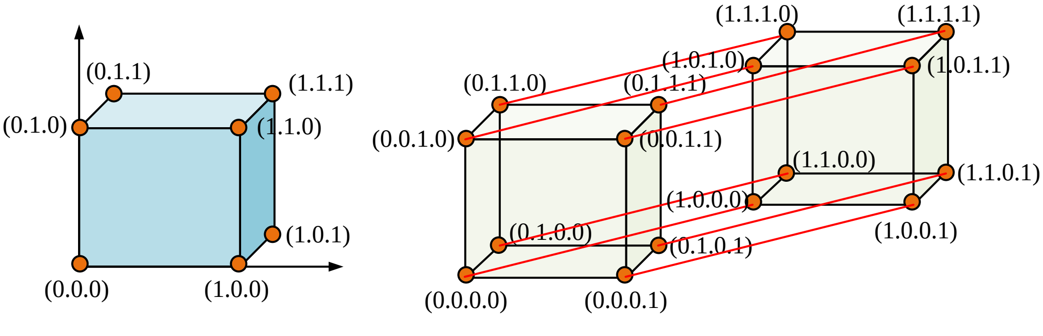
B. *Construction of a Communication Image Detection Model based on Improved Deep-Re-Hash*

DRH provides an effective way to process and retrieve large amounts of image data in image signal detection. The hash center is a point in the hash space that represents the center position of a certain class or cluster of data [20]. The hash center is represented by Eq. (6).

$$C_j = \arg\min_{C_j} \sum_{i=1}^{N} y_{ij} \cdot D(H(x_i), C_j) \tag{6}$$

In Eq. (6), C_j refers to the hash center of category j. x_i represents the i th data point in the training dataset. $H(x_i)$ is a hash representation of data point x_i, typically generated by deep networks. y_{ij} represents an indicator variable. If the data point x_i belongs to category j, then y_{ij} is equal to 1, otherwise it is 0. D represents distance measurement. N refers to the total data points in the training dataset. Fig. 4 shows the hash centers at two different dimensions.

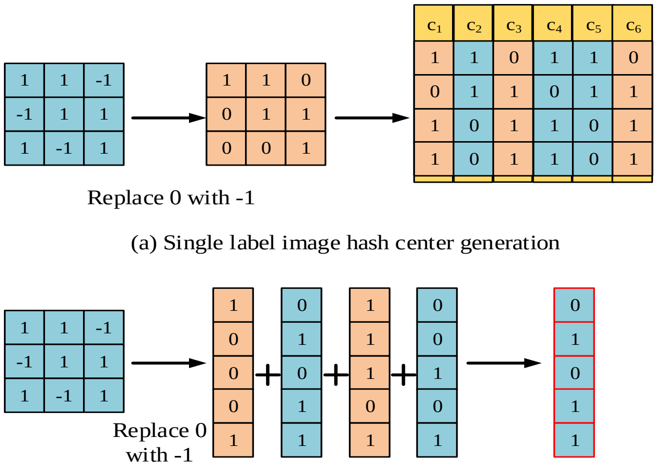
Fig. 4(a) and Fig. 4(b) are schematic diagrams of hash centers at both three-dimensional and four-dimensional scales. The distance between the individual hash code and the hash center is 1. In space, there also exists a consistency relationship with the three-dimensional hash center and hash code. In addition, considering the complex environment of image signals, this study conducts hash center generation calculations for single label and multi-label image signal data, respectively. Fig. 5 shows the generation of hash centers in both environments at this time.



(a) 3D hash center

(b) Four-dimensional hash center

Fig. 4. Schematic diagram of multi-dimensional hash center position.



(a) Single label image hash center generation

(b) Multi label image hash center generation

Fig. 5. Hash center generation for single-label and multi-label image data.

Fig. 5(a) shows the generation of hash centers for single label and multi label image data in various scenarios. In single label image data, each image belongs to only one category. Hash centers are typically representative hash codes of categories. The hash center can be directly obtained by calculating the average value of all image hash codes within the same category. In multi-label image data, each image can belong to multiple categories simultaneously. In this case, the generation of hash centers becomes even more complex, as the impact of multiple labels on each image needs to be considered. The hash center of single label image data is represented by Eq. (7).

$$C_j^s = \frac{1}{N_j}\sum\nolimits_{i=1}^{N} y_{ij}\cdot H(x_i) \qquad (7)$$

The hash center of multi-label image signal data is represented by Eq. (8).

$$C_j^m = \frac{1}{\sum\nolimits_{i=1}^{N} y_{ij}}\sum\nolimits_{i=1}^{N} y_{ij}\cdot H(x_i) \qquad (8)$$

In the case of multiple labels, y_{ij} can be a binary indicator variable or a real value representing the confidence or weight belonging to that category. The calculation of hash centers requires accumulating the contribution of each category on all images, taking into account that each image may contribute to multiple categories. At this point, whether it is single label or multi-label image data, the hash center should have independence and balance, represented by Eq. (0).

$$\begin{cases} B^T B = I \\ B^T 1 = 0 \end{cases} \qquad (9)$$

In Eq. (9), I refers to an identity matrix. In model training, the learning efficiency of hash codes is highest when the hash code is infinitely close to the hash center. To further guide the implementation of this process, this study introduces Cauchy functions to improve the generation process of hash codes. This enables the generated hash code to better reflect the similarity between images while preserving image features. By using the cumulative distribution function of Cauchy distribution, points in deep feature space can be mapped to points in hash space [21], represented by Eq. (10).

$$Y_{cj}(Z`) = \frac{1}{\pi}\arctan(\frac{Z`-C_i}{\gamma}) + \frac{1}{2} \qquad (10)$$

In Eq. (10), $H_{cj}(Z`)$ refers to the mapping value of the feature vector $Z`$ relative to the hash center C_i of class i. γ represents a scale parameter of the Cauchy distribution. After completing the generation of continuous hash codes with probability characteristics, the processing of hash codes continued to be explored in the experiment. Hadamard matrix can be used as a special hash function to generate hash codes. Hadamard matrix is a specific orthogonal matrix whose elements only contain 1 and -1, and its orthogonality ensures the dispersion of hash codes, represented by Eq. (11).

$$H = Had \cdot Z \qquad (11)$$

In Eq. (11), Had represents the Hadamard matrix. The Hadamard matrix serves as the hash center training and sampling matrix, which satisfies Eq. (12).

$$\begin{cases} Had^T \cdot Had = KI_K \\ Had \cdot Had^T = KI_K \end{cases} \qquad (12)$$

In Eq. (12), I_K refers to a K order identity matrix. In the K dimensional Hamming space, if the feature vectors of a set of images are orthogonal to each other, the Hamming distance between the two also meets the requirements described in Eq. (9). Finally, the result generated by the user hash center after combining the Cauchy probability function with Hadamard is represented by Eq. (13).

$$B = sign(\frac{1}{\pi} arc\tan(\frac{Had \cdot Z - C}{\gamma})) \tag{13}$$

In Eq. (13), B refers to the final generated binary hash code. C is a Cauchy center. In summary, after combining Cauchy function and Hadamard matrix to optimize the generation of hash center points, an improved DRH image signal detection search model using deep learning framework is proposed in Fig. 6.

In Fig. 6, firstly, CIS undergoes standardized preprocessing to adapt to the input requirements of the deep learning network. Then, the preprocessed image is feature extracted through a deep neural network. Next, the extracted continuous feature vectors are mapped to a new feature space by the Cauchy probability function, increasing resolution and enhancing the dispersion of the hash code through Hadamard matrix transformation. By utilizing these transformed features, this model calculates the hash center points for each category through optimization algorithms, which guide the generation of compact and discriminative hash codes during the binarization. During model training, similarity learning is also included to ensure that hash codes can reflect the similarity between images. Finally, during the retrieval phase, for a given query image signal, this model will quickly retrieve the image signal in the database with the smallest Hamming distance from its hash code. This model may undergo subsequent sorting and filtering to display the most relevant search results.

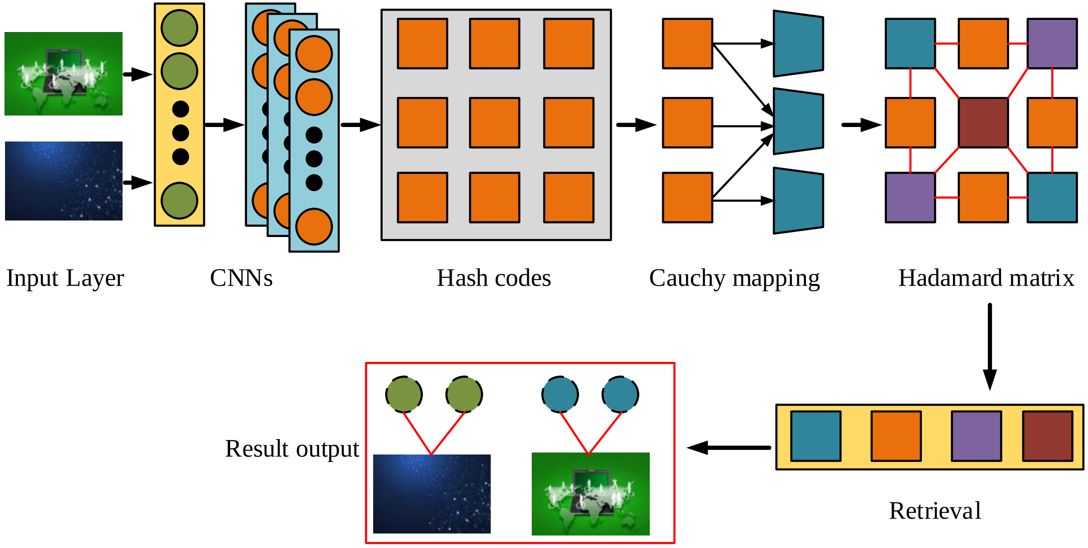


Fig. 6. A communication image signal detection model under improved DRH.

IV. Experimental Testing

A suitable experimental environment was established. Firstly, performance testing was conducted on the algorithm module in the improved DRH, and the training results of the algorithm were compared with algorithms of the same type. Secondly, simulation tests were conducted on the improved DRH detection model to verify its superiority in image signal detection. Tests show that the improved DRH algorithm is particularly suitable for datasets with multidimensional features, such as communication image signal data. Its advantage lies in the processing of large-scale, high-diversity data types, especially in the case of high data dimensionality, the improved DRH algorithm can effectively reduce the computational complexity and improve the retrieval efficiency through the optimized generation of hash codes. Therefore, the algorithm is more suitable for image signal datasets containing a large number of image features with multiple labels, such as CSCOCO and Caltech 256.Especially when dealing with this type of multi-label, high-dimensional data, the algorithm is able to significantly improve the checking rate and retrieval accuracy through the hash center optimized by the Hadamard matrix and the Cauchy function, and the MAP value is significantly higher than that of other algorithms.

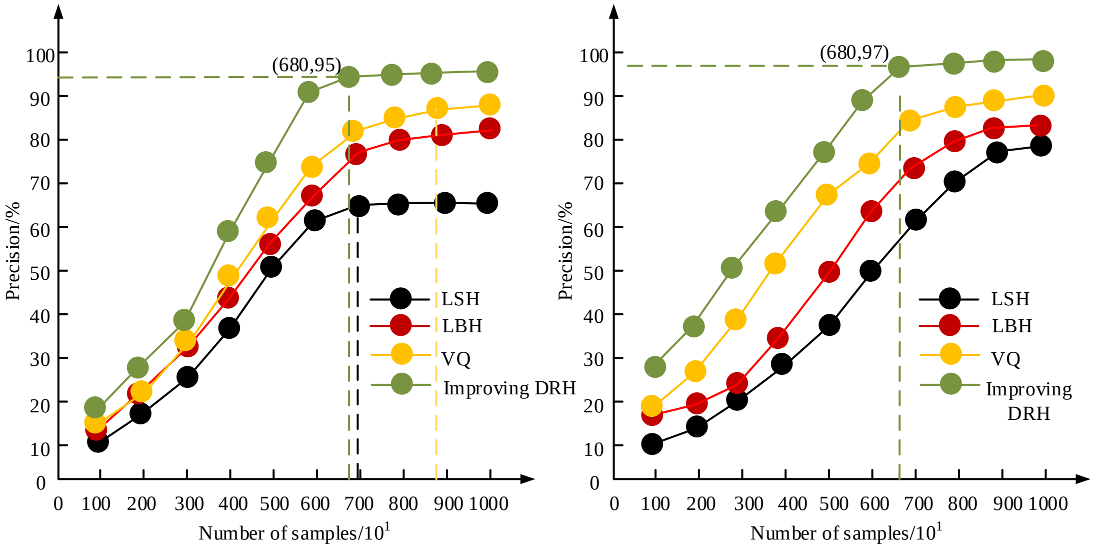
A. Model Performance Testing

A suitable experimental environment was established to verify the effectiveness of the proposed improved DRH image signal detection. Two widely used CIS datasets were introduced, namely CSCOCO and Caltech 256. CSCOCO is a dataset used for image keypoint detection tasks, containing over 200000 images and corresponding annotations. Caltech 256 is a larger dataset that includes 100000 images of different categories. Table I shows the specific experimental testing environment configuration.

Based on the above parameter configuration, the dataset was divided into training and testing sets in an 8:2 ratio. To better fit the characteristics of the test object, the study used precision as the testing indicator. Meanwhile, similar communication image information detection algorithms were introduced, such as Locality-Sensitive Hashing (LSH), Learning-Based Hashing (LBH), and Vector Quantization (VQ). Fig. 7 shows the test results.

TABLE I. ENVIRONMENTAL CONFIGURATION AND PARAMETERS

Environment	**Name**	**Configuration**
Hardware	CPU	Intel Core i7
	Graphics card	NVIDIA GeForce GTX 1060
	Memory	8GB RAM
	Hard disk	15T
Software	Operating environment	Windows 10
	Python	Python 3.8
	Deep learning framework	TensorFlow
	Data set	CSCOCO、Caltech 256
	Development tool	Visual Studio Code



(a) Training set

(b) Test set

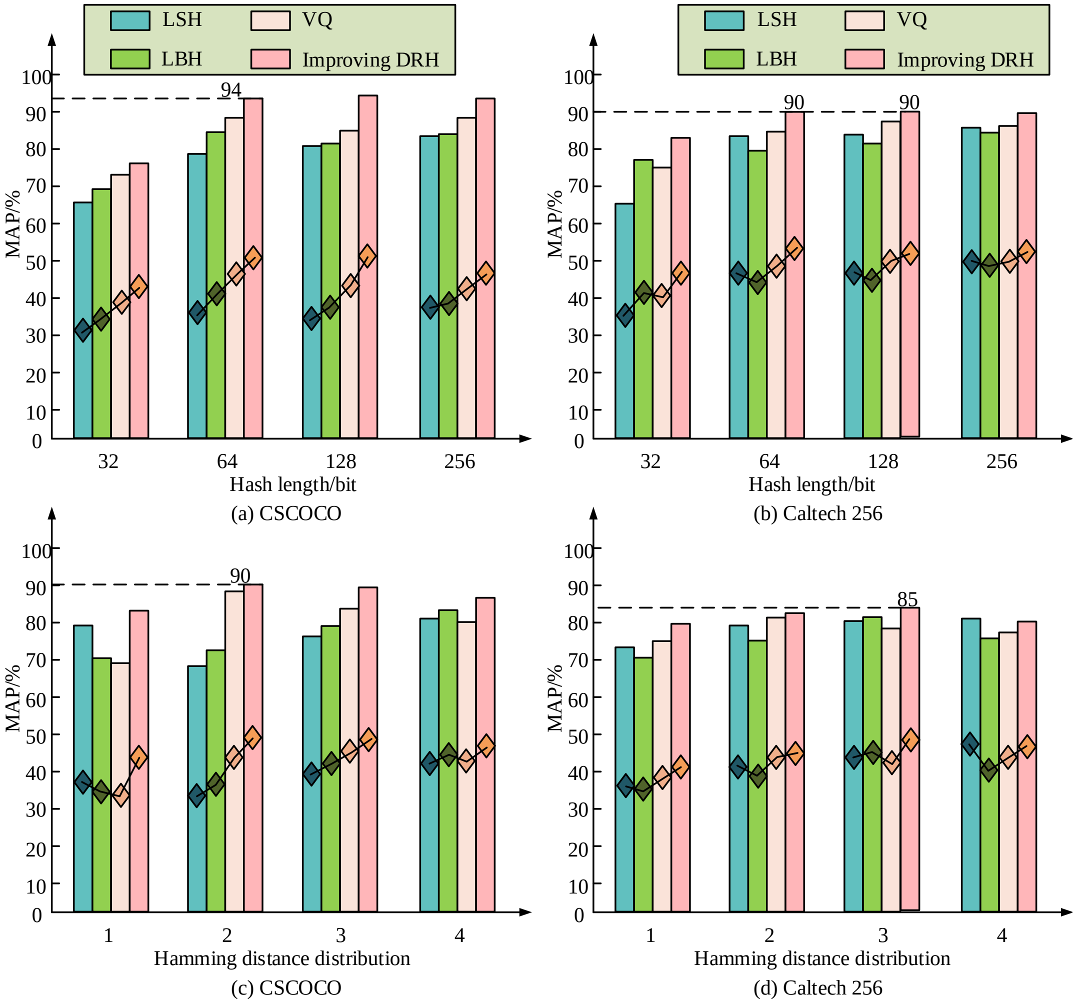
Fig. 7. Performance effects of different algorithms.

Fig. 7(a) shows the performance test results of four algorithms on the training set. Fig. 7(b) shows the performance test results of four algorithms on the test set. In both datasets, the precision of these four algorithms tends to stabilize with increasing sample size. Overall, this improved DRH had the best performance, with a highest precision of 95% in the training set and a minimum sample size of 6800. In the test set, the highest precision of the improved DRH was 97%, with a sample size of 680. Therefore, the improved DRH showed better algorithm performance in both training and testing results. The study continued to use Mean Average Precision (MAP) as an indicator and the length of hash codes and Hamming distance distribution as test variables to discuss the changes in the four algorithms in Fig. 8.

Fig. 8(a) shows the hash length detection results of four algorithms in CSCOCO. Fig. 8(b) shows the hash length detection results of the four algorithms in Caltech 256. Fig. 8(c) shows the Hamming length distribution detection results of four algorithms in CSCOCO. Fig. 8(d) shows the Hamming length distribution detection results of the four algorithms in Caltech 256. As the hash length continued to increase, the MAP of these four algorithms continued to increase. But when the hash length approached an extreme value, the MAP was at its maximum. In the above data, the improved DRH generally had a higher MAP as the hash length changed, with a maximum value of 94%. The hash length at this point was 64 bits. In addition, under different Hamming distance distributions, the improved DRH's MAP was significantly better than other methods, with the optimal MAP of 90% and a Hamming distance distribution value of 2. In theory, longer hash codes can provide more refined feature representations, which may result in higher accuracy during retrieval. However, excessively long hash codes can also lead to dimensional disasters, resulting in reduced retrieval efficiency. In addition, if a retrieval system can effectively map the similarity of images to the Hamming distance of hash codes, then retrieval within a smaller Hamming distance range should achieve higher accuracy and MAP.

B. Model Performance Simulation Testing

The study randomly selected four images each from CSCOCO and Caltech 256 for simulation testing to verify the improved DRH's CIS detection performance. Fig. 9 shows the combined image.



(a) CSCOCO

(b) Caltech 256

(c) CSCOCO

(d) Caltech 256

Fig. 8. Test results of indicators for different models under two types of data.

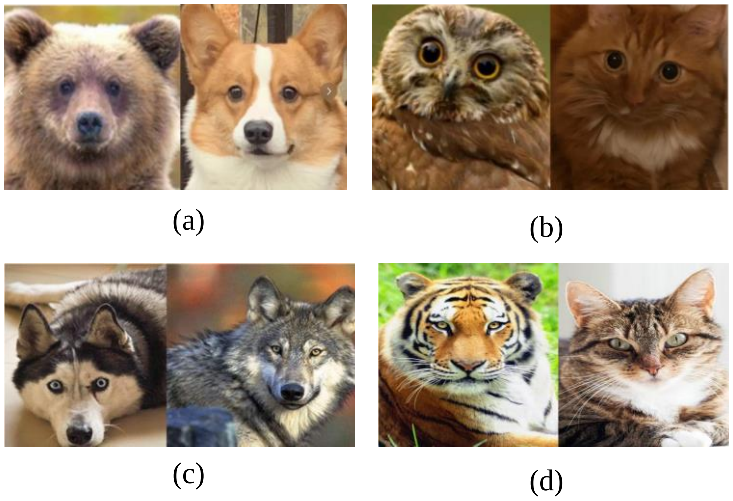


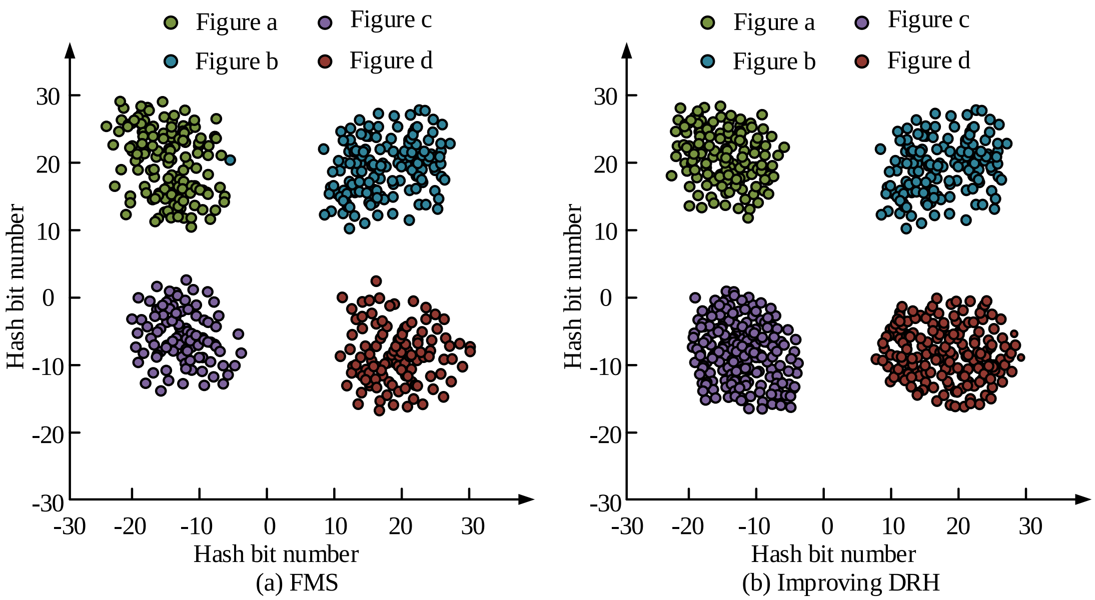
Fig. 9. Four communication image signals to be detected.

Combining these four image signals to be detected in Fig. 9, this study introduced CIS detection models of the same type, such as Content-Based Image Retrieval (CBIR), Feature Matching Search (FMS), and traditional DRH. The previously mentioned previous studies such as the method proposed by Orsuti D et al, the method proposed by Kumar S et al., the method proposed by Zhu L et al. and the method proposed by Sabahi F et al. were also selected for comparison and tested in terms of image signal retrieval similarity and detection time and the results of the test are shown in Table II.

TABLE II. THE DETECTION RESULTS OF FOUR MODELS ON DIFFERENT IMAGES

Image	Molde	Image signal retrieval similarity/%	Retrieval time/s
Figure a	CBIR	93.7	15.4
	FMS	80.7	13.1
	DRH	89.4	10.5
	The method proposed by Orsuti D et al.	88.6	9.7
	The method proposed by Kumar S et al.	85.3	8.3
	The method proposed by Zhu L et al.	85.7	8.6
	The method proposed by Sabahi F et al.	86.9	8.9
	Improving DRH	83.3	7.1
Figure b	CBIR	92.1	16.5
	FMS	83.4	13.4
	DRH	87.4	8.7
	The method proposed by Orsuti D et al.	88.2	9.3
	The method proposed by Kumar S et al.	86.3	8.9
	The method proposed by Zhu L et al.	83.1	7.7
	The method proposed by Sabahi F et al.	83.5	7.2
	Improving DRH	81.1	5.3
Figure c	CBIR	84.9	19.5
	FMS	81.3	19.7
	DRH	85.8	14.6
	The method proposed by Orsuti D et al.	85.4	13.7
	The method proposed by Kumar S et al.	86.6	12.2
	The method proposed by Zhu L et al.	83.9	13.1
	The method proposed by Sabahi F et al.	83.5	12.8
	Improving DRH	81.7	10.7
Figure d	CBIR	79.4	24.7
	FMS	72.4	16.7
	DRH	72.4	14.2
	The method proposed by Orsuti D et al.	70.2	8.9
	The method proposed by Kumar S et al.	69.8	10.2
	The method proposed by Zhu L et al.	65.7	9.7
	The method proposed by Sabahi F et al.	67.4	8.8
	Improving DRH	64.8	7.6

Combining the four groups of extremely similar different species of animals in Fig. 9, if the detection similarity is higher, it indicates that the model's ability to distinguish CIS is insufficient. On the contrary, it indicates that the model can distinguish animal images from each group and meet the requirements of image signal differentiation. Compared with other models such as CBIR, FMS, and conventional DRH, the improved DRH model showed the shortest detection time, especially in Fig. 9(a) and Fig. 9(d), which were shortened by about 8.3 and 17.1 seconds, respectively, and demonstrated a high detection efficiency. In addition, compared with several previous studies, such as the models proposed by Orsuti D et al., Kumar S et al., Zhu L et al., Sabahi F et al., the detection of the improved DRH model is still faster, with the time consumed ranging from 7.1 to 10.7 seconds, respectively, which is lower than that of all the compared models. In terms of image signal retrieval similarity, although the similarity of the improved DRH model is slightly lower than that of the CBIR and FMS models, it still performs stably with a range of retrieval similarity between 64.8% and 83.3%, which shows the better differentiation ability of this model, especially in complex environments where it can maintain a low false detection rate. In the image of Figure d, the similarity of the improved DRH model is 64.8%, which is about 14.6% less compared to the CBIR model, but its detection time is shortened by 17.1 seconds, indicating that the improved DRH model is more suitable for application in time-sensitive scenes. The hash code has a length of 64 bits and a Hamming distance distribution of 2. Fig. 10 shows the visualization result at this time.

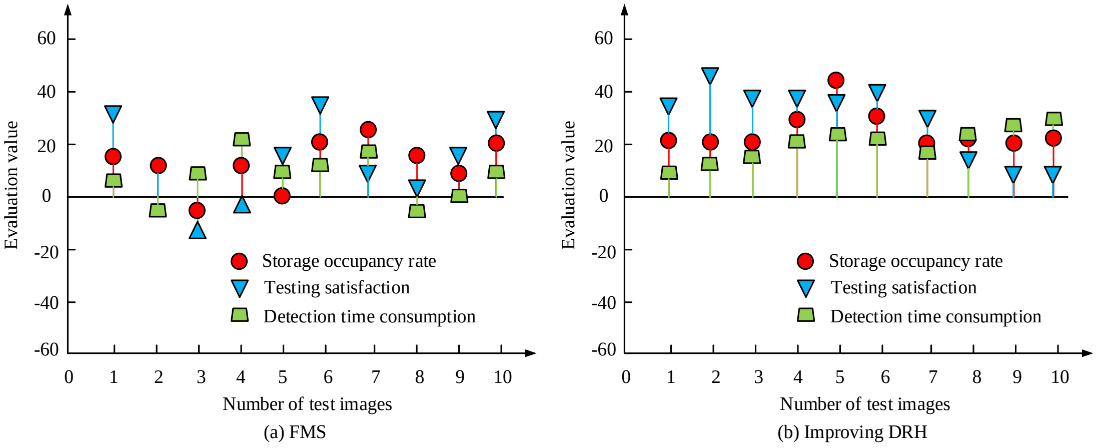


(a) FMS

(b) Improving DRH

Fig. 10. Image visualization results of two algorithms.

Fig. 10 (a) shows the visualization results of four images under FMS. Fig. 10 (b) shows the visualization results of four images under improved DRH. The hash code of FMS was relatively scattered and chaotic, with poor discrimination. On the other hand, the improved DRH generated hash codes that were highly aggregated and had very clear edges. The reason for this is that the improved DRH utilizes Cauchy functions for scaling, while also using Hadamard matrices to generate hash centers and guide hash learning. Finally, the study used a comprehensive scoring method to evaluate storage occupancy rate, detection satisfaction, and time consumption as indicators. Indicator tests were conducted on 10 images under two models, with evaluation values ranging from (-60,60) in Fig. 11.



(a) FMS

(b) Improving DRH

Fig. 11. Multi indicator scoring results of two types of algorithms.

Fig. 11(a) shows the three indicator scores of FMS. Fig. 11(b) shows the three indicator scores of the improved DRH. In FMS, when the detected image signals were 2, 3, 4, 8, and 9, the model storage occupancy, detection satisfaction, and detection time ratings were low, showing unstable fluctuations. On the other hand, the improved DRH had a positive rating for all three indicators. At this point, the storage occupancy rate score was close to 45 points, the detection satisfaction score was close to 50 points, and the detection time score was close to 30 points. In summary, the proposed improved image signal model has certain advantages and stability.

V. Discussion

The study proposes an improved deep hashing algorithm that combines the Cauchy function and Hadamard matrix, and the method optimizes the retrieval performance of image signals by generating more compact and discriminative hash codes. The experimental results show that the checking accuracy of the proposed method under study reaches 97% and 95% on CSCOCO and Caltech 256 datasets, respectively. In addition, the MAP value of the model reaches up to 90% when the hash code length is 64bit, which indicates that the research has significantly improved the accuracy while ensuring the retrieval speed. By optimizing the distribution of Hamming distances, the model is able to achieve high-precision similarity search within a small range of Hamming distances. In comparison with other related studies, the improved DRH algorithm shows unique advantages. Orsuti D et al. showed that the feature extraction of image signals using CNN, although it can obtain better results on small-scale datasets, the computational overhead of the model increases significantly in large-scale, multi-labeled datasets, resulting in a decrease in the retrieval speed. In contrast, the study reduces the computational complexity in the convolutional network and improves the processing speed through the optimal generation of hash codes. In addition, it is found that the improved DRH algorithm achieves the highest MAP value of 90% when the Hamming distance is 2, while the MAP value of the other methods is only about 70% under the same conditions. This result shows that the proposed method of the study is still able to maintain a high similarity retrieval accuracy in the low-dimensional hash space of large-scale datasets. This is in contrast to the results of Wu G et al. whose proposed algorithm rapidly decreases the retrieval accuracy when the Hamming distance is small [22]. It can be shown that the study effectively solves such problems through the improved hash center generation mechanism, which enables better performance to be maintained when performing low-dimensional retrieval in large-scale communication image signals.

In summary, the improved DRH algorithm proposed by the study successfully improves the retrieval efficiency and accuracy of large-scale communication image signals through the optimization of the Cauchy function and Hadamard matrix. By comparing with other studies, it can be seen that the algorithm performs superiorly in dealing with multi-label and high-dimensional data, especially in the optimization of retrieval accuracy and Hamming distance with unique advantages. Future work can continue to explore the application of the model in different types of data and further optimize its computational performance to cope with the processing demands of larger-scale signal datasets.

VI. Conclusion

With the developing communication technology, the transmission and processing of CIS becomes more important in various fields. In this context, intelligent detection and search of CIS becomes a highly focused research direction. In view of this, firstly, this study constructed an image signal detection search model for traditional DRH. Secondly, the Cauchy concept function and Hadamard matrix were introduced to optimize the generation of hash centers. Finally, an improved DRH image signal detection model was proposed. When the test samples were 680, the highest precision of the new model was 97%. Compared with the same type of algorithm, when the Hamming distance is 2, at this time, the MAP value of the improved DRH algorithm is up to 90%, and the hash length is 64bit, which is significantly better than other algorithms. Simulation tests show that the detection similarity of the improved DRH model can be as low as 64.8%, and the fastest detection speed is 7.6 s. At this time, the similarity is reduced by about 14.6% compared with the CBIR model, and the time is shortened by 17.1 seconds. In addition, in the visualization results, the hash code generated by the improved DRH model is extremely aggregated with very clear edges, which is clearly superior. The highest storage occupancy score, detection satisfaction score, and detection time consumed score for this new model are 45, 50, and 30, respectively, which are higher than all other three types of models. In summary, the proposed improved DRH should not only have efficient detection and search capabilities, but also consider the diversity and dynamism of CIS in practical applications. In the future, this model can be further optimized to consider more practical application scenarios to promote the CIS processing.

References

[1] [1] Chiedu Okwudili Maduekeh, Ifeoma Nancy Obinwa. Impacts Of The Public Procurement Act 2007 On The Procurement Of Public Projects In Nigerian Tertiary Institutions. Malaysian E Commerce Journal. 2022; 6 (2): 89-95.

[2] Hasbi Mubarak Suud. An Image Processing Approach For Monitoring Soil Plowing Based On Drone Rgb Images. Big Data in Agriculture (BDA), 2022, 5(1).

[3] Azman Brahim Llaguno Garciaa, Arturo Orellana Garciab, Vivian Estrada Sentic. Thermal Images Pre-Processing For Early Detection Of Breast Cancer: A Progressive Review. Acta Informatica Malaysia. 2024; 8(1): 26-31.

[4] Hu B, Wang J. A weighted multi-source domain adaptation approach for surface defect detection. IET Image Processing, 2022, 16(8):2210-2218.

[5] Nguyen D T, Park Y. Performance enhancement of optical camera communication system using optical camera communication coding and region-of-interest detection. IET Optoelectronics, 2021, 15(6):255-263.

[6] You Y H, Park J H, Ahn I Y. Complexity Effective Sequential Detection of Secondary Synchronization Signal for 5G New Radio Communication Systems. IEEE Systems Journal, 2020, 15(3):3382-3390.

[7] Li K, Zhou W, Lia H, Mark A A. Assessing the Impact of Deep Neural Network-based Image Denoising on Binary Signal Detection Tasks. IEEE Transactions on Medical Imaging, 2021, 40(9):2295-2305.

[8] Wen X, Lai H, Gao G, Zhao Y. Video abnormal behaviour detection based on pseudo-3D encoder and multi-cascade memory mechanism. IET image processing, 2023, 17(3):709-721.

[9] Orsuti D, Antonelli C, Chiuso A, Santagiustina M, Mecozzi A, Galtarossa A, Palmieri L. Deep learning-based phase retrieval scheme for minimum-phase signal recovery. Journal of Lightwave Technology, 2023, 41(2): 578-592.

[10] Kumar S, Pal A K, Islam S K H, Hammoudeh M. Secure and efficient image retrieval through invariant features selection in insecure cloud environments. Neural Computing and Applications, 2023, 35(7): 4855-4880.

[11] Qin Q, Wei Z, Huang L, Xie K, Zhang W. Deep Top Similarity Hashing with Class-wise Loss for Multi-label Image Retrieval. Neurocomputing, 2021, 439(11):302-315.

[12] Shuai C, Wang X, He M, Ouyang X. A presentation and retrieval hash scheme of images based on principal component analysis. The Visual Computer, 2021, 37(8):2113-2126.

[13] Shen X, Zhang H, Li L, Zhang Z, Chen D, Liu L. Clustering-driven Deep Adversarial Hashing for scalable unsupervised cross-modal retrieval. Neurocomputing, 2021, 459(10):152-164.

[14] Zhang D, Wu X, Yin H, Kittler J. MOON: Multi-Hash Codes Joint Learning for Cross-Media Retrieval. Pattern Recognition Letters, 2021, 151(11):19-25.

[15] Zhu L, Zheng C, Guan W, Li J. Multi-modal hashing for efficient multimedia retrieval: A survey. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(1): 239-260.

[16] Sabahi F, Ahmad M O, Swamy M N S. RefinerHash: a new hashing-based re-ranking technique for image retrieval. Multimedia Systems, 2024, 30(3): 119-121.

[17] Mengzhu Y, Zhenjun T, Zhixin L. Robust Image Hashing with Saliency Map and Sparse Model. The Computer Journal, 2022, 66(5):1241-1255.

[18] Cai T, Gao P, Niu D. NEHASH: high-concurrency extendible hashing for non-volatile memory. Frontiers of Information Technology & Electronic Engineering, 2023, 24(5):703-715.

[19] Chen Z, Tang Z, Zhang X, Sun R, Zhang X. Efficient video hashing based on low-rank frames. IET image processing, 2022, 16(2):344-355.

[20] Kamal M A S, Hashikura K, Hayakawa T. Look-Ahead Driving Schemes for Efficient Control of Automated Vehicles on Urban Roads. IEEE Transactions on Vehicular Technology, 2022, 71(2):1280-1292.

[21] Preethi P, Mamatha H R. Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images, Artif Intell Appl, 2023, 1(2):119-127.

[22] Wu G, Jin E, Sun Y, Tang B, Zhao W. Deep Attention Fusion Hashing (DAFH) Model for Medical Image Retrieval. Bioengineering, 2024, 11(7): 673-675.

Content-Based Image Retrieval Using Transfer Learning and Vector Database

Li Shuo, Lilly Suriani Affendey, Fatimah Sidi
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

***Abstract*—Content-based image retrieval (CBIR) systems are essential for efficiently searching large image datasets using image features instead of text annotations. Major challenges include extracting effective feature representations to improve accuracy, as well as indexing them to improve the retrieval speed. The use of pre-trained deep learning models to extract features has elicited interest from researchers. In addition, the emergence of open-source vector databases allows efficient vector indexing which significantly increases the speed of similarity search. This paper introduces a novel CBIR system that combines transfer learning with vector databases to improve retrieval speed and accuracy. Using a pre-trained VGG-16 model, we extract high-dimensional feature vectors from images, which are stored and retrieved using the Milvus vector database. Our approach significantly reduces retrieval time, achieving real-time responses while maintaining high precision and recall. Experiments conducted on ImageClef, ImageNet, and Corel-1k datasets demonstrate the system's effectiveness in large-scale image retrieval tasks, outperforming traditional methods in both speed and accuracy.**

Keywords—Content-based image retrieval (CBIR); image retrieval; transfer learning; convolutional neural networks; VGG-16; vector database; milvus; feature extraction; high-dimensional vectors; real-time image search

I. Introduction

In the Internet era, massive amounts of multimedia data (including text, images, audio, video, etc.) are continuously generated and stored. How to effectively retrieve relevant information from a huge image dataset has become an urgent problem to be solved. Traditional text-based image retrieval methods can no longer meet this demand, while content-based image retrieval (CBIR) technology retrieves through image features, significantly improving the accuracy and efficiency of retrieval. Practical applications of CBIR include e-commerce, web search, and medical image analysis.

The core of CBIR technology includes two key steps: image feature extraction (indexing stage) and similarity matching (retrieval stage). Image feature extraction is to convert the original image content into feature vectors, while similarity matching is to compare the feature vectors of the query image with the image in the dataset, calculate their similarity, and identify similar images. Research on CBIR technology has a long history, nevertheless most work mainly focus on model training and evaluation of image classification tasks, measuring the accuracy, and not considering the time consumption for indexing or retrieval.

This paper proposes a new CBIR retrieval system that combines transfer learning technology and vector database. Image features are extracted through convolutional neural networks (CNNs), and high-dimensional vectors are stored in vector databases to achieve efficient image retrieval. We evaluated the performance of the indexing and retrieval stages and verified the feasibility and effectiveness of the scheme. This study aims to explore how to build a high-performance CBIR system to meet the real-time retrieval requirements in large-scale image datasets.

This paper is organized as follows: Section II discusses related research in the field of CBIR and the techniques used in this study; Section III details the methodology adopted in this study; Section IV describes the proposed CBIR system utilizing transfer learning and vector database; Section V presents the experimental results, where we evaluate the performance of our system on various benchmark datasets. In Section VI, we conclude the paper by summarizing the findings, discussing the implications of the proposed system, and suggesting potential directions for future work.

II. Related Work

The theory of image retrieval has been around for a long time. Initially, text annotation was used to store image descriptive text in a database for retrieval. However, this method is manual and subjective as well as inefficient and inaccurate when faced with large amounts of data [1]. Subsequently, researchers began to study CBIR technology, using low-level features such as color, shape, and texture to represent images [2]. This method is automated and efficient, but prone to errors and has low accuracy in complex image recognition.

Modern CBIR technology extracts features from images through deep learning, uses convolutional neural networks to vectorize images, and performs matching through similarity calculations [3]. Currently, commonly used similarity matching algorithms include Euclidean distance, cosine similarity, etc.

Based on these technologies, many CBIR applications have emerged, mainly focusing on the construction, training, and evaluation of models in classification tasks, and evaluating their performance by classifying images from different data sets [4] Thus, to improve the performance of CBIR systems, numerous research on feature extraction, similarity matching, as well as storage architecture have been conducted by different researchers.

Traina et al. [5] proposed the SIFT (Scale-Invariant Feature Transform) machine vision algorithm for feature extraction and used K-means to calculate similarity distances. Simran et al. [6] introduced the use of deep learning techniques for extracting image features. Kumar et al. [7] proposed using DarkNet-19 and

DarkNet-53 models for feature extraction, along with PCA for dimensionality reduction. Sikandar et al. [8] proposed the use of ResNet50 and VGG16 for feature extraction, employing KNN for similarity calculations.

Similarity calculation is also very important. Alsmadi [11] used the memetic algorithm method to calculate similarity. Sikandar et al. [8] used the KNN method to calculate similarity. Niu et al. [12] used the Residual Vector Product Quantization for approximate nearest neighbor method to calculate.

Retrieval performance is an important indicator of CBIR system, which is usually evaluated by accuracy. Some relevant studies have been collected for retrieval performance. Chughtai et al. [13] used transfer learning to call VGG16, VGG19, EfficientNetB0, ResNet50 and other models for CBIR, with a retrieval accuracy of up to 96%. Mohammed et al. [14][14] used two pre-trained deep learning models ResNet50 and VGG16 and a machine learning model KNN implementation to achieve a maximum accuracy of 100%. Gautam et al. [15] used VGG16 and ResNet-50 architectures to obtain a maximum accuracy of 90.18%. Sadiq et al. [16] combined NASNetMobile, DenseNet121 and VGG16 models to achieve a maximum accuracy of 98%. Thanikachalam et al. [17] proposed Tokens-to-Token Vision Transformer (T2T-ViT), a novel CBIR method with an accuracy of up to 99.42%.

The above literature has conducted in-depth research on CBIR, with different retrieval methods and good performance. However, most work do not measure the time consumption for indexing, nor do they report the retrieval time, which is also an important indicator for measuring the performance of CBIR systems.

Mezzoudj et al. [18] on the other hand used a big data solution to retrieve and index CBIR, and reported the time for the indexing and retrieval stages of CBIR using the ImageClef and ImageNet datasets. Their solution greatly improved the efficiency of CBIR. However, they did not report the retrieval accuracy.

Recently, Stata et al. [9][9] and Singla et al. [10] focused on utilizing vector databases for vector extraction and evaluated the performance of these databases.

From the literature, we attempt to implement CBIR that uses transfer learning feature extraction, and utilize a vector database to store the high-dimensional vectors in, so that the vector data volume undertakes the retrieval task and improves the retrieval efficiency.

III. Methodology

This section describes the experimental design process in detail, including experimental steps, application architecture design, software and hardware configuration, data set description, experimental evaluation indicators, and experimental verification methods.

A. Application Architecture Design

Fig. 1 is the application architecture design diagram. First, the pre-trained VGG model is used to process each image in the dataset, converting these images into high-dimensional vectors. This step is called vectorization, which converts the pixel information of the image into feature vectors to facilitate subsequent similarity calculations. The generated high-dimensional vectors are then stored in the Milvus vector database for efficient storage and retrieval operations. Milvus is a high-performance distributed vector database that can store high-dimensional vectors and quickly retrieve massive amounts of vector data. VGG16 is a classic convolutional neural network (CNN) model that efficiently extracts image features. Transfer learning technology allows the VGG model to be used directly for vectorizing images, reducing the model training process and improving work efficiency.

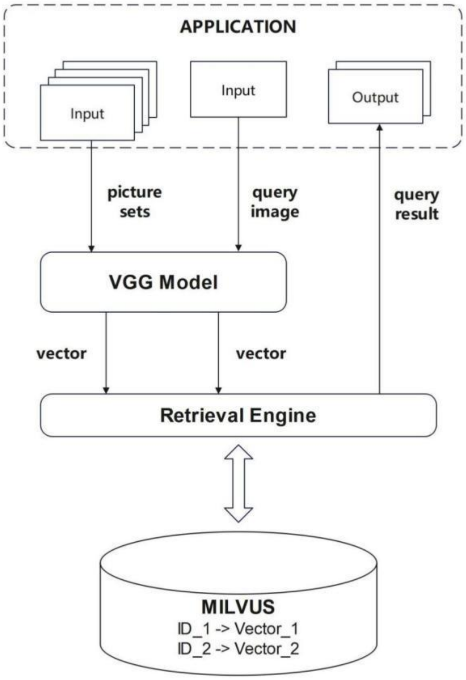


Fig. 1. Application architecture design diagram.

When a user submits an image for retrieval and matching, the input image is processed in the same manner. Specifically, the input image will also be vectorized through the VGG model to generate its corresponding high-dimensional feature vector. This feature vector is then used to query the Milvus database to find the vector that is most similar to it.

The Milvus database calculates vectors that are similar to the query vector and returns a list of results sorted according to the similarity score. The similarity score represents the similarity distance between the query vector and the vectors stored in the Milvus database. The lower the similarity score, the more similar the two images are. Finally, the system retrieves the images corresponding to these vectors, sorts them by similarity, and displays them to the user. In this way, users can easily find other images similar to their input image to meet the needs of image retrieval and matching.

B. Hardware

This experiment uses hardware resources of Intel Quad Core i5-4210U CPU, 1.70 GHz, and 8 GB memory. Table I shows the hardware required for this project.

TABLE I. HARDWARE INFORMATION

Hardware	Specification
Central Processing Unit (CPU)	Intel Core i5 4210 @ 1.7 GHz
Primary Memory (Random Access Memory)	8 GB
Secondary Memory (Hard Disc Drive)	1024 GB Solid State Drive

C. Software and Frameworks

Python was chosen as the development language. The reason is that it contains many mature function libraries and frameworks. Table II shows the software and framework required for this project.

TABLE II. SOFTWARE INFORMATION

Software	Version	Purpose
Python	3.6.0	Provide a basic development environment for the project
Tensorflow	1.15.4	An open-source deep learning framework that integrates many algorithms and models to facilitate model training.
Keras	2.3.1	A machine learning framework repackaged based on TensorFlow, an open-source high-level API neural network framework. Simplifies use and facilitates development.
Numpy	1.16.5	Integrated function library that supports a large number of dimensional array and matrix operations.
Matplotlib	3.7.2	A two-dimensional drawing library developed in Python language.
Pillow	7.1.0	Get the image based on the image path and
Flask	2.0.3	Start a service to accept and process request requests
Milvus	1.0.0	Vector database stores high-dimensional vectors after vectorization of images. And used to query and retrieve similar vectors.
Docker	19.03	Install and run Milvus and the CBIR project

D. ImageClef Dataset

ImageClef is a series of events, and a different dataset is released each year, so the size of the dataset will vary with the event and year. The ImageCleF dataset contains images and related annotation information, and researchers can conduct research on various image understanding tasks, such as image classification, image annotation, image retrieval, etc. The dataset has a total of 20,000 images, which are numerically classified into folders. This is a photo book, taken from all over the world. These datasets usually cover different subject areas and multiple languages, providing a rich research foundation.

E. ImageNet Dataset

The ImageNet dataset is a computer vision dataset. It is a large image dataset established to promote the development of computer image recognition technology. Many well-known models have been trained on this dataset, such as VGG-16, VGG-19, Restnet-50, etc. The images in the ImageNet dataset cover most of the image categories seen in daily life. This is a dataset with more than one million images.

F. Corel-1k Dataset

The Corel-1000 dataset is widely used in the CBIR field, as many researchers use this dataset to evaluate the quality of image retrieval tasks. It contains a total of 1000 images in 10 categories, namely: Africa, Beach, Building, Bus, Dinosaur, Elephant, Flower, Food, Horse, and Monument. There are 100 images in each category. This dataset was used to evaluate the quality of retrieval for CBIR applications.

G. Experimental Evaluation Metrics

To evaluate the indexing performance, the time consumption for indexing of images of different sizes in the imageClef data were measured.

To evaluate the retrieval performance, the retrieval time for images of different sizes from the imageClef dataset were taken.

In addition, the time consumption of indexing and retrieval using the ImageNet dataset was measured to evaluate the performance in a large dataset.

Finally, the retrieval quality was evaluated using the precision and recall metrics on the Corel-1k dataset to measure the retrieval accuracy.

IV. CBIR USING TRANSFER LEARNING AND VDBMS

In this section, the implementation of the CBIR experiment will be discussed in detail. First, the relevant environment is deployed, the VGG-16 model is built, and the dataset is preprocessed. Subsequently, the execution time of the indexing and retrieval stages with different data amounts is recorded on the imageClef dataset, and the accuracy of the retrieval stage is tested on the Corel-1k dataset. Finally, the experiment is repeated using the ImageNet dataset and the relevant data is recorded.

A. Data Preprocessing

Since the images are stored in various folders and are nested, it is necessary to extract the images from various directories and then merge them into the same directory folder. This makes it easier to process the images. At the same time, filter out non-image files. Finally, the dataset is compressed packaged and uploaded to the specified path on the server for subsequent use.

B. Experimental Environment Setup

In this step, the operating environment needs to be prepared. All software and the related dependent libraries as shown in Table II are installed. The project code can be found at https://github.com/LI-SHUO-lee/CBIR.git.

C. Initialize VGG Model

In this work, transfer learning is utilized to directly employ the VGG model for feature extraction. Initially, the VGG model needs to be defined. The VGG16 model requires the input image size to be 224x224 pixels with three channels. Pre-trained weights on the ImageNet dataset are utilized here. These pre-trained weights facilitate faster convergence and generally enhance the model's performance. Max pooling is employed, which selects the maximum value in each region as the output. The parameter includes top=False indicating that the top layer (fully connected layer) of the model is excluded, as a custom output layer is intended to be added on top. An array of zeros with shape (1, 224, 224, 3) is created using np.zeros((1, 224, 224, 3)) as input, and the model's prediction method is called to make a prediction. This is typically performed to ensure that the model is loaded correctly and can make a normal prediction.

D. CBIR Indexing Phase

The indexing stage refers to the process of directly using existing models and parameters to extract features from images through transfer learning technology and storing these high-dimensional vectors describing image features in the vector database so that subsequent retrieval processes can be carried out effectively.

Fig. 2 shows the process of image indexing. All images are preprocessed to ensure that their suffix is jpg or png, and at the same time unify the size of the images to ensure that they are 224*224*3. Then the features of the image are extracted through the VGG model. In order to reduce the pressure on the vector database Milvus, high-dimensional vectors are temporarily stored in the server cache. After waiting for the image batch vectorization to be completed, the vectors in the server cache are flushed into the vector database Milvus to ensure that the database is only requested once. This ensures the function and improves efficiency.

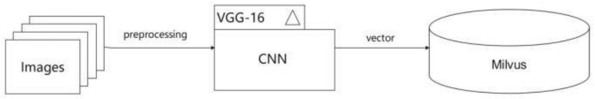


Fig. 2. Indexing process.

Table III shows the time taken for indexing different image sizes using the ImageClef dataset.

TABLE III. THE TIME CONSUMPTION OF INDEXING

Image size	2011	5952	10000	11963	17953	20000
Time consumption	446 s	1066 s	1689 s	1991 s	3089 s	3412 s

E. CBIR Retrieval Phase

Conducting an experimental study on the retrieval stage of the CBIR application is essential, as it is the core component of the project. The retrieval phase's success directly impacts the overall project outcome. Therefore, an in-depth experimental exploration is necessary to address its challenges.

Firstly, we will evaluate retrieval time, a crucial performance metric. Assessing retrieval time helps understand the system's efficiency in handling a large volume of images, reflecting architectural improvements.

Secondly, accuracy evaluation is vital in CBIR. We will compare search results with standard dataset to assess the accuracy.

1) Retrieval time consumption evaluation: Fig. 3 illustrates the image-retrieving process. When the data set is fully vectorized, the database stores the feature vector of each image in the data set. The user only needs to input a query image, and then use the same model and algorithm to vectorize and extract features of the image to obtain a feature vector. Only the feature vector and the number of data pieces needing fuzzy matching need to be communicated to Milvus. Milvus will automatically search from its own library through the vector, and then return the closest Top k vectors and vector IDs. Next, the table is consulted to find the correspondence between the vector ID and the actual image, enabling the retrieval of the fuzzy-matched image to be returned to the user. This enables users to search for images by image. The time consumed in this stage is used as a key indicator for evaluating the CBIR application.

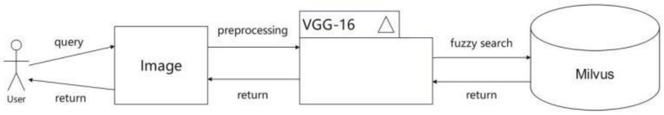


Fig. 3. Image retrieval.

First, the ImageClef dataset is used to perform the retrieval process according to the image size in Table IV. Then the consumed time is recorded respectively to evaluate the performance of the retrieval process of the CBIR application.

TABLE IV. THE TIME CONSUMPTION OF RETRIEVAL

Image size	2011	5952	10000	11963	17953	20000
Time consumption	0.27 s	0.26 s	0.33 s	0.23 s	0.20 s	0.30 s

2) Retrieval accuracy evaluation: Retrieval accuracy is also an important indicator of retrieval performance. Since the vector database is an approximate search, the search results may not be 100% correct. However, accuracy and speed are contradictory indicators. We cannot blindly pursue speed and ignore the accuracy of CBIR. Therefore, at this stage, we will evaluate and record the retrieval accuracy on the Corel-1k dataset, which will also be used to evaluate the performance of CBIR applications. Fig. 4 illustrates the process of evaluating the accuracy.

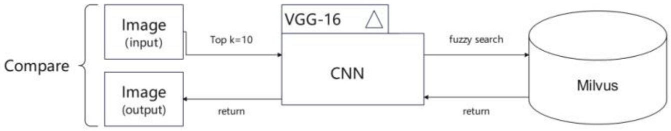


Fig. 4. The accuracy process.

In this experiment, precision and recall will be used to evaluate the retrieval quality. The Corel-1k dataset will be used because many papers use this dataset to calculate precision and recall for evaluating CBIR applications. There are 10 categories in Corel-1k, each with 100 images, for a total of 1,000 images. One image is randomly selected from each category for query testing, and the remaining 99 are used for training. Therefore, 990 images need to be vectorized, converted into vectors through the VGG-16 model, and stored in Milvus. The Euclidean distance (L2) is used to calculate the distance between vectors. Setting Top k = 10, which means that 10 images will be returned for each query. For the returned images, we extract the classification label of the image and then compare it with the classification label of the query image. If the classification label is consistent, the retrieval is considered correct, and the application has successfully detected the image of this class. Otherwise, it is an error. Record the number of correctly classified images and compare them with the returned images and all images of this class in the database to calculate the precision and recall. Finally, the process is repeated for each image in each category, and the precision and recall rate of each category are calculated to verify the retrieval performance of the application.

Precision formula:

$$\text{Precision} = \frac{\text{number of similar images retrieved}}{\text{total number of images retrieved}} \tag{1}$$

Recall formula:

$$\text{Recall} = \frac{\text{number of similar images retrieved}}{\text{total number of similar images in the database}} \tag{2}$$

Table V shows the calculation results of precision and recall through the experiment.

TABLE V. THE ACCURACY FOR DIFFERENT IMAGE CATEGORIES

	Precision	**Recall**
africans	0.869	0.711
beaches	0.847	0.592
buildings	0.898	0.666
buses	1.000	1.000
dinosaurs	1.000	0.956
elephants	1.000	0.853
flowers	0.994	0.804
foods	0.983	0.805
horses	1.000	0.828
monuments	0.935	0.716

F. *Time Consumption in the ImageNet Dataset*

To verify time consumption on the ImageNet dataset, we will use the same method. Specifically, we will record the time taken for both the indexing and retrieval phases. Repeating the experiment on different datasets will help verify the robustness and wide applicability of the proposed method. Table VI shows the experiment results.

TABLE VI. THE RESEARCH RESULTS IN THE IMAGENET DATASET

ImageNet size	**Time for indexing**	**Time for retrieval**
1,461,406	249,400 s	0.5 s

V. RESULTS AND DISCUSSION

A. *Introduction*

This section will provide a detailed analysis of the project experiment data. We will collect, organize, visualize, and statistically analyze the experimental data. Using the reference paper [18] as a benchmark, we will compare and discuss various aspects such as architecture design, algorithm comparison, development language, index performance, and retrieval performance to demonstrate the superiority of our architecture.

B. *Architecture Comparison*

The architecture has the following advantages over Mezzoudj [18].

First, the VCBIR architecture is simpler, requiring no big data platform; it can be deployed and maintained on a single server, facilitating horizontal scaling.

Second, image vectorization uses the VGG-16 model, which provides more accurate feature extraction while maintaining high efficiency. If higher accuracy or a lighter model is needed, the vector model can be easily replaced without altering other components.

Finally, vector storage uses Milvus instead of Hadoop, eliminating the need for big data components like Spark for vector calculations, thus simplifying the architecture. Milvus supports horizontal scaling, allowing for upgrades to a distributed cluster in case of storage bottlenecks, without impacting upper-layer applications. Table VII summarizes the architectural differences.

TABLE VII. ARCHITECTURE COMPARISON TABLE

	Architecture from reference paper [18]	**VCBIR Architecture**
Vector model	CS-LBP extracts image features based on color, texture, etc., but it has low accuracy.	VGG extracts image features based on neural networks and has more layers with high accuracy.
Operating environment	Rely on the big data platform.	The ordinary operating system is sufficient.
Resource consumption	High resource consumption.	Low resource consumption.
Deployment	Hard to deploy.	Easy to deploy.
Maintain	The entire big data platform needs to be maintained. High maintenance costs.	Only Milvus and programs need to be maintained. Low maintenance costs.
Expand	Easy to extend	Easy to extend

C. *Performance of the Indexing Module*

1) Data analysis: The experimental data was collected from the CBIR indexing phase and compared with the study [18]. Table VIII shows the comparison of retrieval time results.

TABLE VIII. TIME CONSUMPTION OF THE INDEXING FOR DIFFERENT DATASET SIZES AND METHODS

Images size	**HDFS**	**Tachyon**	**VCBIR**
2011	240 s	120 s	446 s
5952	660 s	420 s	1066 s
11963	1260 s	720 s	1991s
17953	1380 s	1140 s	3089 s
20000	1500 s	1200 s	3412 s

Fig. 5 visualizes the graph for comparison. The asterisk line graph represents the time consumed by the method (VCBIR) in the indexing phase. As can be seen from the figure, the performance of this method in the indexing phase is lower than that of the method in study [18] and presents a linear distribution. The reason will be analyzed in detail later.

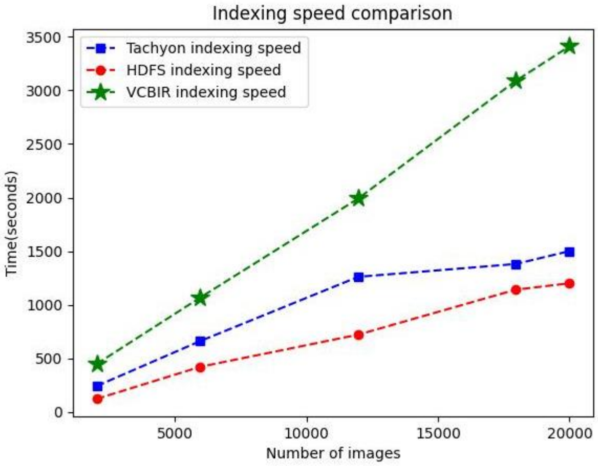


Fig. 5. Indexing speed for different methods.

The research on indexing performance from relevant literature was reviewed and compared. Table IX presents the index performance from different studies and Fig. 6 visualizes the comparison.

TABLE IX. PERFORMANCE COMPARISON OF DIFFERENT INDEXING APPROACHES FOR 10,000 IMAGES

Approach	Description	Time(s)
Centralize method [20]	Sequential method on 1 node	600,000
DIRS method [20]	HBase system+MapReduce on 9 Hadoop nodes	200,000
Luca C. et R. method [21]	HDFS+MapReduce on 1 Hadoop node	2,820
VCBIR method	Milvus +VGG16 on 1 node	1689
Reference paper method [18]	Tachyon+MapReduce on 1 Spark node	460

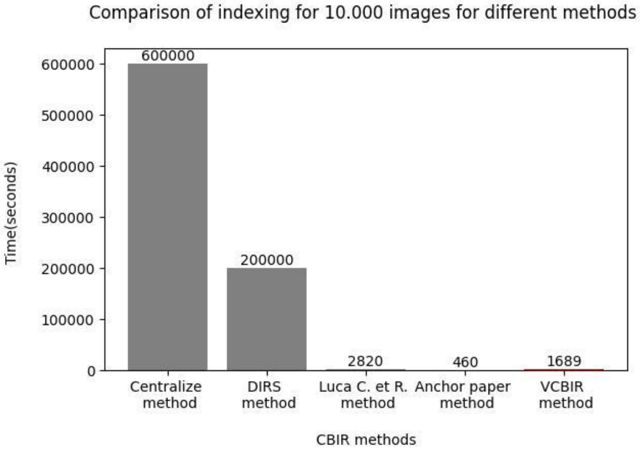


Fig. 6. The time cost visualization for different methods.

From Fig. 6. and Table IX, it can be observed that although the CBIR index performance based on the vector database is a little worse, the performance is still greatly improved in this field. It validates the feasibility of the approach.

2) Result analysis: The reasons why this solution is weaker than the big data solution in study [18] at the indexing stage are as follows:

- Since the literature [18] uses a big data solution, it is easy to implement distributed applications based on the big data platform, so efficiency can be greatly improved. However, the VCBIR is developed using Python. Global Interpreter Lock (GIL) is a mechanism in the Python interpreter. This lock greatly limits the multi-threading performance of the program, which is the main reason why the performance of Python programs is weaker than that of Java. The VGG model relies on tensorflow, and it does not support multi-process execution. Therefore, the entire program can only be executed in a single thread and a single process, and the machine resources cannot be fully utilized. Leading to inefficiency.
- Since the VGG-16 in the indexing stage is used, the model has 16 layers, which is more time consumption than the ordinary texture-based feature extraction algorithm, sacrificing time for quality. Therefore, it is not as efficient as the indexing efficiency in the reference paper [18]. However, the extracted features better characterize the original image. Although the indexing efficiency of this solution is lower than that of the study [18], it is significantly higher than other studies in the same field. It is completely acceptable. Moreover, tasks of this stage can be run in batches or asynchronously at night without disrupting normal operations. There are many solutions to this problem.

D. Performance of the Retrieval Module

1) Data analysis: The retrieval phase of CBIR is the focus of this project and the core problem that needs to be solved. In order to verify its performance, experimental data from the CBIR retrieval phase is collected. Table X lists the experimental results.

TABLE X. COMPARISON OF THE AVERAGE COMPUTING TIME IN THE SEARCHING MODULE

Image dataset size	Parallel k-NN without cache	Parallel k-NN with cache	VCBIR
2011	180 s	120 s	0.27 s
5952	540 s	240 s	0.26 s
11963	1140 s	540 s	0.23 s
17953	1740 s	840 s	0.20 s
20000	1980 s	960 s	0.30 s

For ease of comparison, the chart was visualized, as depicted in Fig. 7. The asterisk line shows the time consumed by the method (VCBIR) for different amounts of data in the retrieval phase. The other two show the time consumed by the solutions in study [18] for different amounts of data in the retrieval phase.

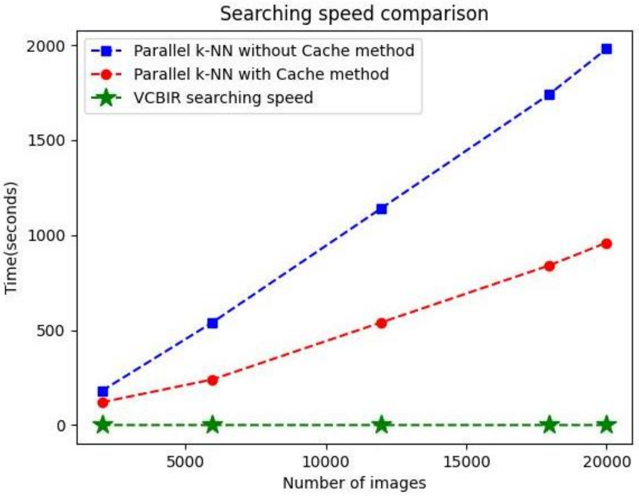


Fig. 7. Searching speed for different methods.

From the figure, it is evident that the performance of the VCBIR solution in the retrieval stage is very outstanding, far higher than the big data solution, and it can almost achieve a response in seconds and real-time return. Moreover, as the amount of data increases, the performance is almost not affected.

As shown in Table XI, the retrieval performance of existing literature was organized to facilitate a better comparison of the current research status in this field.

TABLE XI. PERFORMANCE COMPARISON OF DIFFERENT RETRIEVAL APPROACHES FOR 20,000 IMAGES

Approach	Description	Time (s)
Centralize method [20]	Sequential method on 1 node	25,000
DIRS method [20]	HBase system + MapReduce on 9 Hadoop nodes	15,000
Sakr et al. Method [22]	Parallel retrieval on 1 node Hadoop	1200
reference paper method [18][18] without cache	Parallel k-NN on 1 node Spark without cache	1980
reference paper method [18] with cache	Parallel k-NN on 1 node Spark with cache	960
VCBIR method	Milvus +VGG16 on 1 node	0.3

Again, to visualize the performance improvements, a histogram was generated. Fig. 8 depicts the results of research from various literature sources in this field. The VCBIR in the figure represents the time consumption of the solution in the retrieval field. It is evident that the performance improvement is substantial.

2) Result analysis: From the above results, it is evident that the performance improvement of CBIR applications based on vector databases in the field of retrieval is very huge, even several orders of magnitude higher than the performance of the previous solution. And as the number of images increases, the performance remains basically unchanged. It can be said that real-time response is achieved. Leading the way in performance research in this field.

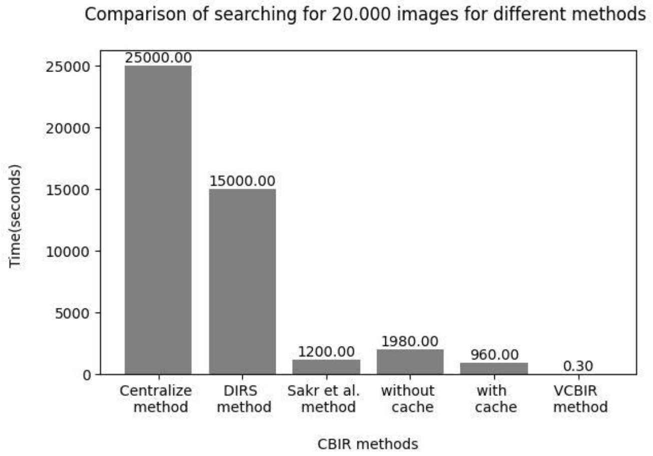


Fig. 8. Searching speed visualization for different methods.

E. Performance Comparison Between Different Datasets

Similarly, the indexing and retrieval times consumption between the ImageClef and ImageNet dataset also were recorded and compared. CBIR applications based on vector databases are comprehensively evaluated. Table XII and Table XIII are the performance comparisons of indexing and retrieval.

TABLE XII. INDEX PERFORMANCE ON TWO DATASETS

Dataset	Nbr of images	Using 1 node	Using 5 nodes	VDBMS
ImageClef	20,000	1200 s	720 s	3412 s
ImageNet	1,461,406	51,283 s	32,000 s	249,400 s

TABLE XIII. RETRIEVAL PERFORMANCE ON TWO DATASETS

Dataset	Sequential k-NN Maillo et al.2015	Parallel k-NN using 1 node	Parallel k-NN using 5 nodes	VDBMS
ImageClef	/	960 s	790 s	0.3 s
ImageNet	107,735 s	42,250 s	34,265 s	0.5 s

F. Performance of the Retrieval Accuracy

However, accuracy is also an important indicator of retrieval performance. In the vector database of this project, the IVF_FLAT index type is utilized, which considers both performance and accuracy. The retrieval function of this application was tested using the Corel-1k dataset, and its evaluation was based on precision and recall. The experimental results have been documented in Table V of Section IV. Fig. 9 shows the precision and recall results.

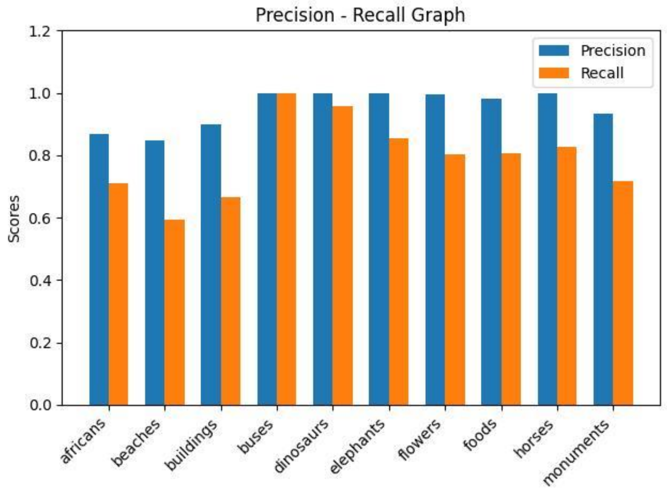


Fig. 9. The precision and recall in Corel-1k for different categories.

From the above figure, it is evident that the retrieval precision of this application is very high, even reaching 100% for some categories, and the lowest is still 60%. The recall is also very well. For the buses category, all images of these image categories can be retrieved based on the query image. It can be concluded that the retrieval accuracy of this application is relatively high and can meet the needs of most daily applications.

VI. Conclusion and Future Research

A. Introduction

Based on the design, a CBIR application utilizing transfer learning, and a vector database was developed. The performance of the application was evaluated by recording the time and accuracy of the indexing and retrieval stages. Experimental results indicated that the performance for the indexing stage did not outperform the big data solution, however for the retrieval stage, the performance of the VCBIR application is much better than that of the reference paper [18] and other solutions in this field. Furthermore since indexing is normally done offline, the performance is still acceptable. Whether it is the time consumption or accuracy of the retrieval stage, the performance is very outstanding, especially the retrieval time is reduced to about 1s, which is improved by several orders of magnitude and can achieve real-time response.

B. Contribution

The introduction of the vector database significantly improves the performance of the CBIR retrieval system, not only increasing the retrieval speed but also ensuring the accuracy. This new technology provides a powerful high-dimensional vector management tool for image retrieval, effectively solving the problems of low retrieval performance and accuracy, thereby bringing faster and more accurate query responses. This improvement improves the user experience and enhances the reliability and trust of the system, indicating that the integration of the vector database and the CBIR system has opened up new space for the development of image retrieval technology.

C. Future Work

- This project currently uses CPU for experiments but considering that GPU has higher performance and efficiency in the field of image processing, we plan to change the program to a GPU version in the future. In this way, we expect to significantly improve the performance and response speed of CBIR applications, allowing them to process large-scale image data faster and more accurately. This improvement will bring more opportunities and advantages to the project, provide users with a better image retrieval experience, and promote the development and application of CBIR technology in practical applications.

- Currently, the project is limited to implementation in a single-machine environment, but there are plans to study how to transform it into a distributed project in the future. This improvement aims to better utilize the advantages of distributed systems and improve the performance and efficiency of CBIR. Additionally, the distributed architecture can enhance the stability and reliability of the system, making it better able to cope with the challenges of large-scale data processing and high concurrent access. This step will bring broader development opportunities to the project, provide users with faster and more reliable image retrieval services, and promote the application and development of CBIR technology in distributed environments.

Acknowledgment

The authors wish to thank The Faculty of Computer Science and Information Technology, Universiti Putra Malaysia for supporting this research and funding this publication.

References

[1] Min, H., & Shuangyuan, Y. (2010, August). Overview of content-based image retrieval with high-level semantics. In 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE) (Vol. 6 , pp. V6-312). IEEE.

[2] Humeau-Heurtier, A. (2019). Texture feature extraction methods: A survey. IEEE access, 7, 8975-9000.

[3] Singh, P., Hrisheekesha, P. N., & Singh, V. K. (2021). CBIR-CNN: content-based image retrieval on celebrity data using deep convolution neural network. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(1), 257-272.

[4] Kapadia, M. R., & Paunwala, C. N. (2018, July). Improved CBIR system using Multilayer CNN. In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 840-845). IEEE.

[5] Traina, A. J., Brinis, S., Pedrosa, G. V., Avalhais, L. P., & Traina Jr, C. (2019). Querying on large and complex databases by content: Challenges on variety and veracity regarding real applications. Information Systems, 86, 10-27.

[6] Simran, A., Kumar, P.S., and Bachu, S. (March 2021). Content-based image retrieval using deep learning convolutional neural networks. IOP Conference Series: Materials Science and Engineering (Vol. 1084, No. 1, pp. 012026). IOP Publishing.

[7] Kumar, S., Singh, MK, and Mishra, MK (August 2022). Improving content-based image retrieval using deep learning models. In Journal of Physics: Conference Series (Vol. 2327, No. 1, pp. 012028). IOP Publishing.

[8] Sikandar, S., Mahum, R. and Alsalman, A. (2023). A novel hybrid approach for content-based image retrieval using feature fusion. Applied Sciences, 13(7), 4581.

[9] Stata, R., Bharat, K. and Maghoul, F. (2000). Term vector database: Fast access to web index terms. Computer Networks, 33(1-6), 247-255.

[10] Singla, S., Eldawy, A., Diao, T., Mukhopadhyay, A., and Scudiero, E. (2021, April). Experimental study of large raster and vector database systems. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 2243-2248). IEEE.

[11] Alsmadi, M. K. (2017). An efficient similarity measure for content based image retrieval using memetic algorithm. Egyptian journal of basic and applied sciences, 4(2), 112-122.

[12] Niu L., Xu Z., Zhao L., He D., Ji J., Yuan X., Xue M. (2023). Residual vector product quantization for approximate nearest neighbor search. Expert Systems with Applications, 120832.

[13] Chughtai, I. T., Naseer, A., Tamoor, M., Asif, S., Jabbar, M., & Shahid, R. (2023). Content-based image retrieval via transfer learning. Journal of Intelligent & Fuzzy Systems, 44(5), 8193-8218.

[14] Mohammed, M. A., Oraibi, Z. A., & Hussain, M. A. (2023, August). Content based Image Retrieval using Fine-tuned Deep Features with Transfer Learning. In 2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE) (pp. 108-113). IEEE.

[15] Gautam, G., & Khanna, A. (2024). Content Based Image Retrieval System Using CNN based Deep Learning Models. Procedia Computer Science, 235, 3131-3141.

[16] Sadiq, S. S. (2024). Improving CBIR Techniques with Deep Learning Approach: An Ensemble Method Using NASNetMobile, DenseNet121, and VGG12. Journal of Robotics and Control (JRC), 5(3), 863-874.

[17] Thanikachalam, R., Thavasimuthu, R., Arulkumar, V., Prabin, S. M., Saranya, N., & Devi, R. (2024). T2T-ViT: A Novel Semantic Image Mining Approach for Improving CBIR Using Vision Transformer.

[18] Mezzoudj, S., Behloul, A., Seghir, R., & Saadna, Y. (2021). A parallel content-based image retrieval system using spark and tachyon frameworks. Journal of King Saud University-Computer and Information Sciences, 33(2), 141-149.

[19] Traina, AJ, Brinis, S., Pedrosa, GV, Avalhais, LP and Traina Jr, C. (2019). Querying large complex databases by content: Diversity of practical applications and challenges in accuracy. Information Systems, 86, 10-27.

[20] Zhang, J., Liu, X., Luo, J., & Lang, B. (2010, December). Dirs: Distributed image retrieval system based on mapreduce. In 5th International Conference on Pervasive Computing and Applications (pp. 93-98). IEEE.

[21] Costantini, L., & Nicolussi, R. (2015). Performances evaluation of a novel Hadoop and Spark based system of image retrieval for huge collections. Advances in Multimedia, 2015, 11-11.

[22] Sakr, N. A., ELdesouky, A. I., & Arafat, H. (2016). An efficient fast-response content-based image retrieval framework for big data. Computers & Electrical Engineering, 54, 522-538.

Towards Accurate Detection of Diabetic Retinopathy Using Image Processing and Deep Learning

K. Kalindhu Navanjana De Silva[1], T.Sanduni Kumari Lanka Fernando[2], L.D.Lakshan Sandaruwan Jayasinghe[3], M.H.Dinuka Sandaruwan Jayalath[4], Dr. Kasun Karunanayake[5], B.A.P. Madhuwantha[6]
University of Colombo School of Computing, Serupita, Kaluthara, Sri Lanka[1]
University of Colombo School of Computing, Colombo 06, Sri Lanka[2]
University of Colombo School of Computing, Moronthuduwa, Sri Lanka[3]
University of Colombo School of Computing, Homagama, Colombo, Sri Lanka[4]
University of Colombo School of Computing, Colombo 07, Sri Lanka[5]
University of Colombo School of Computing, Colombo, Sri Lanka[6]

***Abstract*—Diabetic retinopathy (DR) is a critical complication of diabetes, characterized by pathological changes in retinal blood vessels. This paper presents an innovative software application designed for DR detection and staging using fundus images. The system generates comprehensive reports, facilitating treatment planning and improving patient outcomes. Our study aims to develop an affordable computer assisted analysis system for accurate DR assessment, leveraging publicly available fundus image datasets. Key objectives include identifying relevant features for DR staging, developing robust image processing algorithms for lesion detection, and implementing machine learning models for accurate diagnosis. The research employs various pre-processing techniques to enhance image quality and optimize feature extraction. Convolutional Neural Networks (CNNs) are utilized for stage classification, achieving an impressive accuracy of 93.45%. Lesion detection algorithms, including optic disk localization, blood vessel segmentation, and exudate identification, demonstrate promising results in accurately identifying DR-related abnormalities. The developed software product integrates these advancements, providing a user-friendly interface for efficient DR diagnosis and management. Evaluation results validate the effectiveness of the CNN model in stage classification and lesion detection, with high sensitivity and specificity. The study discusses the significance of image augmentation and hyperparameter tuning in improving model performance. Future research directions include enhancing the detection of microaneurysms and hemorrhages, incorporating higher resolution images, and standardizing evaluation methods for lesion detection algorithms. In conclusion, this research underscores the potential of technology in revolutionizing DR diagnosis and management. The developed software product offers a cost-effective solution for early DR detection, emphasizing the importance of accessible healthcare solutions. The findings contribute to advancing the field of DR analysis and inspire further innovation for improved patient care.**

Keywords*—Diabetic retinopathy; fundus images; computer-assisted analysis; deep learning; image processing; convolutional neural networks component

I. Introduction

The rapid advancement of diabetes poses a significant challenge to modern healthcare. Medical literature suggests that diabetes is associated with the emergence of serious long-term complications resulting in a variety of conditions, including cardiovascular disease, retinal complications, and retinopathy. According to a survey conducted by the International Diabetes Federation in 2015, around 410 million people worldwide suffer from diabetes. The disease also caused over 5 million deaths in the same year [1]. The disease is progressive, and its classification is based on the presence of different clinical abnormalities. Our research's main aim is to develop an affordable computer-assisted analysis system that accurately assesses the status of diabetic retinopathy (DR). By offering a cost-effective solution for DR analysis, we aim to prevent the progression of DR to its final stage and improve patient outcomes. This project highlights the potential benefits of technological innovation in healthcare and the importance of accessible solutions for DR analysis.

To achieve our aim, we have defined specific objectives. Firstly, we identified relevant features that can effectively determine each stage of DR, which were used to create a reliable diagnostic model. Secondly, we developed an image processing component capable of extracting these features from fundus images, enabling accurate detection of DR lesions. Thirdly, we build a robust and accurate machine learning model to diagnose the patient’s stage of DR based on the identified features. Finally, we developed a software product that can identify the stage of DR and detect the presence of lesions in each fundus image. These goals collectively helped develop a comprehensive computer-assisted analysis system for DR diagnosis and management.

This research project is delimited to the consideration of Diabetic Retinopathy (DR) exclusively in the human population, and other retinal diseases are excluded from the scope of the study. By excluding other retinal diseases, the study is designed to have a focused and specific approach in the development of a diagnostic tool for Diabetic Retinopathy (DR).

The research utilized publicly available datasets of fundus images, annotated with DR grades, for training and testing machine learning models. The datasets utilized in this study include Kaggle dataset for DR detection [2], Messidor-1 dataset [3], Messidor-2 dataset [3], and the Indian Diabetic Retinopathy Image Dataset (IDRiD) [4]. The research project will deliver a web application for DR diagnosis, accompanied by a user manual. A research paper documenting the process, methodology, and results will also be produced. These deliverables aim to provide a cost-effective and reliable tool for the diagnosis and management of DR, improving treatment programs for diabetic patients.

II. Related work

Diabetic Retinopathy is a complication of diabetes that is caused by pathological changes of the blood vessels which nourish the retina and lead the blood vessels of the retina to swell and to leak fluids and blood. In an advanced stage of diabetic retinopathy, it can lead to a loss of vision. Diabetic retinopathy is the most common cause of blindness in people aged 30–69 years [5]. According to Diabetes Control and Complication Trail (DCCT), DR is considered one of the four leading causes of blindness. A macular is a condition that affects vision and occurs when fluid leaks from blood vessels in the retina, leading to the formation of lesions. It is the leading cause of blindness among people with diabetes [6]. Retina regular screening is crucial for people with diabetes to detect and treat DR in its initial stages, as this can help prevent the risk of blindness [7]. The detection of DR involves identifying distinct types of lesions in a retina image. Microaneurysms (MA), Hemorrhages (HM), soft and hard Exudates (EX) [8-10].

K. Xu et al. [11] automatically classified the images of the Kaggle [12] dataset into normal images or DR images using CNN. The researchers selected 1000 images from the Kaggle dataset, and applied data augmentation and resizing to 224 x 224 x 3 before feeding the images to their CNN model. Their CNN architecture consisted of eight convolution layers, four max-pooling layers, and two fully connected layers. The SoftMax function was utilized in the final layer of the CNN for classification. This method achieved an accuracy of 94.5%. M. T. Esfahan et al. [13] used a known CNN, which is ResNet34 [14] in their study to classify DR images of the Kaggle dataset into normal or DR image. To enhance the quality of the images, they applied a set of image preprocessing techniques, such as Gaussian filtering, weighted addition, and normalization. The reported accuracy of their method was 85%, and the sensitivity was 86%.

The work by V. Gulshan et al. [15] introduced a method to detect DR and diabetic macular edema (DME) using the CNN model. They used Messidor-2 [3] and eyepacs1 datasets which contain 1748 images and 9963 images, respectively to test the model. They trained 10 CNNs with the pre-trained Inceptionv3 [16] architecture with a various number of images and the result was computed by a linear average function. However, the study did not focus on explicitly detecting non-DR or the five stages of DR. H. Pratt et al. [17] proposed a method based on a CNN to classify images from the Kaggle dataset into five DR stages. The researchers used a custom CNN architecture to classify 80,000 test images after performing color normalization and resizing them to 512x512 pixels. However, it is important to note that their CNN was unable to detect the lesions in the images, and they only evaluated their CNN on a single dataset. S. Dutta et al. [18] detected and classified DR images from the Kaggle dataset into five DR stages. The researchers evaluated the performance of three distinct types of neural networks - back propagation neural network (BNN), deep neural network (DNN), and convolutional neural network (CNN) - using a dataset of 2000 images.

Ege et al. [19] located exudates and cotton wool spots in 38 color images. The authors used a combination of template matching, region growing, and thresholding techniques to detect abnormalities. Wang et al. [20] addressed the same problem by using a minimum-distance discriminant classifier to identify the retinal bright lesions such as exudates and cotton wool spots. They represented color features in a spherical color space.

Our study identifies a gap in the existing tools available for generating comprehensive reports on the status of patients with diabetic retinopathy. To address this gap, we propose the development of a software product capable of storing patient medical history, detecting DR, and improving treatment programs. This software product would generate detailed reports summarizing the patient’s treatment program and store relevant information in a database. By providing doctors with a more complete understanding of the patient’s medical history and DR stage, the proposed software product would enable more effective treatment planning and ensure appropriate care. This research emphasizes the significance of technological innovation in healthcare and highlights the potential benefits of software products in enhancing patient outcomes.

III. Methodology

The rapid advancement of diabetes poses a significant challenge to modern healthcare. Medical literature suggests that diabetes is associated with the emergence of serious long-term complications resulting in a variety of conditions, including cardiovascular disease, retinal complications, and retinopathy. According to a survey conducted by the International.

A. *Pre-processing*

In this research study, a series of pre-processing steps were conducted to improve the quality and uniformity of retinal fundus images. These steps included capturing the retina by removing the background using adaptive thresholding and contour detection, minimizing lens flares by cutting off the outermost area of the image, and resizing the images for quicker processing. Additionally, normalization was applied to the images with respect to a reference image to match their histogram distribution. Image enhancement techniques were employed using contrast-limited adaptive histogram equalization (CLAHE) on both RGB and HSI channels. For RGB channels, the images were divided into their components, enhanced using CLAHE, and filtered to reduce noise. For HSI channels, the images were transformed, enhanced, converted back to RGB, and filtered. The HSI approach was favored over RGB due to minimal color shift. These pre-processing steps aimed to enhance lesion detection and were crucial in the subsequent stage classification experiments.

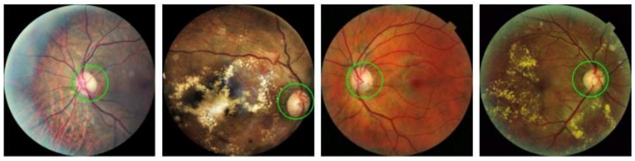
B. *Stage Classification*

In our research, the focus was on the classification of various stages of Diabetic Retinopathy (DR). Convolutional Neural Networks (CNNs) were chosen as the primary method due to their effectiveness in previous studies. Transfer learning was employed, utilizing pre-trained models available in the Keras deep learning framework to improve classification performance.

We experimented with various CNN models, including ResNet50, InceptionV3, VGG, and custom models, exploring different parameters and learning rate adjustment approaches. Initially, a fixed learning rate was used, but we later implemented a dynamic approach using the “ReduceLROnPlateau” method based on validation loss. To prevent overfitting, we employed early stopping using the “EarlyStopping” function. Evaluating the performance of different pre-trained models, such as ResNet50, InceptionV3, and VGG, showed no significant differences. Hence, hyperparameter tuning was performed on selected models, considering their previous performance and relevant literature. Additionally, the possibility of implementing an ensemble method was explored if multiple models demonstrated satisfactory performance.

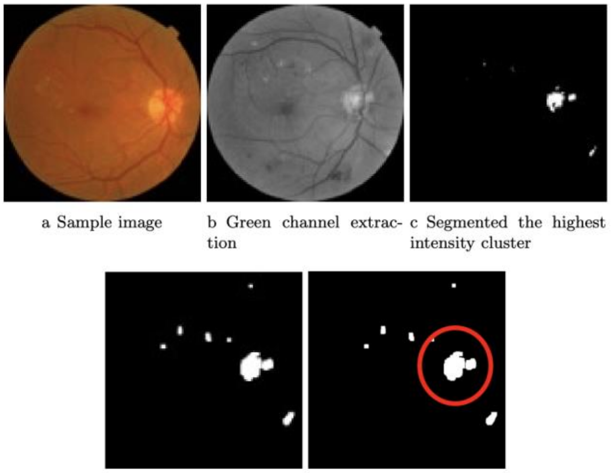
C. Stage Classification

1) Localization of optic disk: Two approaches were experimented with for the localization of the optic disk: the Kmeans clustering-based approach (see Fig. 2) and the intensity variation-based approach (see Fig. 1). In the K-means clustering-based approach, the algorithm is used to group pixels in the image based on their color intensities. The highest intensity cluster, representing the brightest pixels, is identified as the optic disk. The accuracy of this approach was evaluated on the Messidor and Kaggle datasets, achieving approximately 80.5% and 63.2% accuracy, respectively. However, issues such as lens flares and determining the optimal value of k for clustering can affect the accuracy of optic disk localization using this method.



a Image 01 b Image 02 c Image 03 d Image 04

Fig. 1. Results from localization of optic disk by intensity variation-based approach.



a Sample image b Green channel extraction c Segmented the highest intensity cluster

d After the dilation and erosion e Location of optic disk

Fig. 2. Optic disk localization using K-means.

The intensity variation-based approach identifies the optic disk by detecting rapid intensity variations caused by dark blood vessels and bright nerve fibers. The average intensity variance within a window is calculated, and the highest intensity point is marked as the optic disk center. This approach demonstrated high accuracy, correctly identifying the optic disk in approximately 98.13% of the images from the Messidor dataset and 93.6% of the images from the Kaggle dataset, including those with lens flares which is summarized in Table 1.

TABLE I. SUMMARY ON OPTIC DISK LOCALIZATION APPROACHES

Approach	**Messidor-2 Dataset (375 images)**	**Kaggle Dataset (250 images)**
K-means clustering based approach	80.5 %	63.2 %
Intensity variation-based approach	98.13 %	93.6 %

Both approaches have their advantages and limitations, and the intensity variation-based approach shows promise in improving the accuracy of optic disk detection.

2) Localization of blood vessels: To detect blood vessels in the retinal fundus images, a series of steps were followed. First, the green channel was extracted to enhance the contrast. Then, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to further enhance the image. Morphological operations involving opening and closing were performed using ellipsoidal structuring elements of varied sizes to remove small objects and smooth larger ones. The result of the CLAHE step was subtracted from the input image to isolate the blood vessels. A threshold was applied to binarize the blood vessel image. Finally, small contours were removed from the binarized image using an area-based criterion.

3) Detection of exudates: The detection of exudates from fundus images can be accomplished through various approaches. In the k-means clustering-based approach, the intensity values of the optic disk on the green channel image are set to zero, followed by dilation to remove lens flare. The highest intensity clusters are then identified using kmeans, with the brightest pixels representing the exudates. Experimentation with different k-values and normalization with a reference image are performed to enhance accuracy. However, false positive detections can occur if high-intensity pixel clusters are falsely identified as exudates. Alternatively, the edge detection approach involves extracting the green channel and applying the Canny edge detection algorithm to identify all edges. To isolate exudate edges, localization of blood vessels is first performed. The edges corresponding to blood vessels and the optic disk are then subtracted from the result. Contour analysis is employed to address the inclusion of tiny blood vessels during the extraction process. Although the edge detection approach is more reliable than the k-means approach, it may detect additional features like blood vessels, microaneurysms, and hemorrhages, making exudate identification more challenging. Consequently, another approach, namely Recursive Region Growing, is experimented with, which focuses on identifying regions with similar color or intensity likely to be exudates. The steps involved in this approach include extracting the green channel, applying the region growing and merging segmentation algorithm, producing a binary image using thresholding, localizing the optic disk, and subtracting it from the result.

4) Detection of detection of microaneurysms and hemorrhages: The process of detecting microaneurysms and hemorrhages using a k-means clustering approach in fundus images involves the following steps. First, the green channel is extracted as it provides better contrast for identifying red lesions. The images are then normalized with a reference image to ensure homogeneity and align with the selected value of k in the k-means algorithm. Next, k-means clustering is applied to identify clusters of similar pixel intensity values. The lowest intensity cluster is selected as it is expected to contain the microaneurysms and hemorrhages. The pixels in this cluster are thresholder and binarized to separate them from the rest of the image. Finally, the blood vessels are localized and removed from the segmented cluster.

IV. Implementation

The rapid advancement of diabetes poses a significant challenge to modern healthcare. Medical literature suggests that diabetes is associated with the emergence of serious long-term complications resulting in a variety of conditions, including cardiovascular disease, retinal complications, and retinopathy. According to a survey conducted by the International.

A. *CNN model for Stage Classification*

After conducting hyperparameter tuning on the pre-trained ResNet50v2 model, we found that setting the learning rate to 0.00001 and using three hidden layers, each with 230 neurons, resulted in the best performance. Based on these optimized hyperparameters, we finalized our ResNet50v2 model. The architecture of the model can be seen in Fig. 3.

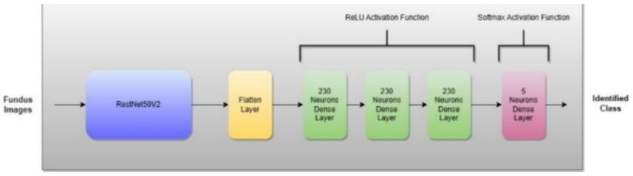


Fig. 3. CNN model for stage classification.

B. *System for Lesion Detections*

In our research, we focused on lesion detection in fundus images and successfully integrated the method into our application, Seer, using the OpenCV library. Seer not only accurately classifies various stages of Diabetic Retinopathy (DR) but also provides comprehensive reports on the DR class and identified lesions. By incorporating lesion detection into our analysis system, we have significantly enhanced the potential for early detection and effective management of DR, highlighting the importance of innovative solutions in healthcare. Through experimentation, we compared different approaches for localizing the optic disk and chose the intensity variation-based method for its superior accuracy. We also evaluated multiple techniques for exudate detection and determined that the recursive region-growing-based approach yielded the best performance. Additionally, we employed the k-means clustering approach to detect microaneurysms and hemorrhages and conducted experiments on localizing blood vessels. These advancements in lesion detection contribute to the overall effectiveness and reliability of Seer in diagnosing and managing DR.

C. *Implementation of Software product*

The implemented software product is a comprehensive and sophisticated solution designed to streamline the management of Diabetic Retinopathy (DR) in healthcare institutions. It offers a range of powerful functionalities that enhance various aspects of DR management. The software provides secure user authentication and authorization, ensuring that only authorized individuals can access the system. It utilizes advanced algorithms to predict the stage of DR based on fundus images, assisting in early detection and treatment planning. The software also offers intuitive visualization tools for DR lesions, enabling healthcare professionals to analyze and interpret the images with ease. It generates detailed reports that summarize the DR class and provide insights into the identified lesions, facilitating documentation and communication among healthcare teams. Furthermore, the software allows for efficient management of patient DR history, ensuring comprehensive records and tracking of their condition over time. User and institution management features are included to provide administrative control and customization options. Overall, this software product simplifies and streamlines DR management, empowering healthcare providers to deliver more effective care through its user-friendly interface and comprehensive range of functionalities.

1) Tools and Technologies: The proposed solution is being developed on the basis of the following tools:

- Application Framework: Python Flask
- Database: MySQL Database
- Additional tools: OpenCV

V. Evaluation

The rapid advancement of diabetes poses a significant challenge to modern healthcare. Medical literature suggests that diabetes is associated with the emergence of serious long-term complications resulting in a variety of conditions, including cardiovascular disease, retinal complications, and retinopathy. According to a survey conducted by the International.

A. *Evaluation of CNN Model for Stage Classification*

We evaluated our model against following dataset variations which derived from the Messidor-2 and Kaggle datasets to class balance and improve the quality.

- Dataset 03: Dataset generated by applying image augmentation to pre-processed images with minimized lens flares
- Dataset 04: Dataset generated by applying image augmentation to pre-processed images by applying CLAHE on RGB channels

- Dataset 05: Dataset generated by applying image augmentation for pre-processed images by applying CLAHE on HSI channels
- Dataset 06: Dataset generated by class balancing with approximately 600 images per class.

1) Evaluation of model with respect to Dataset 03: The results gained by testing the stage classification model on different variations of Dataset 03 and the confusion matrix and ROC curve are calculated to evaluate the performance of the model (see Fig. 4).

The accuracy and sensitivity of the test have been calculated for each class individually, and the results are presented below.

- Class 0: Accuracy = 0.806, Sensitivity = 0.979
- Class 1: Accuracy = 0.958, Sensitivity = 0.868
- Class 2: Accuracy = 0.982, Sensitivity = 0.948
- Class 3: Accuracy = 0.987, Sensitivity = 1.000
- Class 4: Accuracy = 1.000, Sensitivity = 0.962

The overall performance is listed below.

- Overall Accuracy: 0.935
- Overall Precision: 0.951
- Overall Specificity: 0.9

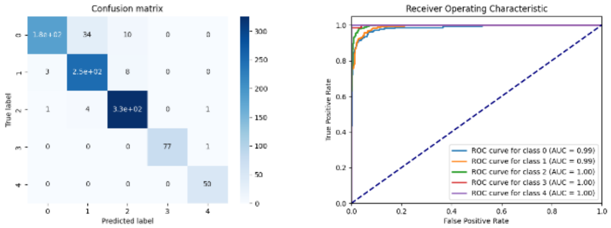


Fig. 4. Evaluation results with respect to Dataset 03.

2) Evaluation of model with respect to Dataset 04: The results gained by testing the stage classification model on different variations of Dataset 03 and the confusion matrix and ROC curve are calculated to evaluate the performance of the model (see Fig. 5).

The accuracy and sensitivity of the test have been calculated for each class individually, and the results are presented below.

- Class 0: Accuracy = 0.819, Sensitivity = 0.870
- Class 1: Accuracy = 0.916, Sensitivity = 0.839
- Class 2: Accuracy = 0.919, Sensitivity = 0.947
- Class 3: Accuracy = 0.987, Sensitivity = 0.987
- Class 4: Accuracy = 0.980, Sensitivity = 1.000

The overall performance is listed below.

- Overall Accuracy: 0.930
- Overall Precision: 0.928
- Overall Specificity: 0.953

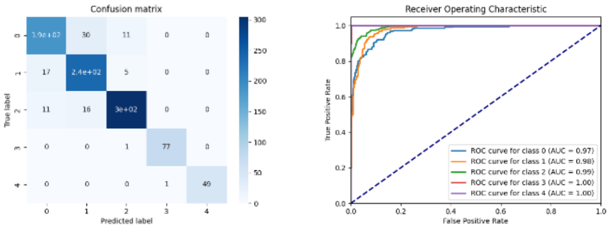


Fig. 5. Evaluation results with respect to Dataset 04.

3) Evaluation of model with respect to Dataset 05: The results gained by testing the stage classification model on different variations of Dataset 03 and the confusion matrix and ROC curve are calculated to evaluate the performance of the model (see Fig. 6).

The accuracy and sensitivity of the test have been calculated for each class individually, and the results are presented below.

- Class 0: Accuracy = 0.894, Sensitivity = 0.839
- Class 1: Accuracy = 0.912, Sensitivity = 0.881
- Class 2: Accuracy = 0.880, Sensitivity = 0.954
- Class 3: Accuracy = 0.974, Sensitivity = 0.950
- Class 4: Accuracy = 0.980, Sensitivity = 0.980

The overall performance is listed below.

- Overall Accuracy: 0.905
- Overall Precision: 0.921
- Overall Specificity: 0.950

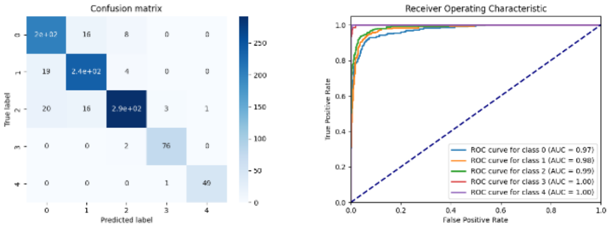


Fig. 6. Evaluation results with respect to Dataset 05.

4) Evaluation of model with respect to Dataset 63: The results gained by testing the stage classification model on different variations of Dataset 03 and the confusion matrix and ROC curve are calculated to evaluate the performance of the model (see Fig. 7).

The accuracy and sensitivity of the test have been calculated for each class individually, and the results are presented below.

- Class 0: Accuracy = 0.717, Sensitivity = 0.860
- Class 1: Accuracy = 0.825, Sensitivity = 0.780
- Class 2: Accuracy = 0.892, Sensitivity = 0.754
- Class 3: Accuracy = 0.908, Sensitivity = 0.732

- Class 4: Accuracy = 0.647, Sensitivity = 0.951

The overall performance is listed below.

- Overall Accuracy: 0.798
- Overall Precision: 0.815
- Overall Specificity: 0.939

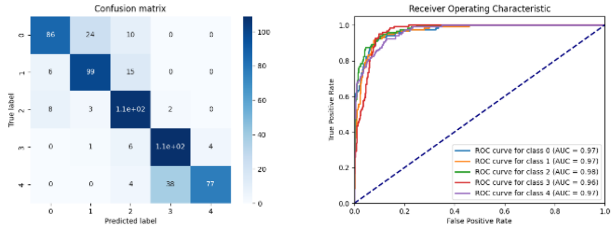


Fig. 7. Evaluation results with respect to Dataset 06.

B. *Evaluation of Lesion Detection*

To evaluate the effectiveness of a lesion detection method on a dataset of fundus images containing retinal lesions, an initial attempt was made to engage a medical professional to manually inspect the lesions. However, this approach proved unsuccessful due to several reasons. As an alternative, the Indian Diabetic Retinopathy Image Dataset (IDRiD) was discovered, which provided annotated ground truths of the lesions by a clinician. A 5x5 pixelation approach was adopted to facilitate the comparison between the detected lesions and the ground truth markings. The method's performance was evaluated by measuring true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and accuracy was calculated accordingly. Masks were generated using the IDRiD dataset, and these masks were compared to the ground truth markings to assess the performance of the lesion detection method.

1) Evaluation results by detection of exudates: This test was done on 27 images of the Indian Diabetic Retinopathy Image Dataset (IDRiD). Evaluation results by Detection of Exudates are given in Table 2.

TABLE II. EVALUATION RESULTS ON DETECTION OF EXUDATES

	Segments classified as Exudates	**Segments classified as normal**
Exudates	1005	790
Normal	1332	230728

TABLE III. EVALUATION RESULTS ON DETECTION OF MICROANEURYSMS AND HEMORRHAGES

	Segments classified as Microaneurysms and Hemorrhages	**Segments classified as normal**
Microaneurysms and Hemorrhages	546	2820
Normal	1173	344443

- Sensitivity: 0.560
- Specificity: 0.994
- Accuracy: 0.990

2) Evaluation results by detection of evaluation results by detection of microaneurysms and hemorrhages: This test was done on 27 images of the Indian Diabetic Retinopathy Image Dataset (IDRID). Evaluation results by Detection of Microaneurysms and Hemorrhages are given in Table 3.

- Sensitivity: 0.318
- Specificity: 0.992
- Accuracy: 0.989

C. *Discussion*

To enhance the features in fundus images, there are several effective pre-processing techniques that can be employed. A crucial step involves removing the background and centering the image, which helps to bring attention to the important regions while eliminating unnecessary distractions. Another valuable method is normalizing the fundus images using a reference image, ensuring consistent brightness and contrast for accurate comparisons. Enhancing the image on RGB channels and intensity channel in HSI format can improve overall image quality by employing CLAHE. By applying median blur, any remaining noise or artifacts can be further reduced, resulting in a cleaner image. Lastly, image augmentation, which involves techniques like random rotation, shifting the image horizontally and vertically, Shearing and Horizontal flips introduces variations to the training set, enabling the model to learn more robust features. By incorporating these pre-processing methods, the features in fundus images can be effectively enhanced, leading to more accurate and reliable analysis.

Based on the information provided, a ResNet50V2 model, which is a pre-trained CNN, was used for stage classification. This model achieved an impressive accuracy of 93.45% when evaluated on 900 augmented images from the Messidor-2 and Kaggle datasets. This suggests that a CNN-based model like ResNet50V2 can be highly effective for stage classification tasks in this context.

After conducting hyperparameter tuning on the pre-trained ResNet50V2 model for stage classification, the most suitable combination of hyperparameters was determined. The best results were obtained with a learning rate of 0.00001, 3 hidden layers, and 230 neurons in each hidden layer. This configuration maximized the model's learning capacity and allowed it to capture the complex patterns and relationships present in the stage classification task. Hyperparameters such as learning rate and network architecture can significantly impact the performance of a CNN model. By carefully selecting and optimizing these hyperparameters, the model's accuracy and ability to classify different stages can be enhanced.

After experimenting with three different approaches - K-means approach, Canny Edge detection approach, and Recursive Region Growing approach - it was determined that the Recursive Region Growing approach is the most suitable method for correctly identifying the locations of Exudates. Among the three approaches, Recursive Region Growing demonstrated superior performance in accurately detecting and delineating the regions of Exudates in the images. The Recursive

Region Growing method, based on a seed point and specific criteria for region expansion, effectively identifies and segments the Exudates areas with a higher level of precision.

To accurately identify the locations of Microaneurysms and Hemorrhages, the Kmeans clustering approach can be employed, along with the Elbow method to dynamically determine the optimal value of K for each fundus image. By finding the” elbow” point on the resulting plot, the optimal K value can be determined for accurate identification of the locations of Microaneurysms and Hemorrhages.

VI. Conclusion

In conclusion, we have successfully developed an affordable computer-assisted analysis system for accurate assessment of diabetic retinopathy status by utilizing Kaggle and Messidor-2 public datasets, along with an augmented dataset. We trained a ResNET50v2 model was trained using Bayesian Optimization for hyperparameter tuning. Our model achieved an accuracy of more than 0.9 for stage classification. Our application, Seer, utilizes this model to classify fundus images and provide a detailed report on DR class and lesions, which has the potential to significantly improve the detection and management of DR. Our research highlights the need for accessible solutions for DR analysis and the potential benefits of technological innovation in healthcare. We hope that our work inspires further development of cost-effective solutions for DR analysis, leading to improved patient outcomes.

VII. Future work

In future work, one of the key areas to focus on is the accurate detection and differentiation of Microaneurysms and Hemorrhages from blood vessels. This requires addressing the challenge of their similarity and developing new techniques that can effectively localize blood vessels and distinguish these lesions. Advanced image processing algorithms, feature extraction methods, and deep learning architectures could be explored to achieve this goal. Improving the accuracy of detecting hemorrhages and microaneurysms is another important aspect for future research. This can involve refining the existing algorithms to better identify red-colored and addressing limitations such as the algorithm’s inability to detect hemorrhages near blood vessels. Further studies can be conducted to enhance the performance of these detection methods.

For stage classification, incorporating higher resolution images in the model training process could be a potential improvement. Although it would require more computational resources, training the model on higher resolution images can enable the capture of finer details and subtle features associated with distinct stages. This can potentially lead to improved accuracy in stage classification. To facilitate standardized evaluation of lesions, the implementation of a standardized method is crucial. This would ensure consistency in the evaluation process across different studies and research efforts. A standardized method for evaluating lesions would allow for better comparison and benchmarking of different algorithms, models, or techniques, advancing the field.

These future works highlight the ongoing efforts and areas of focus in improving the detection and classification of Microaneurysms, Hemorrhages, and other relevant abnormalities in fundus images. By addressing these challenges and advancing the state-of-the-art techniques, we can achieve more accurate diagnoses and provide improved patient care in the field of Diabetic Retinopathy management.

Acknowledgment

We use this opportunity to express our greatest gratitude to Mr. A. Dinesh Madushanka Silva (Optometrist) who dedicated his valuable time to this research and helped us to enhance our domain knowledge in this area. We would also like to acknowledge with much appreciation to our parents and colleagues for their consistent support and their inclusion.

References

[1] Kapur, A., Harries, A. D., L¨onnroth, K., Wilson, P. & Sulistyowati, L. S. Diabetes and tuberculosis co-epidemic: the bali declaration. The Lancet Diabetes & Endocrinology 4, 8–10 (2016).

[2] Kaggle. Diabetic Retinopathy Detection https://www.kaggle.com/c/diabeticretinopathydetection

[3] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., … Klein, J.-C.(2014). FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. Image Analysis and Stereology, 33(3), 231–234. https://doi.org/10.5566/ias.1155

[4] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F., April 24, 2018, "Indian Diabetic Retinopathy Image Dataset (IDRiD)", IEEE Dataport, doi: https://dx.doi.org/10.21227/H25W98.

[5] Klein, R., Klein, B., Moss, S., Davis, M. & Demets, D. The Wisconsin epidemiologic study of diabetic retinopathy II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. Arch. Ophthalmol. 102, 520–526 (1984).

[6] Faust, O., Acharya U, R., Ng, E. Y., Ng, K. H., & Suri, J. S. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. Journal of medical systems, 36(1), 145–157 (2012). https://doi.org/10.1007/s10916-010-9454-7

[7] Harper, C. A. & Keeffe, J. E. Diabetic retinopathy management guidelines. Expert review of ophthalmology 7, 417–439 (2012).

[8] Taylor, R. & Batey, D. Handbook of Retinal Screening in Diabetes: Diagnosis and Management 2nd (John Wiley & Sons, Ltd Wiley-Blackwell, 2012).

[9] Group, E. R. Grading diabetic retinopathy from stereoscopic color fundus photographs– an extension of the modified Airlie House classification. Ophthalmology 98, 786–806 (1991).

[10] Scanlon, P. H., Wilkinson, C. P., Aldington, S. J. & Matthews, D. R. A Practical Manual of Diabetic Retinopathy Management (Wiley-Blackwell, 2009).

[11] Xu, K., Feng, D. & Mi, H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. Molecules 22, 2054 (2017).

[12] Erd´elyi, A., Magnus, W., Oberhettinger, F. Tricomi, F. G. Tables of Integral Transforms Vol. I (New York, NY: McGraw-Hill Book Company, Inc., 1954).

[13] Esfahani, M. T., Ghaderi, M. & Kafiyeh, R. Classification of diabetic and normal fundus images using new deep learning method. Leonardo Electronic Journal of Practices and Technologies 17, 233–248 (2018).

[14] Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, inception-ResNet and the impact of residual connections on learning in The thirty-first AAAI conference on artificial intelligence (2016), 4278–4284.

[15] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in

Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402-2410. doi: 10.1001/jama.2016.17216. PMID: 27898976.

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision in IEEE Conference on Computer Vision and Pattern Recognition (2016), 2818–2826.

[17] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P. & Zheng, Y. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science 90, 200–205 (2016).

[18] Dutta, S., Manideep, B., Basha, S. M., Caytiles, R. D. & Iyengar, S. S. N. Classification of diabetic retinopathy images by using deep learning models. International Journal of Grid and Distributed Computing 11, 99–106 (2018).

[19] Ege, B., Larsen, O. & Hejlesen, O. Detection of abnormalities in retinal images using digital image analysis in Proc. 11th Scand. Conf. Image Process. (1999), 833–840.

[20] Wang, H., Hsu, W., Goh, K. & Lee, M. L. An effective approach to detect lesions in retinal images in Proc. IEEE Conf. Comput. Vis. Pattern Recogn. 2 (Hilton Head Island, SC, 2000), 181–187

Machine Learning Approaches for Predicting Occupancy Patterns and its Influence on Indoor Air Quality in Office Environments

Amir Hamzah Mohd Shaberi[1], Sumayyah Dzulkifly[2]*, Wang Shir Li[3], Yona Falinie A. Gaus[4]
Faculty of Computing and Meta-Technology, Sultan Idris Education University, Perak, Malaysia[1, 2, 3]
Centre of Embedded Education Green Technology (EduGreen@UPSI), Sultan Idris Education University, Perak, Malaysia[2]
Department of Computer Science, Durham University, Durham, United Kingdom[4]

***Abstract*—It is normal for the modern population to spend 12 hours or more daily indoors where the level of comfort can be moderated. Yet, indoor occupants are similarly exposed to various air pollutants just as outdoors. Indoor air pollution could be detrimental toward the occupant's health noted by the United Nation Environment Programme (UNEP) in the Pollution Action Note, published on 7th of September 2021. According to the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standards, occupancy patterns could influence indoor air quality. Hence, this paper investigates the utilisation of machine learning algorithms in predicting occupancy patterns against indoor air quality (IAQ) variables such as humidity, temperature, light, and carbon dioxide (CO_2). This study compares the performance of selected machine learning approaches, namely deep learning (LSTM, CNN), regression (ANN) and (SVR) models. In addition, it explores the diverse range of evaluation metrics utilized to evaluate the performance of machine learning in the specific context of Mean Squared Error (MSE) and Mean Absolute Error (MAE). In the training phase, the SVR model achieved the lowest MAE of 0.0826 and MSE of 0.0280 as compared to the other algorithms. The ANN model demonstrated slightly better generalization capabilities in the testing phase, while the LSTM model demonstrated robust performance in the test phase. Overall, the results highlighted the significant impact of occupancy behaviour on Indoor Air Quality (IAQ) variables and underscored the importance of advanced modelling techniques in IAQ monitoring and management, emphasizing the need for tailored approaches to address the complex relationship between occupancy patterns and IAQ variables.**

Keywords*—Indoor air quality; occupancy patterns; machine learning; deep learning; regression models; Mean Squared Error; Mean Absolute Error; IAQ monitoring; IAQ management

I. INTRODUCTION

Indoor air quality (IAQ) is a critical component in maintaining occupants' health, impacting the well-being of both the humans and the interior ecosystems. Prolonged exposure to harmful substances in the air could lead to persistent discomfort, severe illnesses, and could lead to respiratory attributed deaths annually [1]. Compromised air quality encompasses of various factors, which includes, but not limited to the concentration of pollutants, such as the cleaning supplies and the building materials [2], the indoor humidity levels [3], the temperature control [4] and the adequacy of the ventilation systems [5]. Unhealthy working space for indoor occupants due to poor IAQ can lead to a range of health issues, such as respiratory problems [6-7], asthma [8] and fatigue [9], which can impact their productivity and overall well-being.

The ability to monitor and control IAQ is essential for building managers to identify potential issues and implement corrective measures [9]. Yet, to some, comprehensive installation of indoor sensing system could incur a hefty cost which could discourage the motivation to maintain healthy indoor air quality. The variation of indoor air quality with respect to the occupancy pattern could lead to complex data analysis which requires the use of machine learning to study their linear or non-linear relationships. In recent years, the integration of machine learning techniques has emerged as a promising avenue for predicting and optimizing air quality, in general. Additionally, as societies grapple with the consequences of air pollution, understanding the effectiveness of machine learning models in this context is imperative [10].

As indoor air pollution (IAP) poses a major risk to human health and is responsible for millions of deaths annually, preserving a good IAQ is significant for the health sector [11]. The intersection studies of IAQ and machine learning have garnered significant attention in recent years as researchers explore innovative ways to leverage machine learning techniques to enhance indoor air quality monitoring and management. For example, authors in the study [11], state that machine learning technologies are highly capable of providing real-time indoor air quality monitoring, which is essential for determining and managing indoor air pollutants. In addition, [12] also states that sets of algorithms are utilized to extract and filter general principles from massive datasets, allowing for the automated learning of user preferences in relation to the IAQ.

Despite of this, there is a gap in current literatures, specifically in terms of the employment of machine learning techniques to predict occupancy patterns within the context of indoor air quality management. While previous studies have highlighted the potential of machine learning for real-time monitoring and general data analysis, few have focused on its application to predict occupancy patterns, which can be crucial for understanding indoor air quality dynamics. Therefore, this study aims to fill this gap by investigating the effectiveness of machine learning approaches in predicting occupancy patterns based on variables such as humidity, temperature, light, and carbon dioxide (CO_2). In addition, this study will determine which algorithms performed better and using selected evaluation metrics, namely, the Mean Squared Error (MSE) and Mean Absolute Error (MAE).

The paper is organized into five main sections, namely the Introduction in Section I**,** where the overall background of the research is elaborated; the Literature Review in Section II, which provides extensive reviews on deep learning, regression, classification model, and the relationship between IAQ and occupancy behaviour; the Methodology in Section III, which elaborates on the data acquisition, data training, data testing and selected machine learning algorithms' evaluation approaches; the Results and Analysis in Section IV, which provides in depth evaluations and discussions on the overall analysis of this study; and Conclusion in Section V, which sums up the investigation and highlights the key topics of this study.

II. Literature Review

A. Comparison between Deep Learning, Regression and Classification Models.

IAQ monitoring is a critical aspect of ensuring healthy and comfortable indoor environments, particularly in settings such as homes, offices, and schools [13]. With the increasing prevalence of indoor air pollutants and their impact on human health, there is a growing need for advanced predictive models to accurately monitor and forecast occupancy patterns because occupancy patterns are closely related to the variables of IAQ. In this comparison models, this study delves into the field of deep learning, regression, and classification models, exploring their method and capabilities in the context of occupancy patterns prediction.

1) Deep learning model. Deep learning, a subset of machine learning, has emerged as a powerful tool for processing complex data and extracting meaningful patterns [14]. By leveraging deep neural networks, deep learning models can effectively analyze large volumes of IAQ data, including particulate matter (PM) [15–18] volatile organic compounds (VOCs) [17], CO_2 [17], and sulfur dioxide (SO_2) [19, 20], to provide real-time insights into IAQ levels. These models can learn intricate relationships within the data, enabling them to make accurate predictions and identify potential IAQ issues before they escalate. However, for specific occupancy patterns [21] mentioned that predictions for occupancy were carried out using various deep learning architectures, such as Deep Neural Network (DNN), Long Short-Term Memory (LSTM), Bi-directional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and Bi-directional GRU (Bi-GRU), in different settings like an office, library and lecture room. The results demonstrated that the feature selection algorithm proposed performed better than a commonly used one, leading to higher model performance while requiring fewer sensors.

Other than that, from 11 studies [15–20, 22–26] employing various architectures such as long short-term memory (LSTM), Convolutional Neural Network (CNN), unique combinations like Combined Self-Attention (SA) mechanism, Empirical Mode Decomposition (EMD) algorithm and LSTM network (SA–EMD–LSTM) with Ensemble Empirical Mode Decomposition-Sparrow Search Algorithm (EEMD-SSA-LSTM). The performance metrics for these studies include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Percentage Error (MAPE), Index of Agreement (IA), Theil Inequality Coefficient (TIC), coefficient of determination (R2) and Absolute Average Deviation. This diversity in metrics underscores the complexity and multifaceted nature of evaluating deep learning models in the context of air quality prediction, considering both spatial and temporal aspects.

2) Regression model. Regression models, on the other hand, are well-suited for predicting continuous IAQ variables, such as PM concentration [27], temperature [28] or occupancy pattern [29] based on historical data and other relevant variables. By fitting a regression model to IAQ data, regression model can create a mathematical equation that describes the relationship between the input variables and the output [30], allowing to make predictions with a high degree of accuracy.

For example, six studies [27, 28, 31–34] employ diverse techniques such as Artificial Neural Network (ANN), Extra Trees Regressor, Wavelet Artificial Neural Network, and Support Vector Regression (SVR). Evaluation metrics for these models include MAE, RMSE, SMAPE, MAPE, and Pearson correlation coefficient (R). The inclusion of regression models adds a valuable dimension to the literature, as these models provide insights into continuous air quality prediction. In addition, the study in [29] also states linear and logistic regression models were created using the variables to forecast occupant activity schedules and the probability of occupant presence.

3) Classification models. Classification models are particularly useful for identifying discrete IAQ states, such as air quality categories (e.g., good, moderate, poor) or the presence of specific pollutants [35]. These models classify IAQ data into different classes based on their characteristics, enabling us to categorize IAQ levels and take appropriate actions to improve indoor air quality. Previous study primarily leverages random forest [31, 36–38] and decision tree [32] while employing a range of evaluation metrics such as specificity, precision, recall, accuracy, F1 score, area under the ROC Curve (AUC), and sensitivity. This reflects a robust assessment strategy to measure the performance of these models in predicting air quality.

While it is possible for a classification model to estimate the occupancy patterns, there are some important considerations and limitations. In a typical classification problem, the goal is to predict the class label of an observation based on its features. So, this study will not include classification models, as the focus is on utilizing deep learning and regression models, which are deemed more important for the study.

B. Relationship between IAQ and Occupancy Behavior

The relationship between IAQ and occupancy behavior is complex and symbiotic. The way in which occupants use spaces, their activities, and their preferences all have an impact on the IAQ, which includes temperature, humidity, and pollutants [39]. Occupancy behavior plays a crucial role in determining pollutant emissions, ventilation needs and desired comfort levels. For instance, spaces with more occupants engaged in activities that generate pollutants may necessitate higher ventilation rates [40]. Managing IAQ effectively involves comprehending these interconnections and devising strategies that could harmonise occupants' behavior while maintaining optimal indoor air quality.

To encourage a healthier indoor environment, a comprehensive strategy that considers both IAQ variables and occupancy behavior is imperative. According to [41], architectural designs should cater to varied activities and occupant densities, optimizing factors like temperature control to curb discomfort and air stagnation. Incorporating intelligent building technologies to monitor and regulate IAQ variables in real time based on occupancy patterns can further amplify the interplay between IAQ and occupant behavior [42]. Ultimately, cultivating an atmosphere where occupants are conscientious about their actions and where IAQ commands shared attention can culminate in an enhanced IAQ and overall well-being.

Given the intricate relationship between IAQ and occupancy behavior, this study aims to predict occupancy patterns in relation to selected IAQ variables. Understanding how occupants utilize spaces and engage in activities that affect IAQ variables like temperature, humidity, and pollutant levels is crucial for effective IAQ management. By leveraging the predictive modelling, this study seeks to develop insights into how occupancy patterns influence the IAQ dynamics, ultimately contributing to strategies that enhance the indoor environmental quality and occupant well-being.

III. Methodology

In the context of predicting occupancy patterns, the model training for deep learning and regression models can be similar, as both approaches aim to predict occupancy based on input variables such as temperature, humidity, light and CO_2 levels. The phase would involve several steps as shown in the Fig. 1 below. However, the specific implementation details and the model architectures would differ between the deep learning and regression, depending on the algorithms used. To choose a better algorithm that suited the aim of this study, data collection, data pre-processing, model training and testing, and model evaluation will be conducted. Fig. 1 below shows the phases involved in this study to predict occupancy pattern in relation to the IAQ variables.

A. Phase 1 (Data Collection)

For the data collection stage, the historical data for the variables such as temperature, humidity, light, CO_2, and occupancy will be gathered. It is crucial to ensure that the data is representative and covers a wide range of values to capture the variability of the IAQ. This comprehensive dataset will serve as the foundation for training and evaluating the deep learning and regression models for predicting occupancy patterns. This study addresses key variables present in office room environment, with a focus on identifying an acceptable range for these variables that may pose health risks to occupants if exceeded.

B. Phase 2 (Data Pre-Processing)

In the data pre-processing phase, the input features were normalized to ensure data consistency. This normalization step is crucial for deep learning and regression models, as it helps to prevent features with larger scales from dominating the training process. Additionally, the data were split into the training (70%) and the testing sets (30%). The training set was used to train the models, while the testing set was used to evaluate their performance. This split is essential to assess how well the generalization of the models were towards unseen data and to prevent overfitting. Overall, these preprocessing steps help to ensure the models can effectively learn from the data.

C. Phase 3 (Model Training and Testing)

This stage implements a custom callback, ‘MetricsCallback’, to calculate and print MAE and MSE during the training and testing phases of LSTM, CNN, and ANN models. This callback was used to monitor the model's performance on the training and the validation sets. However, the SVR model is not a neural network and therefore does not use the same training process. For such case, the MAE and the MSE were calculated manually after the training completed.

The training process involves splitting the dataset into the training and the testing sets, scaling the features, and reshaping the data for the LSTM and CNN models. The models are then compiled and trained using the fit method, with the callback used to print the loss, the MAE, and the MSE at the end of each epoch. The detailed machine learning parameters used for each model are as follows:

LSTM Model

- Sequential model with an LSTM layer (64 units) and a dense output layer.
- Compiled with Mean Squared Error (MSE) loss and Adam optimizer.
- Trained for 50 epochs with a batch size of 32.
- Validation data is specified for monitoring performance during training.
- Callback is used to monitor and print MAE and MSE during training and validation.

CNN Model

- Sequential model with a 2D convolutional layer (32 filters, 2x2 kernel size, ReLU activation), a flattening layer, and a dense output layer.
- Compiled and trained similarly to the LSTM model.

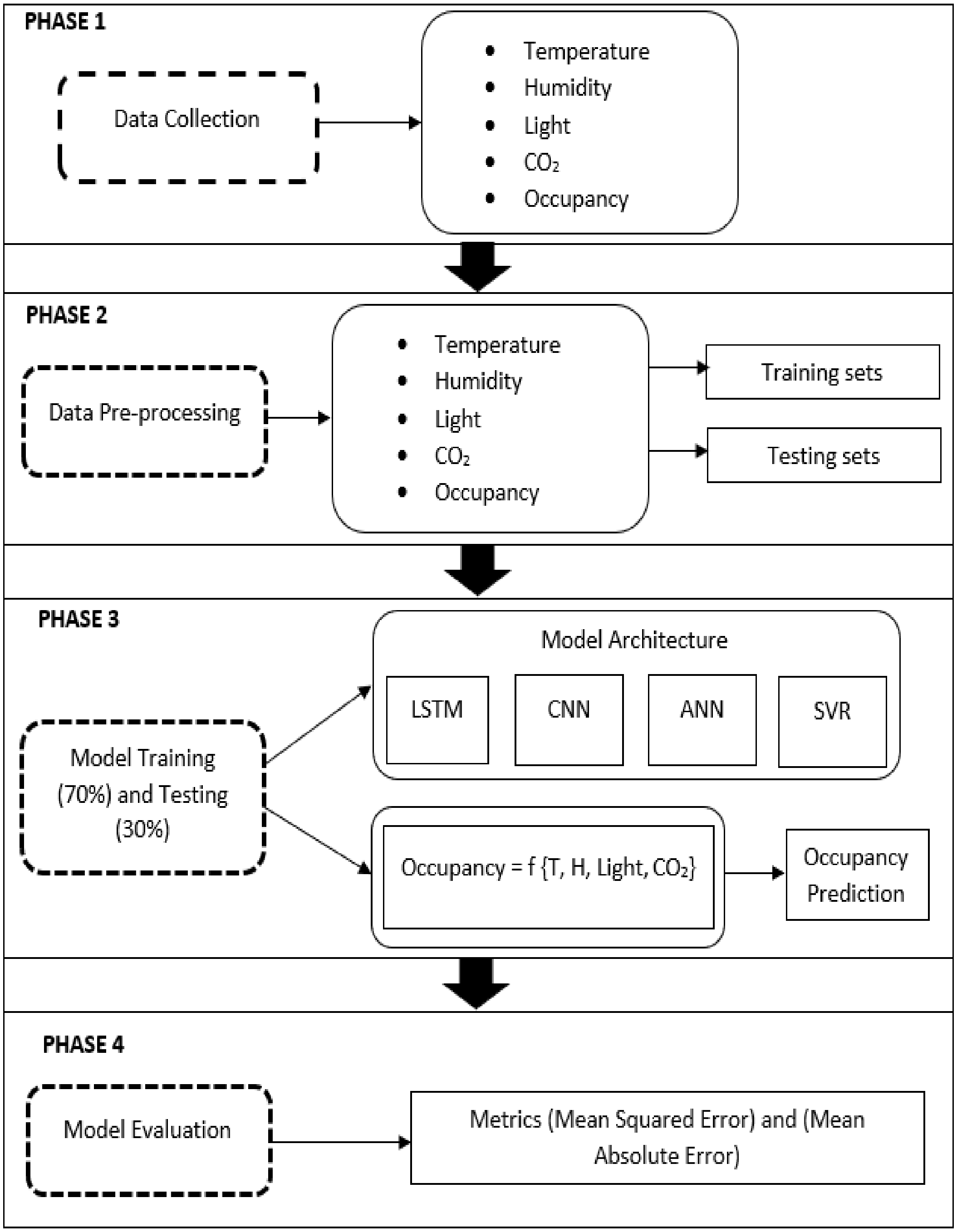


Fig. 1. Phases of predicting occupancy pattern.

ANN Model (Regression)

- Sequential model with two dense layers (64 units, ReLU activation) and a dense output layer.
- Compiled and trained similarly to the LSTM and CNN models.

SVR Model

- SVR model trained separately using the SVR class from scikit-learn.
- Does not use the custom callback as it does not follow the same training process as neural networks.

Once the models completed the training and testing stage, they were evaluated using the MAE and the MSE to observe the difference in term of the performance. This approach provides a comprehensive analysis of the models' performance, allowing a comparison for their ability to predict occupancy pattern based on the given IAQ variables.

D. Phase 4 (Model Evaluation)

For model evaluation, the trained model is assessed using the testing dataset to gauge its performance in predicting occupancy. Metrics such as the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) are used. The MSE provides a measure of the average squared difference between the predicted occupancy values and the actual values, offering insight into the model's overall accuracy.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (yi - \bar{y}i)^2 \quad (1)$$

where:

- n is the number of observations.
- yi is the actual value of the target variable for the i-th observation.
- $\bar{y}i$ is the predicted value of the target variable for the i-th observation.

On the other hand, the MAE measures the average absolute difference between the predicted and the actual values, providing an indication of the model's precision. These metrics collectively offer a comprehensive assessment of the model's performance in predicting occupancy values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |yi - \bar{y}i| \quad (2)$$

where:

- n is the number of observations.
- yi is the actual value of the target variable for the i-th observation.
- $\bar{y}i$ is the predicted value of the target variable for the i-th observation.

IV. Results and Discussion

The discussion for this study focused on comparisons between the prediction results of various machine learning algorithms for occupancy patterns based on IAQ variables including an analysis of their time and memory complexities. In addition, this study also explores the impact of occupancy behavior on IAQ variables. Key findings included the effectiveness of certain algorithms in predicting occupancy patterns and how changes in IAQ variables were influenced by occupancy.

A. Prediction Results for Training and Testing Data Across Different Machine Learning Models

Comparing the training and the testing results is essential in machine learning to assess how well a model performs in analyzing the data. Study by [43] mentioned that during the training, a model learns to map input features to output labels using the provided data. However, this process can lead to overfitting, where the model memorizes the training data instead of learning the underlying patterns. By evaluating the models' performance on a separate testing dataset, one can gauge its ability to generalize. Discrepancies between training and testing results indicate potential overfitting, highlighting the need for adjustments such as hyperparameter tuning [43]. Additionally, comparing results helps in model selection, as the best-performing model on testing data is typically chosen for deployment.

Tables 1 and 2 compare the performance of four different models (LSTM, CNN, ANN, SVR) in the training and the testing phases using the MAE and the MSE metrics. In the training phase (see Table 1), the SVR model achieved the lowest MAE, namely 0.0826 and MSE of 0.0280, indicating better performance in predicting occupancy pattern based on the input features (CO_2, Light, Temperature, Humidity) compared to the other models. The ANN model also performed well, with a MAE of 0.0940 and MSE of 0.0353, followed by the CNN and LSTM models.

In the testing phase (see Table 2), the ANN model demonstrated the best performance, achieving the lowest MAE 0.0834 and MSE 0.0364. This indicates that the ANN model was more accurate in predicting occupancy pattern on unseen data compared to the other models. The CNN model also performed well in the testing phase, with an MAE of 0.0866 and MSE of 0.0385, followed by the LSTM and SVR models.

TABLE. I. Comparison of MAE and MSE for Training Data

Model	MAE	MSE
LSTM	0.1153	0.0474
CNN	0.1024	0.0419
ANN	0.0940	0.0353
SVR	0.0826	0.0280

TABLE. II. Comparison of MAE and MSE for Testing Data

Model	MAE	MSE
LSTM	0.0977	0.0420
CNN	0.0866	0.0385
ANN	0.0834	0.0364
SVR	0.0968	0.0411

Overall, the ANN and CNN models demonstrated robust performance in predicting the occupancy patterns, with the ANN model showing slightly better generalization capabilities in the testing phase.

B. Complexity Analysis: Time and Memory

Table 3 below provides the time and memory complexities for each of the models used in this study. For the time complexity, it is measured in seconds and represents the duration taken by each model to complete the training process. The LSTM model took the longest time at 0.2044 seconds, followed closely by the CNN model at 0.1921 seconds. The ANN model was next, at 0.1342 seconds, and the SVR model was the fastest at only 0.0156 seconds. These time complexities give an indication of how efficiently each model can process and learn from the training data.

In terms of memory complexity, measured in MiB (Mebibytes), it represents the peak memory usage during the training process. Interestingly, all three models (CNN, ANN, and SVR) exhibited very similar memory usage, ranging from approximately 723 MiB to 738 MiB. This suggests that these models require a similar amount of memory to store and process the data during training. The LSTM model, on the other hand, showed a slightly higher memory usage of 738.836 MiB, indicating that it may require a bit more memory compared to the other models.

Overall, these complexities provide insights into the efficiency and resource requirements of each model, which can be valuable for selecting the most suitable model for a given application based on computational resources and time constraints.

TABLE. III. TIME AND MEMORY COMPLEXITY COMPARISON FOR EACH MODEL

Model	Time Complexity (s)	Memory Complexity (MiB)
LSTM	0.2044	738.836
CNN	0.1921	723.012
ANN	0.1342	723.246
SVR	0.0156	723.086

C. Exploring the Impact of Occupancy Behavior on Indoor Air Quality Variables

The comparison of machine learning algorithms for predicting occupancy patterns based on IAQ variables provided valuable insights into their effectiveness and efficiency. The study focused on four models: LSTM, CNN, ANN, and SVR, evaluating their performance using MAE and MSE metrics in both training and testing phases.

In the training phase, the SVR model exhibited the lowest MAE and MSE, indicating superior performance in predicting occupancy patterns. The ANN model also performed well, followed by the CNN and LSTM models. However, in the testing phase, the ANN model demonstrated the best performance, achieving the lowest MAE and MSE. This suggests that the ANN model was more accurate in predicting occupancy patterns on unseen data, highlighting its superior generalization capabilities compared to the other models.

Furthermore, the study analyzed the time and memory complexities of each model. The LSTM and CNN models exhibited longer training times compared to the ANN and SVR models. In terms of memory complexity, all models (CNN, ANN, and SVR) showed similar memory usage, while the LSTM model required slightly more memory. These complexities provide insights into the computational efficiency and resource requirements of each model, which are crucial considerations for real-world applications.

Overall, these findings underscore the importance of selecting the right machine learning model for predicting occupancy patterns based on IAQ variables. The study's results can guide future research in optimizing IAQ monitoring systems and prediction algorithms, ultimately leading to improved indoor air quality and occupant comfort.

V. CONCLUSION AND FUTURE ENHANCEMENT

This study demonstrates the importance of considering occupancy behavior in predicting IAQ patterns. The results highlight the effectiveness of machine learning algorithms, particularly ANN and CNN, in accurately predicting occupancy patterns based on IAQ variables. ANN emerged as the most accurate algorithm, followed by CNN, LSTM and SVR. These findings underscore the significance of advanced modeling techniques in IAQ monitoring and management, emphasizing the need for tailored approaches to address the complex relationship between occupancy behavior and IAQ variables. In addition, integrating machine learning models into IAQ management strategies can lead to improved indoor environmental quality and occupant well-being.

This study also highlights the promising potential of relationship between IAQ variables and occupancy pattern. This study uses machine learning algorithms in predicting occupancy patterns within office room environments. This research aims to address this deficiency by examining how machine learning methods can predict occupancy patterns using factors like humidity, temperature, light, and CO_2 levels. If the IAQ variables increase, it aims to determine if there's a corresponding increase in occupancy in room environments. It will also identify the most effective algorithms for this task and explore the various evaluation metrics, particularly focusing on MSE and MAE.

Despite the successes observed, it's essential to acknowledge the limitations inherent in this study, particularly the reliance on data collected from external sources rather than proprietary datasets specific to the office building under investigation. The occupancy value is binary, either 0 or 1, which means the prediction can only predict these two states. Secondly, the data size and variables may not be sufficient to provide highly accurate results. Lastly, in this study, there is no available measurement of the room space, which is imperative when studying the impact of IAQ against occupancy pattern.

Future enhancements could involve obtaining proprietary datasets specific to the office building under investigation. This can provide more accurate and relevant data for analysis. Then, increase the dataset size by collecting more data over a longer period. Additionally, consideration could be made to add more variables that may impact indoor air quality and occupancy patterns. This study must consider using indirect methods to estimate occupancy, such as analyzing patterns of other variables that are correlated with occupancy, like motion sensors. Lastly, to calculate the volume of the space, typically need the dimensions of the room (length, width, and height). If there have access to the physical space, it can measure these dimensions directly.

ACKNOWLEDGMENT

This research is funded by the Malaysian Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS), grant number 2022-0092-104-02 (FRGS/1/2022/TK08/UPSI/02/1).

REFERENCES

[1] 1. Almetwally, A.A., Bin-Jumah, M., Allam, A.A.: Ambient air pollution and its influence on human health and welfare: an overview, (2020). https://doi.org/10.1007/s11356-020-09042-2.

[2] Wall, D., McCullagh, P., Cleland, I., Bond, R.: Development of an Internet of Things solution to monitor and analyse indoor air quality. Internet of Things (Netherlands). 14, (2021). https://doi.org/10.1016/j.iot.2021.100392.

[3] Majdi, A., Alrubaie, A.J., Al-Wardy, A.H., Baili, J., Panchal, H.: A novel method for Indoor Air Quality Control of Smart Homes using a Machine learning model. Advances in Engineering Software. 173, (2022). https://doi.org/10.1016/j.advengsoft.2022.103253.

[4] Sulzer, M., Christen, A., Matzarakis, A.: Predicting indoor air temperature and thermal comfort in occupational settings using weather forecasts, indoor sensors, and artificial neural networks. Build Environ. 234, (2023). https://doi.org/10.1016/j.buildenv.2023.110077.

[5] Taheri, S., Hosseini, P., Razban, A.: Model predictive control of heating, ventilation, and air conditioning (HVAC) systems: A state-of-the-art review. Journal of Building Engineering. 60, (2022). https://doi.org/10.1016/j.jobe.2022.105067.

[6] Jain, N., Burman, E., Stamp, S., Shrubsole, C., Bunn, R., Oberman, T., Barrett, E., Aletta, F., Kang, J., Raynham, P., Mumovic, D., Davies, M.: Building performance evaluation of a new hospital building in the uk: Balancing indoor environmental quality and energy performance. Atmosphere (Basel). 12, (2021). https://doi.org/10.3390/ATMOS12010115.

[7] Felgueiras, F., Mourão, Z., Moreira, A., Gabriel, M.F.: Indoor environmental quality in offices and risk of health and productivity complaints at work: A literature review. Journal of Hazardous Materials Advances. 10, (2023). https://doi.org/10.1016/j.hazadv.2023.100314.

[8] Paleologos, K.E., Selim, M.Y.E., Mohamed, A.M.O.: Indoor air quality. In: Pollution Assessment for Sustainable Practices in Applied Sciences and Engineering. pp. 405–489. Elsevier (2020). https://doi.org/10.1016/B978-0-12-809582-9.00008-6.

[9] Cheng, J.C.P., Kwok, H.H.L., Li, A.T.Y., Tong, J.C.K., Lau, A.K.H.: BIM-supported sensor placement optimization based on genetic algorithm for multi-zone thermal comfort and IAQ monitoring. Build Environ. 216, (2022). https://doi.org/10.1016/j.buildenv.2022.108997.

[10] Karimi, S., Asghari, M., Rabie, R., Emami Niri, M.: Machine learning-based white-box prediction and correlation analysis of air pollutants in proximity to industrial zones. Process Safety and Environmental Protection. 178, 1009–1025 (2023). https://doi.org/10.1016/j.psep.2023.08.096.

[11] Van Tran, V., Park, D., Lee, Y.C.: Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality, (2020). https://doi.org/10.3390/ijerph17082927.

[12] Salih Hasan, B.M., Abdulazeez, A.M.: A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining. 02, (2021). https://doi.org/10.30880/jscdm.2021.02.01.003.

[13] Saini, J., Dutta, M., Marques, G.: A comprehensive review on indoor air quality monitoring systems for enhanced public health, (2020). https://doi.org/10.1186/s42834-020-0047-y.

[14] Al-Amri, R., Murugesan, R.K., Man, M., Abdulateef, A.F., Al-Sharafi, M.A., Alkahtani, A.A.: A review of machine learning and deep learning techniques for anomaly detection in iot data, (2021). https://doi.org/10.3390/app11125320.

[15] Mengara Mengara, A.G., Park, E., Jang, J., Yoo, Y.: Attention-Based Distributed Deep Learning Model for Air Quality Forecasting. Sustainability (Switzerland). 14, (2022). https://doi.org/10.3390/su14063269.

[16] Zhang, K., Yang, X., Cao, H., Thé, J., Tan, Z., Yu, H.: Multi-step forecast of PM2.5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning. Environ Int. 171, (2023). https://doi.org/10.1016/j.envint.2022.107691.

[17] Gabriel, M., Auer, T.: LSTM Deep Learning Models for Virtual Sensing of Indoor Air Pollutants: A Feasible Alternative to Physical Sensors. Buildings. 13, (2023). https://doi.org/10.3390/buildings13071684.

[18] Abirami, S., Chitra, P.: Regional air quality forecasting using spatiotemporal deep learning. J Clean Prod. 283, (2021). https://doi.org/10.1016/j.jclepro.2020.125341.

[19] Prado-Rujas, I.I., García-Dopico, A., Serrano, E., Córdoba, M.L., Pérez, M.S.: A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on Deep Learning. Eng Appl Artif Intell. 127, (2024). https://doi.org/10.1016/j.engappai.2023.107271.

[20] Zaini, N., Ahmed, A.N., Ean, L.W., Chow, M.F., Malek, M.A.: Forecasting of fine particulate matter based on LSTM and optimization algorithm. J Clean Prod. 427, (2023). https://doi.org/10.1016/j.jclepro.2023.139233.

[21] Tekler, Z.D., Chong, A.: Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy. Build Environ. 226, (2022). https://doi.org/10.1016/j.buildenv.2022.109689.

[22] Zeng, Y., Chen, J., Jin, N., Jin, X., Du, Y.: Air quality forecasting with hybrid LSTM and extended stationary wavelet transform. Build Environ. 213, (2022). https://doi.org/10.1016/j.buildenv.2022.108822.

[23] Yuan, E., Yang, G.: SA–EMD–LSTM: A novel hybrid method for long-term prediction of classroom PM2.5 concentration. Expert Syst Appl. 230, (2023). https://doi.org/10.1016/j.eswa.2023.120670.

[24] Sarkar, N., Gupta, R., Keserwani, P.K., Govil, M.C.: Air Quality Index prediction using an effective hybrid deep learning model. Environmental Pollution. 315, (2022). https://doi.org/10.1016/j.envpol.2022.120404.

[25] Gokul, P.R., Mathew, A., Bhosale, A., Nair, A.T.: Spatio-temporal air quality analysis and PM2.5 prediction over Hyderabad City, India using artificial intelligence techniques. Ecol Inform. 76, (2023). https://doi.org/10.1016/j.ecoinf.2023.102067.

[26] Sun, X., Tian, Z., Zhang, Z.: A new decomposition-integrated air quality index prediction model. Earth Sci Inform. 16, 2307–2321 (2023). https://doi.org/10.1007/s12145-023-01028-1.

[27] Pozo-Luyo, C.A., Cruz-Duarte, J.M., Amaya, I., Ortiz-Bayliss, J.C.: Forecasting PM2.5 concentration levels using shallow machine learning models on the Monterrey Metropolitan Area in Mexico. Atmos Pollut Res. 14, (2023). https://doi.org/10.1016/j.apr.2023.101898.

[28] Kim, M.H., Park, H.J.: Application of artificial neural networks using sequential prediction approach in indoor airflow prediction. Journal of Building Engineering. 69, (2023). https://doi.org/10.1016/j.jobe.2023.106319.

[29] Yan, B., Yang, W., He, F., Zeng, W.: Occupant behavior impact in buildings and the artificial intelligence-based techniques and data-driven approach solutions, (2023). https://doi.org/10.1016/j.rser.2023.113372.

[30] Korsavi, S.S., Montazami, A., Mumovic, D.: The impact of indoor environment quality (IEQ) on school children’s overall comfort in the UK; a regression approach. Build Environ. 185, (2020). https://doi.org/10.1016/j.buildenv.2020.107309.

[31] Cheng, C.H., Tsai, M.C.: An Intelligent Time Series Model Based on Hybrid Methodology for Forecasting Concentrations of Significant Air Pollutants. Atmosphere (Basel). 13, (2022). https://doi.org/10.3390/atmos13071055.

[32] Sethi, J.K., Mittal, M.: Ambient Air Quality Estimation Using Supervised Learning Techniques. EAI Endorsed Transactions on Scalable Information Systems. 6, 1–10 (2019). https://doi.org/10.4108/eai.13-7-2018.159406.

[33] Guo, Q., He, Z., Wang, Z.: Simulating daily PM2.5 concentrations using wavelet analysis and artificial neural network with remote sensing and surface observation data. Chemosphere. 340, (2023). https://doi.org/10.1016/j.chemosphere.2023.139886.

[34] Koo, J.W., Wong, S.W., Selvachandran, G., Long, H.V., Son, L.H.: Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models. Air Qual Atmos Health. 13, 77–88 (2020). https://doi.org/10.1007/s11869-019-00772-y.

[35] Rastogi, K., Lohani, D.: Context-aware IoT-enabled framework to analyse and predict indoor air quality. Intelligent Systems with Applications. 16, (2022). https://doi.org/10.1016/j.iswa.2022.200132.

[36] Tella, A., Balogun, A.L., Adebisi, N., Abdullah, S.: Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. Atmos Pollut Res. 12, (2021). https://doi.org/10.1016/j.apr.2021.101202.

[37] Alhathloul, S.H., Mishra, A.K., Khan, A.A.: Low visibility event prediction using random forest and K-nearest neighbor methods. Theor Appl Climatol. (2023). https://doi.org/10.1007/s00704-023-04697-6.

[38] Mahmud, S., Ridi, T.B.I., Miah, M.S., Sarower, F., Elahee, S.: Implementing Machine Learning Algorithms to Predict Particulate Matter (PM2.5): A Case Study in the Paso del Norte Region. Atmosphere (Basel). 13, (2022). https://doi.org/10.3390/atmos13122100.

[39] Kim, J., Bang, J. Il, Choi, A., Moon, H.J., Sung, M.: Estimation of Occupancy Using IoT Sensors and a Carbon Dioxide-Based Machine Learning Model with Ventilation System and Differential Pressure Data. Sensors. 23, (2023). https://doi.org/10.3390/s23020585.

[40] Vassella, C.C., Koch, J., Henzi, A., Jordan, A., Waeber, R., Iannaccone, R., Charrière, R.: From spontaneous to strategic natural window ventilation: Improving indoor air quality in Swiss schools. Int J Hyg Environ Health. 234, (2021). https://doi.org/10.1016/j.ijheh.2021.113746.

[41] Ngarambe, J., Yun, G.Y., Santamouris, M.: The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls, (2020). https://doi.org/10.1016/j.enbuild.2020.109807.

[42] Kim, J., Zhou, Y., Schiavon, S., Raftery, P., Brager, G.: Personal comfort models: Predicting individuals’ thermal preference using occupant heating and cooling behavior and machine learning. Build Environ. 129, 96–106 (2018). https://doi.org/10.1016/j.buildenv.2017.12.011.

[43] Garbin, C., Zhu, X., Marques, O.: Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl. 79, 12777–12815 (2020). https://doi.org/10.1007/s11042-019-08453-9.

Visualization of Personality and Phobia Type Clustering with GMM and Spectral

Ting Tin Tin[1, *], Cheok Jia Wei[2], Ong Tzi Min[3], Lim Siew Mooi[4],
Lee Kuok Tiung[5], Ali Aitizaz[6], Chaw Jun Kit[7], Ayodeji Olalekan Salau[8]
Faculty of Data Science and Information Technology, INTI International University, Kuala Lumpur, Malaysia[1]
Faculty of Computing and Information Technology,
Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2, 3, 4]
Faculty of Social Science and Humanities, Universiti Malaysia Sabah, Sabah, Malaysia[5]
School of Technology, Asia Pacific University, Kuala Lumpur, Malaysia[6]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Malaysia[7]
Department of Electrical / Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria[8]
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India[8]

***Abstract*—Personality traits, the unique characteristics defining individuals, have intrigued philosophers and scholars for centuries. With recent advances in machine learning, there is an opportunity to revolutionize how we understand and differentiate personality traits. This study seeks to develop a robust cluster analysis approach (unsupervised learning) to efficiently and accurately classify individuals based on their personality traits, overcoming the limitations of manual classification. The problem at hand is to create a system that can handle the subjective nature of qualitative personality data, providing insights into how people interact, collaborate, and behave in various social contexts and thus increase learning opportunities. To achieve this, various unsupervised clustering techniques, including spectral clustering and Gaussian mixture models, will be employed to identify similarities in unlabeled data collected through interview questions. The clustering approach is crucial in helping policy makers to identify suitable approaches to improve teamwork efficiency in both educational institutions and job industries.**

***Keywords*—Unsupervised learning; learning opportunities; clustering; personality; machine learning; Gaussian mixture model; spectral clustering**

I. Introduction

Personality traits have been discussed among people and philosophers since the early ages. Personality traits are a set of characteristics and attributes that define the uniqueness of an individual. People have unique personality traits that experts agree are the result of traits mixed with learnt behaviors [1]. Throughout history, fellow philosophers such as Aristotle, Theophrastus, and many others have studied human behaviors and tried to make sense of them. For example, Aristotle used vanity, modesty, and cowardice as factors of moral and immoral behavior [2]. Until now, many efforts have been made to understand how humans function and how their personality and social behaviors correlate with each other using various methods such as psychology testing, surveys and questionnaires, social experiments, computer modeling, and simulations.

With recent advances in Machine Learning technology, there is an opportunity to improve the accuracy of differentiating people with different personality traits. Such systems enable people to understand how others communicate, collaborate, manage stress, and many more. This allows them to have a better grasp of how to resolve conflicts better and communicate more effectively. By creating a personality recognition system, we can create different styles of teaching, coaching, leading, communicating, and many more [3]. For example, where personality traits affect workplaces, by understanding more about our peers, we can build a more cohesive team as we can understand how they work and how we can work with them. A unique personality is like a double-edged sword. Sometimes, it can increase the whole team's synergy and increase efficiency, and sometimes it can be the chisel that tears the whole team apart. It is okay to have people with different personalities working together, but the major concern is the way to communicate with different people to avoid conflicts within the team. Therefore, we must consider personality traits to increase efficiency, synergy, and a more dynamic and cohesive team [4].

Based on previous related work, the researchers had attempted to classify and analyze personality traits manually. This has led to several issues where identifying personality traits requires a large amount of time, making it inefficient to analyze a large group of people [5]. Due to the problem stated above, it is necessary to combine the technology of computer science with the study of personality psychology to increase the efficiency and precision of identifying different personality traits to personalize e-learning [6]. Furthermore, personality traits are difficult to analyze as it is qualitative data that are subjective to individuals rather than quantitative data which represents numerical facts [7], [8], [9]. Today, unsupervised machine learning methods such as clustering are commonly used to group people with similar personality traits into a group of clusters. For example, a study by Feng et al. (2008) used fuzzy clustering analysis based on the common characteristics to create a personalized study strategy [10].

Clustering Analysis is where the computer tries to identify a structure or similar pattern in unlabeled data. Cluster is an important term in cluster analysis, where a cluster is a collection of similar items. There are a few commonly used clustering methods such as Hierarchical Clustering, which identifies different clusters using a tree-shaped structure known as the dendrogram, Partitional Clustering such as k-means clustering, where the algorithm reallocates items to an initially specified number of groups and density-based clustering, such as DBSCAN algorithm which is used to discover randomly shaped clusters [11]. These methods can be used to differentiate the different types of personality based on different characteristics and categories such as questions about phobias and personality traits.

Understanding and characterizing human personalities is critical in today's interconnected society, from human resources and marketing to healthcare and social sciences. However, the existing analysis of personality traits is mainly dependent on manual classification and subjective interpretation, leading to inefficiencies and possible biases, particularly for large and diverse populations. Therefore, there is an urgent need for a more efficient and accurate personality recognition system that uses machine learning technology. The major problem is to create a robust clustering analysis approach that can handle qualitative data, which is inherently subjective and complicated. The proposed clustering techniques can divide people into distinct groups based on their personality traits, providing valuable information on how people interact, collaborate, and behave in different social circumstances.

This technology can alter how we understand human behavior, improve decision-making processes, and promote personalized experiences and services by taking a data-driven and objective approach to personality assessment. However, careful consideration of ethical and privacy problems and validation with psychiatric experts are required for the appropriate development and implementation of such a system. In this context, future research should look at novel ways to integrate qualitative and quantitative data, addressing cross-cultural differences, and ensuring the system's interpretability and fairness to enable significant applications across various areas.

This study aims to identify similarities in unlabeled data related to personality traits using unsupervised clustering methods. Various clustering techniques will be utilized to group different personality traits according to interview questions. For instance, clustering techniques include Spectral Clustering, Ordering Points To Identify the Clustering Structure (OPTICS), K-modes, Gaussian Mixture Model, and Affinity Propagation. We can validate the precision and effectiveness of the personality traits for each clustering technique after training, testing and evaluating those models. Therefore, we can advance our understanding of human behavior and personality traits through data-driven approaches.

This paper is constructed with the following sections. Section II presents the latest studies and research on clustering personality using different algorithms. Section III explains the research methodology design. This is followed by Section IV which discusses the result and finally Section V concludes the study with limitations and future works.

II. Literature Review

A. Spectral Clustering

Spectral clustering has become more common recently as it is a simple algorithm to implement and does not require a large amount of resources to process the algorithm [12]. There are a few pros and cons regarding spectral clustering. For example, this algorithm applies to data with high dimensionality, and it also handles categorical variables well as it can calculate the similarity between data points by using an eigenvector instead of using distance to calculate the data points, which is crucial to our research. On the other hand, spectral clustering is relatively slow compared to other traditional algorithms such as k-mean clustering and we are required to determine the k-value of the spectral clustering algorithm, which may be hard if we do not have an intuition on the number of clusters a dataset should have [13], [14].

The algorithm falls into the category of clustering like k-mean clustering and uses the eigenvector of a matrix obtained from the distance between the data points. Eigenvectors and eigenvalues are relatively important concepts in spectral clustering, where the eigenvector is interpreted as a vector that undergoes pure scaling without any rotation and the eigenvalue is interpreted as the scaling factor of a vector [15].

$$A\upsilon = \lambda\upsilon \tag{1}$$

The first step of spectral clustering is to construct an affinity matrix based on the data set. The affinity matrix is used to construct and determine a matrix of similarity between the data points. The affinity matrix is then normalized and partitioned using the largest K-eigenvectors [16]. Since spectral clustering is considered partitional clustering, it has the evaluation metrics of silhouette score, Calinski-Harabasz index, and Davies-Bouldin Index. On the basis of the previous study, we can see that spectral clustering is commonly applied in various real-life situations. For example, a study conducted by McFee & Ellis (2014) used the algorithm to analyze the structure of songs that detects repeated patterns in songs [13]. Furthermore, in a study conducted by Bach & Jordan (2006), (1) is used to perform speech separation [17].

B. OPTICS

OPTICS (Ordering Points to Identify the Clustering Structure) is categorized as a density-based clustering algorithm. The algorithm is inspired by DBSCAN but with a few better modifications. For example, the OPTICS clustering algorithm can partition data with varying densities and shapes. It is widely used in large data sets with high dimensionality [18]. The OPTICS algorithm works by adding a random data point in the cluster to an order list and then continuing to expand the cluster iteratively based on the closest data point to the selected data points. The OPTICS algorithm also calculates the reachability distance for each data point [19]. Based on previous studies, the algorithm is also commonly used in real-life applications such as detecting outliers in data sets and clustering wireless sensor networks [20], [21], [22].

Reachability distance is a measurement that indicates how easily can a datapoint be reached by another data point can reach a data point. Euclidean distance is used to calculate the distance between two data points and then further be used to determine the reachability distance. The algorithm computes the reachability distance of each data point and the generated data is used to plot out the reachability plot which can help identify clusters and the hierarchical structure of the data [23].

Even though the OPTICS algorithm uses similar concept as DBSCAN algorithm which uses density-based clustering, the OPTICS algorithm is considered a better algorithm as the OPTICS algorithm maintains a priority queue to determine the reachability distance, whereas DBSCAN only uses radius queries. Furthermore, the OPTICS clustering technique requires less maintenance, as it has fewer parameters compared to DBSCAN which requires one to maintain the epsilon parameters but to optimize the parameters of the OPTICS algorithm, domain knowledge is required as it is very sensitive to parameters that define density [24]. In addition to that, the OPTICS algorithm is also good at handling clusters with different densities, whereas the DBSCAN clustering algorithm struggles to handle data sets with different densities of clusters, since it only depends on a single value of epsilon to help determine the cluster size for all data points [18].

C. KMODES

The K mode algorithm is a variant of one of the most popular clustering algorithms, the K mean algorithm. K-modes are designed to specifically handle categorical data instead of numerical data which the K-means algorithm is weak at. The k-mode algorithm calculates the mode instead of the mean of the clusters. This modification allows the K-modes algorithm to cluster large categorical datasets more efficiently compared to the K-means algorithm [25]. Furthermore, based on previous studies, the algorithm is used for customer segmentation in e-Commerce business [26]. In addition to that, the algorithm is also used to detect and prevent crime by combining data mining technology with the algorithm. Equations (2) and (3) show the calculation of KMODES.

$$d(x, y) = \sum_{i=1}^{f} \delta(X_j, Y_j) \tag{2}$$

$$C(Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \tag{3}$$

The K-modes algorithm works by first generating a number of clusters by randomly selecting data points to act as the initial cluster centers with each data point representing the centroid of the cluster where the value k is selected manually by the user. Then, each data point in the dataset is assigned to the cluster whose cluster center is closest to it based on the second equation above. After assignment, the clusters are updated based on the allocated data points, and the cluster centers are recalculated to represent the new centroids of each cluster. This update involves calculating the latest modes for each cluster. These steps and calculations are repeated until convergence is reached or stopped based on predefined criteria [27].

D. Gaussian Mixture Model

The Gaussian mixture model is a parametric density function of probability that can be represented as the weighted sum of Gaussian component densities in many applications, such as image clustering to detect human skin color and image segmentation, identifying restaurant hotspots, weather forecasts, flight safety monitoring, and many others [28], [29], [30], [31], [32], [33]. Before getting into the Gaussian mixture model, two key components of clustering must be known which are hard clustering and soft clustering. Hard clustering is the method where models try to force a data point to one of the clusters, and this means that the data point is assigned a membership degree to either 0 or 1. On the other hand, soft clustering models tend to assign data points to their appropriate membership value. Instead of assigning a membership degree of either 0 or 1, soft clustering assigns the membership degree of a point between 0 and 1. This means that each data point can belong to two or more clusters, and this may be more natural in many situations compared to hard clustering [34].

The Gaussian mixture model falls under the category of soft clustering model as it allows data points to be assigned partially to clusters. The model works by clustering data points into different clusters and thus estimating the probability density of new data points. For example, a data point can be 30% cluster A and 70% cluster B. The Gaussian mixture model is commonly applied in machine learning and pattern recognition due to the ability of the model to handle complex data distributions. Even so, it is difficult to implement this algorithm incorporating categorical variables, as the model assumes that all the features are normally distributed. The model works by applying the EM on top of its algorithm as shown in (4). The E-step uses training data to predict a new datapoint, whereas the M-step uses the information obtained from the E-step to calculate the derivative of the log-likelihood to cluster the new datapoint according to the calculations. This process repeats until convergence and the Q calculated from the EM step can be used for clustering purposes [35].

$$Q^i\left(z_k^{(i)}\right) = p\left(z_k^{(i)} \mid x^{(i)};\, \theta\right) = \frac{\pi_k N\,(x^{(i)};\, \mu_k, \Sigma k)}{\sum_{k=1}^{K} \pi_k N\,(x^{(i)};\, \mu_k, \Sigma k)} \tag{4}$$

E. Affinity Propagation

The Affinity Propagation algorithm is a clustering algorithm that first aims to find a data point as an exemplar for each cluster. It group the data based on either their similarity or distance [36]. This algorithm is commonly used to group images such as human faces, and is also used for forecasting rainfall [37]. Affinity propagation algorithm does not require determining the number of clusters, unlike algorithms like k-mean clustering. This is made possible due to the parameters used in the Affinity Propagation algorithm, such as preference, which handles the number of exemplars being used, and the damping factor, which is responsible for preventing large changes in the Responsibility and Availability matrices and ensuring convergence in the algorithm [38]. This method requires heavy calculations as it needs to calculate the similarity value between each data and reselect data as an exemplar. This makes the algorithm computationally expensive and makes it difficult to scale with larger datasets.

The algorithm performs by first creating a similarity matrix. Given a dataset, the similarity between data points needs to be calculated by using some commonly used equations, such as the negative Euclidean distance. The higher the similarity of the data points, the closer the data points are to each other based on the similarity equation. Then, two matrices, the Responsibility matrix, which quantifies how well-suited a data point is to be the exemplar of another data point, and the Availability matrix, which quantifies how much support a data point receives from another data point as an exemplar is created. Then both matrices are updated and recalculated iteratively until convergence and new data points can be clustered on the existing clusters. This algorithm is useful when the initial number of clusters is unknown, and it can handle complex data structures, too.

III. Methodology

A. Data Collection and Preparation

The data set that this study has chosen represents a list of survey answers regarding preference, interests, habits, opinions, and fears of people ages 15 years old to 30 years old. This survey was conducted by students of the Statistics class at FSEV UK in 2013. It contains 1010 unique records with 150 columns, and was taken from Kaggle, a platform where it consistently hosts various types of forecasting competitions, and it also contains data sets provided by various data scientists. Instead of taking the whole dataset, we have decided to focus on the phobia, personality, and demographic category in this data set. The demographic will be then combined with the personality subset, and phobia subset which may provide more information gain for the clustering algorithms. The phobia subset contains 10 independent variables which are flying, storm, darkness, heights, spiders, snakes, rats, aging, dangerous dogs, and fear of public speaking where it was rated from 1-5 indicating the severity of fear. The description of each variable in each subset is shown in Tables I and II. These variables in each subset will be used in clustering modelling.

TABLE I. Variables in the Personality and Phobia Subset (int64)

Personality Subset			
Daily Events Prioritizing workload Writing Notes Workaholism Thinking ahead Final judgment Reliability Keeping promises Loss of interest Funniness Fake Criminal damage Decision Making Friends versus money	Elections Self-criticism Judgment calls Hypochondria Empathy Eating to survive Giving Borrowed stuff Loneliness Cheating in school Health Change in the past God Dreams Charity Number of friends	Compassion to animals Punctuality Lying New Environment Mood swings Appearance and gestures Socializing Achievements Response to a serious letter Children Assertiveness Getting angry Knowing the right people	Public speaking Unpopularity Waiting Life struggles Happiness in life Energy levels Small – big dogs Personality Finding Lost Values Getting up Interests or hobbies Parent' advice Questionnaires or polls Internet usage
Phobia Subset			
Flying Storm Darkness	Heights Spiders Snakes	Rats Ageing	Dangerous Dogs Fear of public Speaking

TABLE II. Demographic Subset

Column Name	Type of Data
Age	Integer
Height	Integer
Weight	Integer
Number of siblings	Integer
Gender	String
Left – right-handed	String
Education	String
Only child	String
Village - town	String
House – block of flats	String

B. Feature Selection

Fig. 1 shows the method we have used to extract subsets from the dataset obtained from Kaggle. The phobia, personality, phobia + demographic and personality + demographic subset will then be used to fit the clustering algorithm. The dataset is first preprocessed by using methods like feature mapping to convert object data type into numeric datatype. Then the missing values are replaced by using Mean, Mode and Median depending on the distribution of the columns. Fig. 2 indicates that more than 80% of the number of friends' columns in the personality data set are null values. Therefore, the column is disputed due to lack of information in the column.

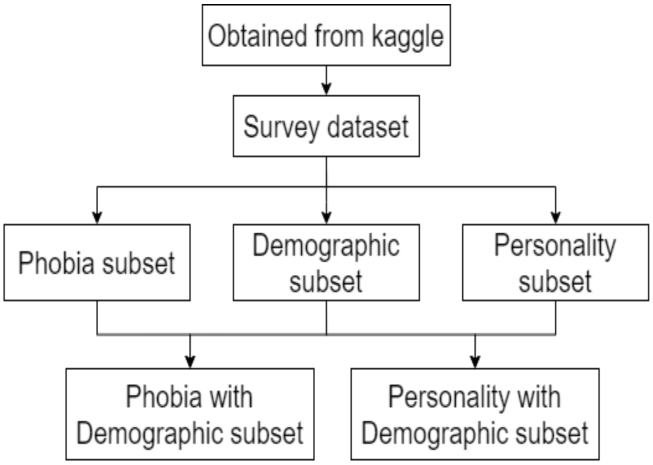


Fig. 1. Data pipeline.

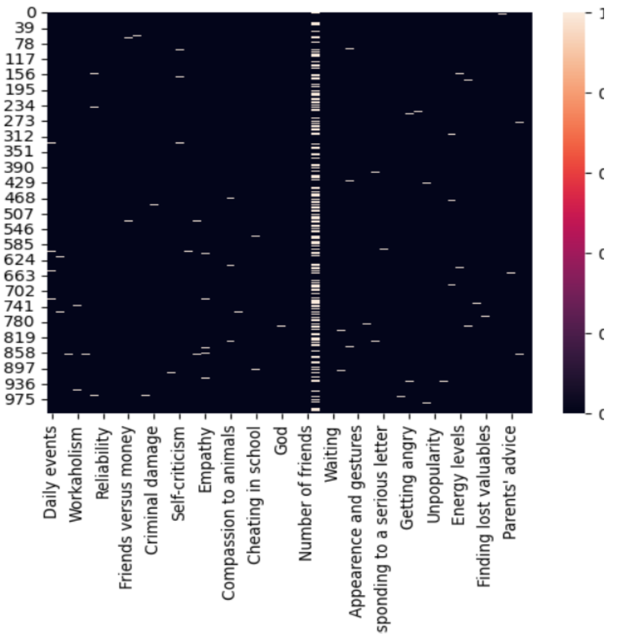


Fig. 2. Heatmap of missing value in the personality data set.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

Eq. (5) shows the formula of variance which will be used for feature selection in our case. Selecting features using variance threshold is a common and efficient way to dispute columns that do not provide information to the model. Feature with low variance indicates that the data in the feature do not vary much and thus not providing good information to the clustering model. 20% is used in our dataset and the variables are removed from the data set. This indicates that both variables do not provide enough information for the model as these two columns only consist of similar values, which decreases the variance of the variables. Features with low variance can sometimes be redundant, meaning they carry information like other features. Removing redundant features can help simplify your model and potentially improve its performance.

C. Algorithms Used

1) Spectral clustering: Spectral clustering has become more common recently as it is a simple algorithm to implement and does not require a large amount of resources to process the algorithm [12]. There are a few pros and cons regarding spectral clustering. For example, this algorithm applies to data with high dimensionality, and it also handles categorical variables well as it can calculate the similarity between data points by using an eigenvector instead of using distance to calculate the data points, which is crucial to our research. The differences between K-means and Spectral Clustering are shown in Fig. 3 where Spectral Clustering can perform on high-dimensionality data whereas K-means perform poorly. On the other hand, spectral clustering is relatively slow compared to other traditional algorithms such as k-mean clustering, and we are required to determine the k-value of the spectral clustering algorithm, which may be hard if we do not have an intuition on the number of clusters a dataset should have [14].

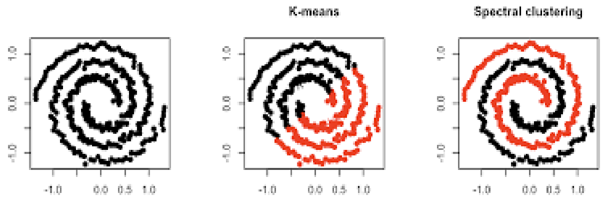


Fig. 3. Difference between K-means and spectral clustering.

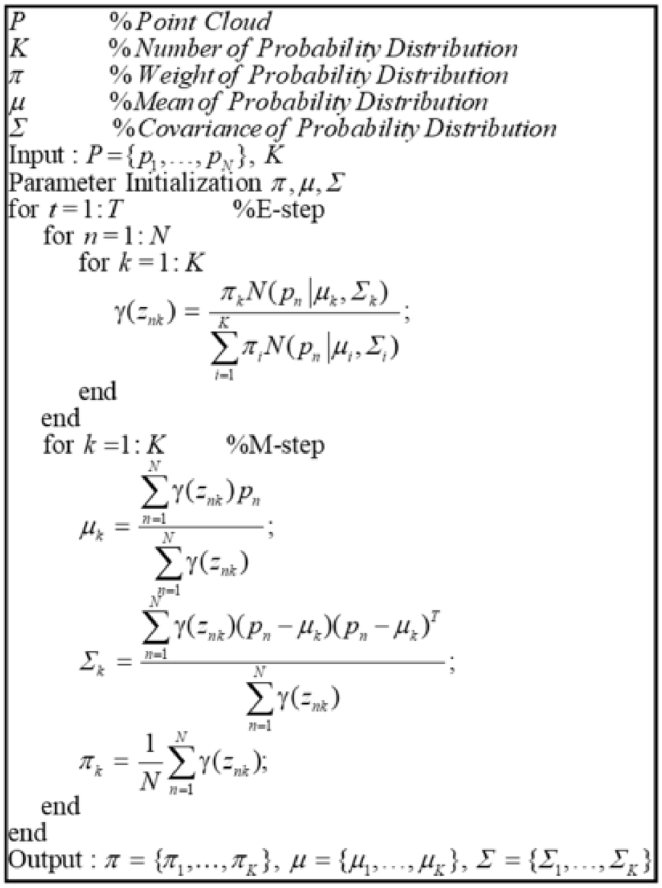
The algorithm falls into the category of clustering like k-mean clustering and uses the eigenvector of a matrix obtained from the distance between the data points. Eigenvectors and eigenvalues are relatively important concepts in spectral clustering, where the eigenvector is interpreted as a vector that undergoes pure scaling without any rotation and the eigenvalue is interpreted as the scaling factor of a vector [15].

$$Au = \lambda u \quad (6)$$

The first step of spectral clustering is to construct an affinity matrix based on the data set. The affinity matrix is used to construct and determine a matrix of similarity between the data points. The affinity matrix (6) is then normalized and partitioned using the largest K-eigenvectors [16]. Since spectral clustering is considered partitional clustering, it has the evaluation metrics of silhouette score, Calinski-Harabasz index, and Davies-Bouldin Index.

2) Gaussian Mixture Model (GMM): The Gaussian mixture model is a parametric density function of probability which can be represented as the weighted sum of the densities of the Gaussian components [28]. Based on the study conducted by other researchers, we can see that the algorithm was applied in image clustering where it detects human skin color and image segmentation [33]. Before getting into the Gaussian mixture model, two key components of clustering must be known which are hard clustering and soft clustering. Hard clustering is the method where models try to force a data point to one of the clusters, and this means that the data point is assigned a membership degree to either 0 or 1. This means that each data point can belong to two or more clusters, and this may be more natural in many situations compared to hard clustering. The Gaussian mixture model falls under the category of soft clustering model as it allows data points to be assigned partially to clusters. This process (Fig. 4) repeats until convergence and the Q calculated from the EM step can be used for clustering purposes [35]. The algorithm used is shown in (6).

$$Q^i\left(z_k^{(i)}\right) = p\left(z_k^{(i)} \,\middle|\, x^{(i)} ;\, \theta\right) = \frac{\pi_k N(x^{(i)}; \mu_k, \Sigma k)}{\sum_{k=1}^{K}\pi_k N(x^{(i)}; \pi_k, \Sigma k)} \quad (6)$$



```
$P$        %Point Cloud
$K$        %Number of Probability Distribution
$\pi$        % Weight of Probability Distribution
$\mu$        %Mean of Probability Distribution
$\Sigma$        %Covariance of Probability Distribution
Input : $P = \{p_1, \ldots, p_N\}$, $K$
Parameter Initialization $\pi, \mu, \Sigma$
for $t = 1 : T$              %E-step
    for $n = 1 : N$
        for $k = 1 : K$
            $\gamma(z_{nk}) = \frac{\pi_k N(p_n | \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i N(p_n | \mu_i, \Sigma_i)};$
        end
    end
    for $k = 1 : K$       %M-step
        $\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) p_n}{\sum_{n=1}^{N} \gamma(z_{nk})};$
        $\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(p_n - \mu_k)(p_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})};$
        $\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk});$
    end
end
Output : $\pi = \{\pi_1, \ldots, \pi_K\}$, $\mu = \{\mu_1, \ldots, \mu_K\}$, $\Sigma = \{\Sigma_1, \ldots, \Sigma_K\}$
```

Fig. 4. Pseudocode of the E-M step in GMM.

D. Model Pipeline

Fig. 5 indicates the clustering method used in this study. The four different subsets will be inputted into the spectral clustering model and the GMM model. Then we will determine whether the subset with demographic content or without demographic content will be used based on its performance in the model. After determining the subset to be used, the model will then be fine-tuned by using tools like GridSearchCV, silhouette score, and elbow graph to determine the best parameter combination. Lastly, each model will generate two clusters which are Phobia cluster and Personality cluster.

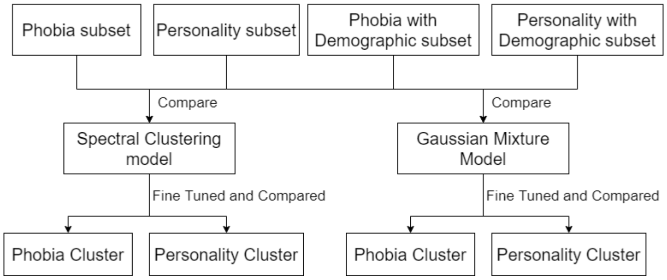


Fig. 5. Model pipeline flow chart.

IV. Result and Discussion

A. Exploratory Data Analysis (EDA)

This section explores the data set using heatmap and line graph, bar chart, violin plot, and cluster map. A heatmap is used to identify potential association between 1) phobias and demographic categories (age, gender, education) as shown in Fig. 6; and 2) personality traits and demographic categories. Based on the phobia vs. Age line graph in Fig. 7, we can see that the level of phobia roughly remains the same throughout the graph. Based on this, we can conclude that the things we fear are the same no matter the age. Fear of height can be the same for a 16-year-old young adult or even a 30-year-old adult.

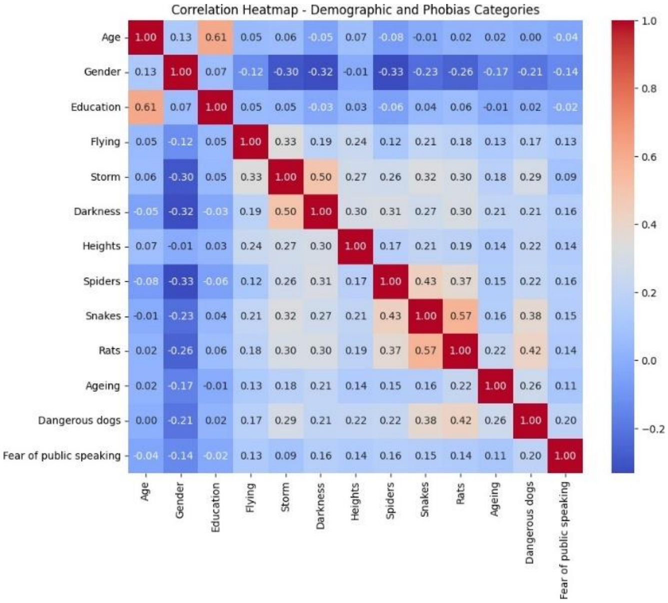


Fig. 6. Heatmap of demographic and phobia categories.

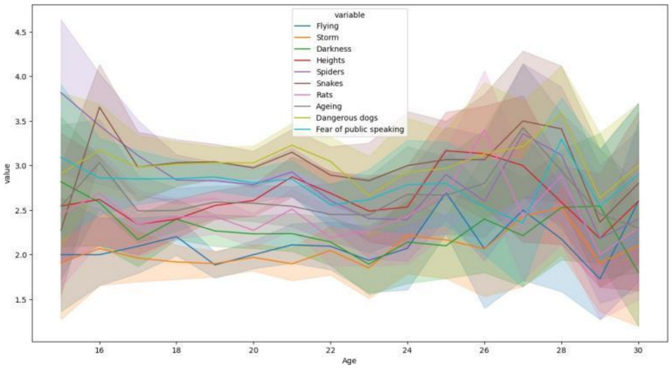


Fig. 7. Phobia vs. Age line graph.

From the bar graph in Fig. 8, 1 represents women and 2 represent males. Due to the large amount of data and bar graphs produced, only a snippet of the bar graph is presented in Fig. 8. The bar graph clearly indicates that female participants gave a neutral value in most views on life and opinion.

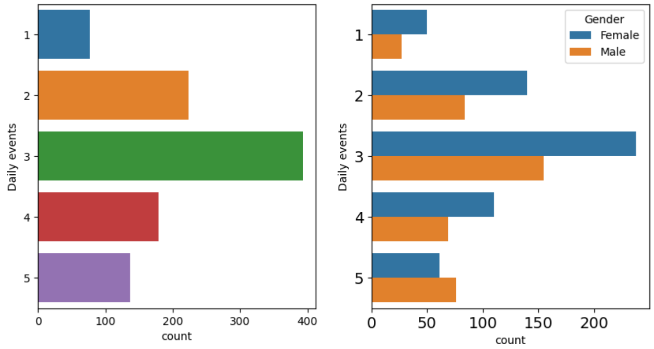


Fig. 8. Snippet of bar plots produced in exploratory data analysis using Python.

The shape of the violin represents the distribution of the data. The wider sections indicate higher density, while narrower sections indicate lower density. The width at a specific point on the violin corresponds to the frequency of data values at that point. Fig. 9 shows the distribution for level of phobia for each phobia based on the questionnaire.

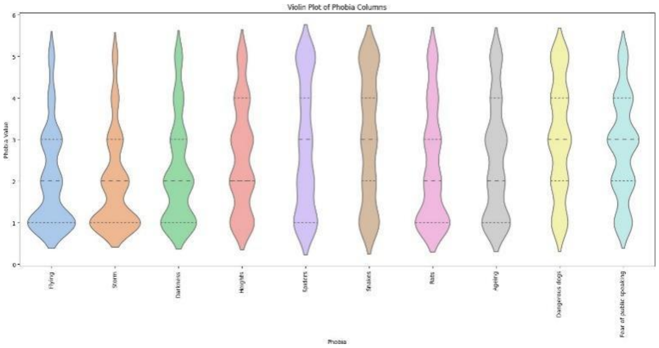


Fig. 9. Violin plot to represent data distribution.

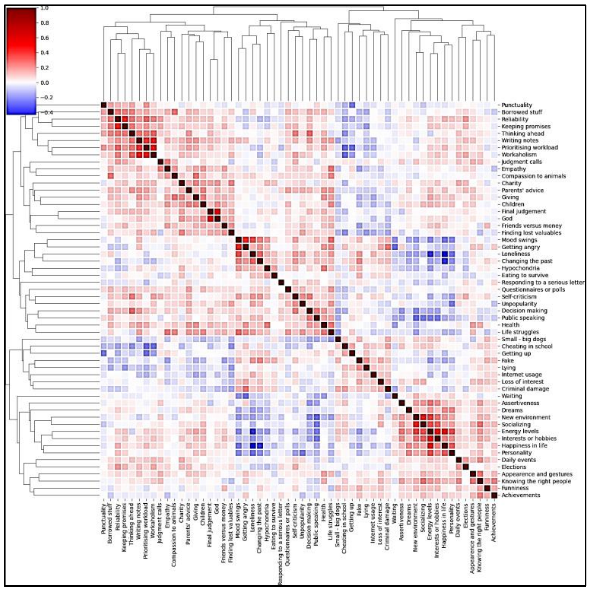


Fig. 10. Cluster map with associated clusters.

The cluster map is useful for identifying patterns not only in individual values but also in the arrangement of rows and columns. Group similar rows and/or columns together to form clusters, which can reveal underlying structures or relationships in the data. The cluster map in Fig. 10 indicates that the data can be grouped into 4, 5 or 6 clusters. Cluster map showing the correlation between each attribute in the personality test. Meanwhile, Fig. 11 cluster map shows that weight, height, and gender are highly correlated, followed by age and education and followed by number of siblings and only child. Therefore, the data can be grouped into 3 clusters. According to Fig. 12 cluster map, there may have 2 clusters for phobia data frame which are storm and darkness, as well as spiders, snakes, and rats.

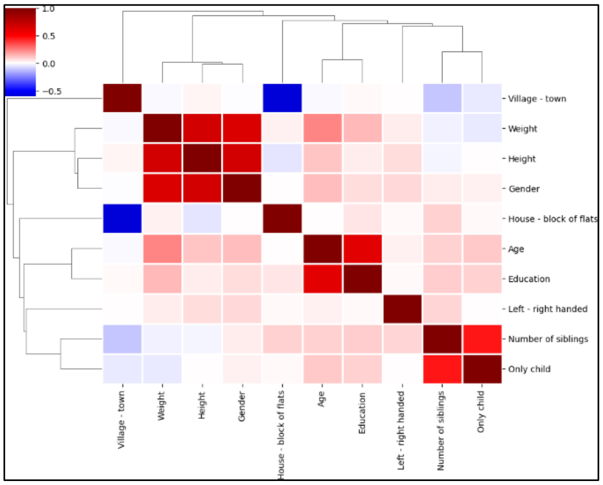


Fig. 11. Cluster map for displaying the correlation between demographics.

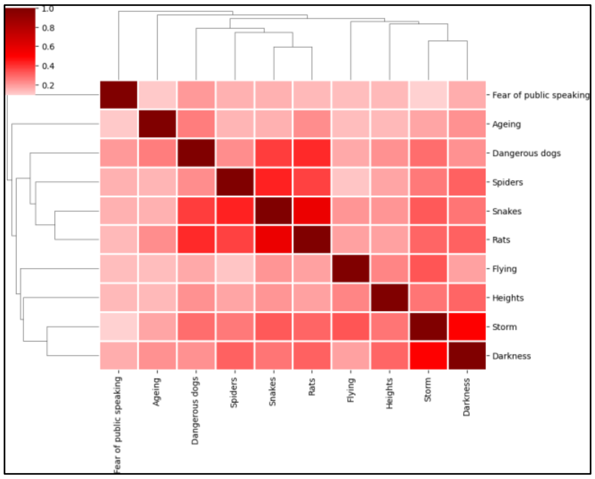


Fig. 12. Cluster map to display the correlation between each types of phobia.

B. Gaussian Mixture Model

By combining the result distortion elbow method and silhouette score graph in different numbers of clusters, we can determine that the optimal dataset and number of clusters are a subset without demographic content, 2 clusters for phobia subset and 4 clusters for personality subset. Based on Fig. 13, even though the graph indicates that 2 clusters is optimal, we have decided to use 4 clusters because it represents four main dominant traits of a person.

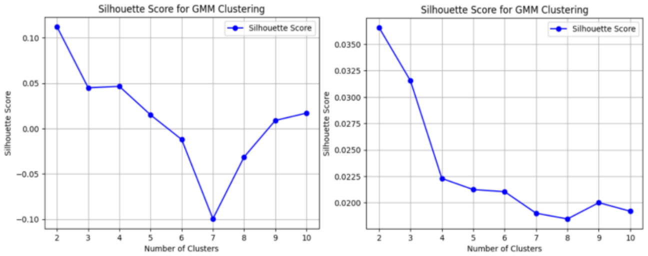


Fig. 13. GMM Silhouette score graph for the phobia subset and the personality subset.

The graphs on Fig. 14 and 15 are obtained by letting the model fit and predict the dimensional reduced data. The dimensionality reduction phase is carried out by using Principal Component Analysis (PCA). Based on Fig. 7, we can see that the yellow groups represent Level 1 phobia while the purple clusters represent Level 2 phobia. Furthermore, the clusters in Fig. 8 also show 4 different clusters of dominant traits. But in the case of GMM, the clusters are not well defined and are all grouped together. Therefore, we can determine that the GMM clustering model is not suitable for clustering personalities as it cannot identify the hidden relationship in the data set.

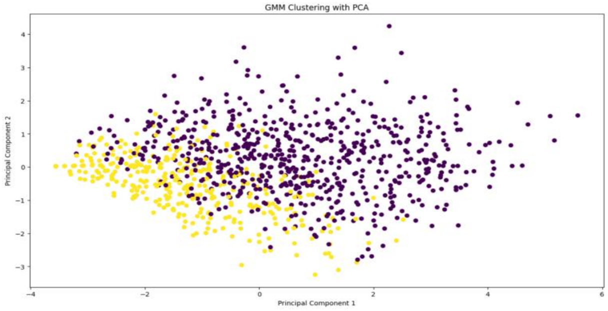


Fig. 14. Groups of GMM phobia subsets.

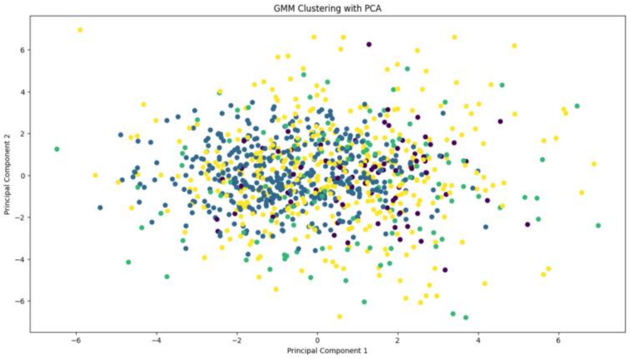


Fig. 15. Groups of personality subsets of GMM.

C. Spectral Clustering

The process of selecting parameters for spectral clustering is the same as GMM. The subset with demographic contents is selected, as the result has shown that subset with demographic contents provide better information gain and higher silhouette score. By observing both the distortion elbow method and silhouette analysis in Fig. 16 and 17 we can see that the optimal cluster is 2 clusters for the phobia subset and 4 clusters for the personality subset. Based on Fig. 9, 4 clusters since it represents 4 main dominant traits of a person. Not only that, we have also used GridSearchCV to adjust parameters like “affinity”, “gamma”, and “eigen_solver”. But the best combination found by GridSearchCV does not cluster the data well, therefore, we have decided to use the default parameters.

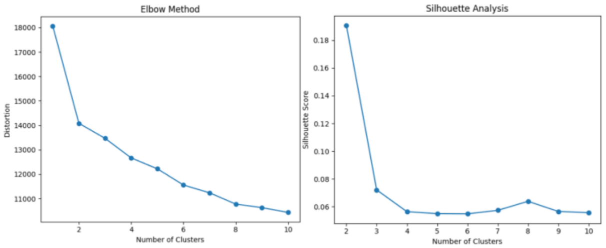


Fig. 16. Spectral clustering distortion and silhouette graph for phobia with demographic subset.

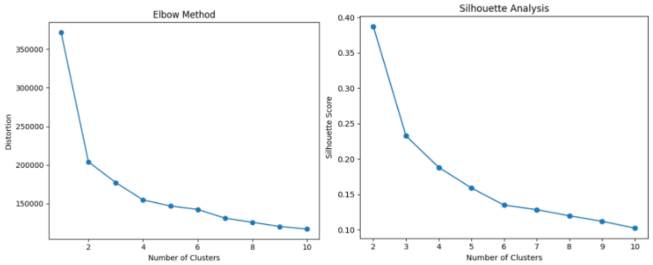


Fig. 17. Spectral clustering distortion and silhouette graph for personality with demographic subset.

The graphs in Fig. 18 and 19 are obtained from fitting dimensionality reduced data by using PCA into the spectral clustering model. Based on Fig. 18, we can see that the yellow clusters represent Level 1 phobia, while the purple clusters represent Level 2 phobia. Furthermore, the clusters in Fig. 19 also show four different clusters of dominant traits where the yellow clusters (Cluster 0) represent Steady personality type, purple clusters (Cluster 1) represent Influential personality type, blue clusters (Cluster 2) represent Compliant Personality Type, and green cluster (Cluster 3) represents dominant personality type.

Compared to previous research of clustering phobias and personality categories, one very relevant research by Anik et al. (2024) clustered the different types of phobia types using BERT-based classification model on Tweet data set. This present study presents different approaches in classification which provides information to alternative methods to analyze different data sets and discover different results or visualization.

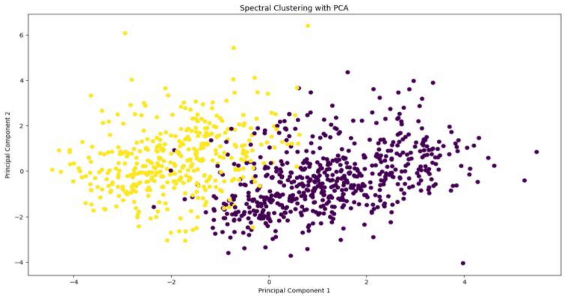


Fig. 18. Spectral clustering phobia with demographic subset clusters.

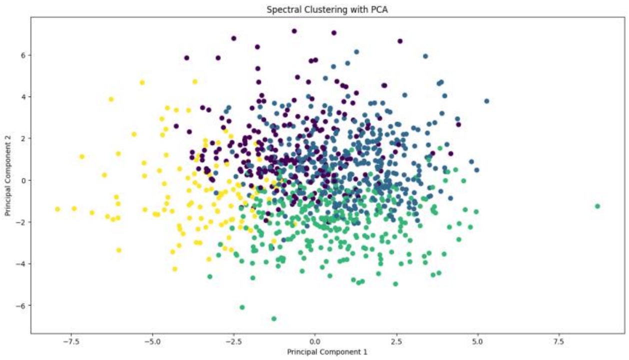


Fig. 19. Spectral clustering personality with demographic subset clusters.

V. Conclusion

As we can see, we have used two different models for two different datasets (phobias and personality categories) (one with demographic content and one without it). The list of models used for GMM are 1) GMM for phobia dataset (with demographic content); 2) GMM for phobia dataset (Without demographic content); 3) GMM for personality dataset (With demographic content); 4) GMM for personality dataset (Without demographic content). Meanwhile, for spectral clustering, the list of models used are: 1) SC for phobia dataset (with demographic content); 2) SC for phobia dataset (Without demographic content); 3) SC for personality dataset (With demographic content); 4) SC for personality dataset (Without demographic content).

For the phobia dataset, the best-performing model is Spectral Clustering, as it has a higher average silhouette score and thus makes well-defined clusters compared to GMM. For the personality dataset, the best-performing model is also spectral clustering, as it can identify the hidden relationship between the questions and thus cluster a more distinct cluster from each other. This can be shown by different clusters having more distinct characteristics, and the clusters are also less intersected with other clusters. Therefore, we can use this model to identify the dominant traits of an individual, to understand more about them. This helps identify strengths and weaknesses of the person, and thus reduces conflicts and makes coordination and cooperation much easier in many domains, especially group work in the school, university, or job. The models can be further tested using different data sets from different countries to ensure generalization of the result.

References

[1] H. Gillette, “Are you born with personality or does it develop later on? Psych Central.” Accessed: Jul. 19, 2024. [Online]. Available: https://psychcentral.com/health/personality-development

[2] G. Matthews, I. J. Deary, and M. C. Whiteman, Personality traits, 2nd ed. Cambridge University Press, 2005.

[3] L. Mosley, “The importance of understanding personality type in the workplace,” LinkedIn. Accessed: Jul. 19, 2024. [Online]. Available: https://www.linkedin.com/pulse/importance-understanding-personality-type-workplace-lauren-copeland/

[4] A. Cabrera, “Personalities in the workplace: Why is it important?,” Peopledynamics. Accessed: Jul. 19, 2024. [Online]. Available: https://peopledynamics.co/personalities-workplace-importance/#:~:text=Understanding%20your%20people%27s%20personalities%20can,chisel%20that%20tears%20it%20apart

[5] A. Talasbek, A. Serek, M. Zhaparov, S.-M. Yoo, Y.-K. Kim, and G.-H. Jeong, “Personality Classification Experiment by Applying k-Means Clustering,” International Journal of Emerging Technologies in Learning (iJET), vol. 15, no. 16, p. 162, Aug. 2020, doi: 10.3991/ijet.v15i16.15049.

[6] R. Karthika, V. E. Jesi, M. S. Christo, L. J. Deborah, A. Sivaraman, and S. Kumar, “Intelligent personalised learning system based on emotions in e-learning,” Pers Ubiquitous Comput, vol. 27, no. 6, pp. 2211–2223, Dec. 2023, doi: 10.1007/s00779-023-01764-7.

[7] S. M. Aslam, A. K. Jilani, J. Sultana, and L. Almutairi, “Feature Evaluation of Emerging E-Learning Systems Using Machine Learning: An Extensive Survey,” IEEE Access, vol. 9, pp. 69573–69587, 2021, doi: 10.1109/ACCESS.2021.3077663.

[8] M. Rahman, H. Sarwar, MD. A. Kader, T. Gonçalves, and T. T. Tin, “Review and Empirical Analysis of Machine Learning-Based Software Effort Estimation,” IEEE Access, vol. 12, pp. 85661–85680, 2024, doi: 10.1109/ACCESS.2024.3404879.

[9] C. W. Puah, W. L. Eng, C. H. Tan, S. C. Tan, and T. T. Ting, “Digital Culture: Online shopping adoption among college students in Malaysia,” in International Conference on Digital Transformation and Applications, 2021, pp. 137–143.

[10] Feng Tian, Shibin Wang, Cheng Zheng, and Qinghua Zheng, “Research on E-learner Personality Grouping Based on Fuzzy Clustering Analysis,” in 2008 12th International Conference on Computer Supported Cooperative Work in Design, IEEE, Apr. 2008, pp. 1035–1040. doi: 10.1109/CSCWD.2008.4537122.

[11] T. S. Madhulatha, “AN OVERVIEW ON CLUSTERING METHODS,” IOSR Journal of Engineering, vol. 02, no. 04, pp. 719–725, Apr. 2012, doi: 10.9790/3021-0204719725.

[12] U. von Luxburg, “A tutorial on spectral clustering,” Stat Comput, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.

[13] B. McFee and D. P. W. Ellis, “Analyzing Song Structure with Spectral Clustering,” in 15th International Society for Music Information Retrieval Conference, 2014.

[14] C. Ellis, “When to use spectral clustering.” Accessed: Jul. 20, 2024. [Online]. Available: https://crunchingthedata.com/when-to-use-spectral-clustering/

[15] H. Abdi, “The Eigen-Decomposition: Eigenvalues and Eigenvectors.” Accessed: Jul. 19, 2024. [Online]. Available: https://personal.utdallas.edu/~herve/Abdi-EVD2007-pretty.pdf

[16] X.-Y. Li and L. Guo, “Constructing affinity matrix in spectral clustering based on neighbor propagation,” Neurocomputing, vol. 97, pp. 125–130, Nov. 2012, doi: 10.1016/j.neucom.2012.06.023.

[17] F. R. Bach and M. I. Jordan, “Learning spectral clustering, with application to speech separation,” Journal of Machine Learning Research, vol. 7, pp. 1963–2001, 2006, Accessed: Jul. 19, 2024. [Online]. Available: https://jmlr.csail.mit.edu/papers/volume7/bach06b/bach06b.pdf

[18] A. Gupta, “ML | OPTICS Clustering Explanation,” geeksforgeeks. Accessed: Jul. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-optics-clustering-explanation/

[19] Z. Deng, Y. Hu, M. Zhu, X. Huang, and B. Du, “A scalable and fast OPTICS for clustering trajectory big data,” Cluster Comput, vol. 18, no. 2, pp. 549–562, Jun. 2015, doi: 10.1007/s10586-014-0413-9.

[20] H. Hassanpour, A. H. Hamedi, P. Mhaskar, J. M. House, and T. I. Salsbury, “A hybrid clustering approach integrating first-principles knowledge with data for fault detection in HVAC systems,” Comput Chem Eng, vol. 187, p. 108717, Aug. 2024, doi: 10.1016/j.compchemeng.2024.108717.

[21] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, “A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins,” Geochemistry, vol. 84, no. 2, p. 126094, May 2024, doi: 10.1016/j.chemer.2024.126094.

[22] P. Lalwani, H. Banka, and C. Kumar, “CRWO: Clustering and routing in wireless sensor networks using optics inspired optimization,” Peer Peer Netw Appl, vol. 10, no. 3, pp. 453–471, May 2017, doi: 10.1007/s12083-016-0531-7.

[23] G. Iván and V. Grolmusz, "On dimension reduction of clustering results in structural bioinformatics," Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, vol. 1844, no. 12, pp. 2277–2283, Dec. 2014, doi: 10.1016/j.bbapap.2014.08.015.

[24] R. Roy, "Optics clustering (intro)," Medium. Accessed: Jul. 19, 2024. [Online]. Available: https://bobrupakroy.medium.com/optics-clustering-intro-76dcdaf94bde#:~:text=Disadvantages%3A&text=It%20fails%20if%20there%20are%20no%20density%20drops%20between%20clusters.&text=It%20is%20also%20sensitive%20to,parameter%20settings%20require%20domain%20knowledge

[25] Z. Huang and M. K. Ng, "A Note on K-modes Clustering," J Classif, vol. 20, no. 2, pp. 257–261, Sep. 2003, doi: 10.1007/s00357-003-0014-4.

[26] D. Kamthania, A. Pawa, and S. Madhavan, "Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm for E-Commerce Business," Journal of Computing and Information Technology, vol. 26, no. 1, pp. 57–68, 2018, doi: 10.20532/cit.2018.1003863.

[27] N. Sharma and N. Gaud, "K-modes Clustering Algorithm for Categorical Data," Int J Comput Appl, vol. 127, no. 17, pp. 1–6, Oct. 2015, doi: 10.5120/ijca2015906708.

[28] L. Li, R. J. Hansman, R. Palacios, and R. Welsch, "Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring," Transp Res Part C Emerg Technol, vol. 64, pp. 45–57, Mar. 2016, doi: 10.1016/j.trc.2016.01.007.

[29] Y. Balakrishna, S. Manda, H. Mwambi, and A. van Graan, "Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models," Front Nutr, vol. 10, Oct. 2023, doi: 10.3389/fnut.2023.1186221.

[30] G. Jouan, A. Cuzol, V. Monbet, and G. Monnier, "Gaussian mixture models for clustering and calibration of ensemble weather forecasts," Discrete and Continuous Dynamical Systems - S, vol. 16, no. 2, pp. 309–328, 2023, doi: 10.3934/dcdss.2022037.

[31] C. O'Sullivan, "Identifying Restaurant Hotspots with a Gaussian Mixture Model," Towards Data Science. Accessed: Jul. 08, 2024. [Online]. Available: https://towardsdatascience.com/identifying-restaurant-hotspots-with-a-gaussian-mixture-model-2a840ab0c782

[32] Amy, "Gaussian Mixture Model (GMM) for Anomaly Detection," GrabNGoInfo. Accessed: Jul. 08, 2024. [Online]. Available: https://medium.com/grabngoinfo/gaussian-mixture-model-gmm-for-anomaly-detection-e8360e6f4009

[33] M.-H. Yang and N. Ahuja, "Gaussian mixture model for human skin color and its applications in image and video," M. M. Yeung, B.-L. Yeo, and C. A. Bouman, Eds., Dec. 1998, pp. 458–466. doi: 10.1117/12.333865.

[34] D. J. Bora and A. K. Gupta, "A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm," International Journal of Computer Trends and Technology, vol. 10, no. 2, pp. 108–113, 2014, Accessed: Jul. 19, 2024. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1404/1404.6059.pdf

[35] Q. Cai, Z. Xue, D. Mao, H. Li, and J. Cao, "Bike-Sharing Prediction System," 2016, pp. 301–317. doi: 10.1007/978-3-319-40259-8_27.

[36] L. Wang, K. Zheng, X. Tao, and X. Han, "Affinity propagation clustering algorithm based on large-scale data-set," International Journal of Computers and Applications, vol. 40, no. 3, pp. 1–6, Jul. 2018, doi: 10.1080/1206212X.2018.1425184.

[37] W. Huang and Y. Li, "Application of Affinity Propagation Clustering Method in Medium and Extended Range Forecasting of Heavy Rainfall Processes in China," Atmosphere (Basel), vol. 13, no. 5, p. 768, May 2022, doi: 10.3390/atmos13050768.

[38] D. Dey, "Affinity Propagation in ML | To find the number of clusters. GeeksforGeeks." Accessed: Jul. 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/affinity-propagation-in-ml-to-find-the-number-of-clusters/

A Comprehensive Study of BIM for Infrastructural Crack Detection and the Vital Strategies

Samuel Adekunle, Opeoluwa Akinradewo, Babatunde Ogunbayo, Andrew Ebekozien, Clinton Aigbavboa
Cidb Centre of Excellence, University of Johannesburg, South Africa

Abstract—**Building information modelling is one of the emerging technologies in the construction industry and is relevant to its productivity and efficiency. Application which affects the product and process of the industry. An underdeveloped area with less attention is its adoption for crack detection and visualisation for infrastructural maintenance. This study provides a thorough perspective on BIM adoption for crack detection and visualisation. It also identified the different strategies that can aid in the adoption and use of BIM for infrastructural monitoring and maintenance in South Africa. The study adopted a quantitative approach, and questionnaires were distributed to industry professionals through an online platform. The collected data was analysed. The results indicate a need for incorporation of this aspect into the HEI curriculum and a teaching approach that is practical and experimental to be adopted.**

Keyword—Developing country; emerging technology; facility management; visualisation; emerging technology; South Africa

I. Introduction

Structural and civil infrastructures lose performance and deteriorate over time; detecting and visualising defects such as cracks on infrastructures is critical to ensure proper maintenance and predict possible failures [1]. Several infrastructure failure accidents in South Africa are related to insufficient crack inspection and condition assessment. For example, a coal silo that collapsed at the Majuba power station in Mpumalanga due to cracking and poor maintenance contributed to excessive power cuts, which negatively impacted the country's economy [2; 3]. Crack detection is part of infrastructure maintenance plans carried out regularly to monitor the health of infrastructures. Visualisation is typically used to sort large images and video data visually to improve crack inspection [4]. Infrastructure cracks are the earliest indication of material deterioration and possible infrastructure failures [5]. Crack detection information can be used to diagnose and guide the appropriate maintenance methods and approaches to prevent catastrophic failures [6]. Crack detection using visual inspections is cumbersome, costly, and inefficient [7].

The traditional methods for crack detection are visual inspections, which are conducted utilising manual observations by maintenance inspectors and engineers. Over the years, several disadvantages have been noted from manual crack detection observations. The disadvantages include time-consuming, expensive, and inaccuracy due to irregular conditions and human errors. In most cases, the reliability and accuracy of the method depend on the specialist's knowledge and experience [5, 8]. Some cracks on engineering structures can only be detected at a microscopic level, which is difficult to detect with manual inspection methods. Therefore, as clearly stated by [6, 8], manual crack detection methods are not economical and have lower accuracy levels. With the increasingly complex engineering infrastructures being developed in South Africa, there is a need to adopt new automated crack detection methods.

Over the years, reliable surface crack detection methods have been developed to effectively detect cracks on infrastructure surfaces. Automated crack detection methods are fast and reliable; hence, they boost the productivity of detecting cracks in surface structures [6]. Automatic crack-detecting methods, such as image processing techniques, have gained popularity in recent years, where crack information is extracted from images for analysis. According to study [9], some of these image processing techniques also produce unsatisfying results, as the crack analysis by computer software depends on the quality of images and the characteristics of surfaces.

The adoption and implementation of BIM have gained popularity over the years, and it has been used for planning and design. However, in recent years, BIM applications have expanded to be used in clash detection, carbon capture, and asset management [10]. The study conducted by study [11] noted that it is an area that requires attention.

II. Development of Cracks in South African Infrastructures

According to the South African Institution of Civil Engineering (SAICE) report, a significant portion of South Africa's infrastructure is in poor condition. Roads, bridges, healthcare facilities, and water management structures have surface cracks [12]. Streets and buildings across the country are cracking due to various factors, and there is no indication of proper maintenance systems and plans in place, as this problem of cracks on infrastructures has worsened over the years [13; 14; 15]. The research in study [4] assessed the adoption of computer vision-based models such as BIM to detect defects in infrastructure regularly as part of maintenance. It has been noted that cracks are the most common defects resulting in infrastructure deterioration. Therefore, adopting and implementing effective crack detection methods such as BIM is critical to visualising cracks and improving the state of infrastructures in South Africa.

One of the recent disasters caused by undetected surface cracks in South Africa involves an incident in one of the largest power stations operated by ESKOM in Mpumalanga. In November 2014, a concrete coal silo collapsed at the Majuba power station due to cracks that were not detected on the surface of the silo as part of regular maintenance of the power station. The power station at the time of the incident supplied about 10% of the country's electricity. Hence, the disaster contributed to the continued rolling blackouts of load shedding in South Africa. During the incident investigations, cracks were discovered on other silos resulting in the complete shutdown of the power station [16; 3]. Concrete silos are some of the most critical structures in the industrial sector; they are normally used to store bulk solids such as coal. Therefore, periodic inspections for signs of distress, such as cracks on these structures, are critical in their lifecycle [2]. However, concrete silos are high-rise structures with limited access to external surfaces. Therefore, it's not feasible to detect cracks on these structures using manual inspections. Hence, adopting advanced methods such as BIM for crack detection in the South African construction industry might be a solution to prevent disasters, such as using BIM applications to maintain infrastructures. BIM applications have the potential to detect and visualise cracks in infrastructures. Therefore, the current study assesses the adoption level of BIM for crack detection and visualisation within the South African construction industry as part of continuous progress toward full integration of BIM.

III. Role of BIM in Infrastructure Maintenance

BIM is a technology that holds immense significance in construction projects and infrastructure maintenance. Its benefits range from better communication and collaboration to enhanced efficiency and reduced errors. Despite these advantages, many countries have yet to adopt BIM for post-construction maintenance fully. By leveraging BIM technology, stakeholders can ensure seamless infrastructure maintenance and upkeep, promoting safe infrastructure maintenance practices [4]. The critical role of BIM in infrastructure maintenance lies in the visualisation of geometrics and the integration of infrastructure information into 3D objects. Hence, defects such as cracks can be detected effectively and accurately, and the damage caused can be assessed.

Detection of cracks in infrastructure such as buildings, roads, and bridges is crucial for monitoring their structural integrity and health [6]. A crack is a fracture wherein the components or parts are not entirely separated [17]. It has been noted in the studies by [5; 6] that cracks on most infrastructures are early indications of deterioration. Cracks create access to the harmful and corrosive chemical that penetrates infrastructures, causing damage to their integrity, strength, and durability [18]. Considerable research has been undertaken into detecting cracks to address the issue of infrastructures developing cracks and to monitor their physical and functional conditions [6]. Incorporating crack detection into maintenance plans is a crucial aspect of the construction industry, particularly before infrastructure upgrades, reconstructions, or repairs. Doing so can help ensure the infrastructure's longevity and safety while mitigating the risk of more extensive and costly repairs down the line. As such, it is highly recommended that those involved in the construction industry prioritize crack detection as part of their overall maintenance strategy.

According to study [19], civil infrastructure must be regularly checked for structural integrity as it nears the end of its lifespan. The most critical checks are for crack detection, which uses various approaches and models. The methods and models for crack detection are regularly improved and updated to ensure accuracy and performance in the construction industry. Digital technologies and the 4th industrial revolution have developed computer-based models for crack detection. Most of these models are Computer-based models developed using digital technologies for crack detection, many of which rely on computer vision-based techniques [5]. It has been highlighted by [19] that computer vision-based models improve the accuracy of crack detection with an average precision of 95%. As such, there is an increasing trend of adopting computer vision-based models to boost the productivity of detecting cracks in structures [4]. Therefore, the South African construction industry stakeholders need to adopt computer-based models for crack detection to improve the country's infrastructure maintenance and state.

IV. Benefits of Adopting BIM for Crack Detection and Visualisation

Building Information Modeling (BIM) has become an increasingly popular tool for infrastructure

Projects, particularly in detecting and measuring cracks. This approach offers a comprehensive and sophisticated solution that considers all aspects of the cracks, including their length, width, and depth [5]. BIM also generates 3D images that allow engineers and stakeholders to visualize the patterns of the cracks, which is extremely helpful in determining the most effective repair strategies [18]. BIM enables easier detection of structural cracks, thus facilitating the prediction of future conditions and determining necessary maintenance plans. As highlighted by [20], this technology also effectively allocates repair resources. By harnessing the power of BIM, organizations can optimize their maintenance strategies and ensure the longevity and safety of their buildings. Another significant advantage of employing BIM is accurately estimating the costs involved in repairing the cracks. Considering all relevant factors, such as materials, labour, and equipment costs, BIM can provide a detailed and reliable estimate of the expenses involved in repairing the cracks [21]. This information is crucial for project managers and stakeholders, allowing them to budget effectively and plan accordingly. Furthermore, BIM can aid in the formulation of effective infrastructure maintenance plans. By identifying the locations and severity of cracks, engineers can prioritize repairs and maintenance activities, ensuring that critical infrastructure is well-maintained and safe for use [22].

V. Case Studies on BIM Implementation for Crack Detection and Visualisation

In this section, studies related to adopting BIM for crack detection are reviewed and discussed to assess the potential of implementing BIM for crack detection in South Africa.

A. *Automated Concrete Defect Detection Using Building Information Model in Hong Kong*

The study by [23] adopted BIM for automated concrete defect detection. The 3D model was developed to detect the position and geometries of the concrete defect to guide subsequent maintenance stages. BIM implementation in this study was to overcome the disadvantage of manual inspection methods and assess high-rise buildings with limited access to external structures. The approach that was used in the study involved using a drone to take aerial photographs of a 10-story residential building near the University of Hong Kong. The 3D reconstruction was conducted to generate a defect point cloud (DPC) visualised in BIM to detect defects. The results from the study indicated that defect detection errors from the manual visualisation method had been reduced by up to 14%.

It is evident from this study that BIM adoption for crack detection improves the reliability of defect detection results; hence, proper maintenance planning can be done on the building. The whole process is much faster than the manual visual inspection of cracks. However, the process could be more precise; hence, proper training and understanding of BIM are needed before undertaking crack detection procedures.

B. *Adoption of BIM for the Inspection of Monte da Virgem Telecommunications Tower in Portugal*

The research in [24] also explored the potential of BIM applications in crack detection by studying detecting cracks on a telecommunications tower that is 177m high in Portugal. In the study, a remote inspection of reinforced concrete was conducted using Unmanned Aerial Vehicles (UAVs), referred to as drones, to capture tower images from different angles and integrate these images into the building information modelling (BIM).

The photographs of the tower were collected in a photographic scan from the bottom of the tower to the top, capturing concrete surface conditions. Captured photographs were then processed and reconstructed into 3D modelling, which was integrated into BIM. The biological colonies are indications of cracks, a BIM-generated model that detects cracks in the tower.

Cracks on structures such as telecommunications towers with poor accessibility to external structures are not easily detected using manual visual inspection methods due to their long heights; as such, only automated technological methods such as BIM can detect cracks on their surfaces. The BIM model has provided efficient and reliable results for cracks in the tower. Hence, informed decisions can be made on the maintenance of the tower. This is a good case study that can be related to the case of a coal silo that collapsed at the Majuba power station in Mpumalanga, South Africa, due to cracking, which may be resulted from limited access to the coal silo to detect surface cracks on the structure. The new technology of detecting cracks improves infrastructure maintenance; hence, South Africa needs to adopt it.

VI. Methodology

A quantitative approach was adopted to identify the strategies required to implement BIM adoption for crack detection in the South African construction industry. A total of 77 responses were retrieved and adopted for the study. The research instrument was randomly distributed online to professionals in the South African construction industry via Google Forms. The Cronbach alpha coefficient has been utilised to evaluate the quality and consistency of the data obtained from the study. This statistical method is widely used in research to measure the internal consistency of a set of variables. The coefficient's value ranges from 0 to 1, with 1 signifying perfect internal consistency and 0 indicating no consistency. The higher the coefficient, the more reliable the data. A Cronbach alpha coefficient of 0.85, which is a higher value, indicates that the data collected is consistent and, hence, reliable to be adopted to improve the adoption of BIM for crack detection. The adopted methodology has been used by studies in the construction industry [25, 26].

A. *Background Information of Respondents*

The study's findings revealed that most respondents held a bachelor's degree, the typical entry-level degree (see Table I). Many respondents possessed a master's degree, while only a few held a doctoral degree. This suggests that most South African construction industry professionals do not pursue education beyond the entry-level degree. Many might have professional degrees, which were not requested in this study. Therefore, there may be a need for greater emphasis on continuing education and professional development to ensure professionals remain current with trends and best practices. The study involved eight professional categories with distinct affiliations. Most respondents belonged to the engineering and construction project management categories, while fewer participants were from the construction health and safety category. The findings indicate that engineers and construction project managers in the South African construction industry are highly interested in staying updated with advancements and technologies in infrastructure maintenance, actively seeking ways to improve their knowledge and skills. The years of experience is crucial in determining industry knowledge and exposure. The study revealed that most respondents have six to 10 years of experience, with fewer having over 20 years. This suggests the industry is dominated by professionals with moderate experience, with relatively few highly experienced professionals. However, it is expected that more experienced professionals are less inclined to adopt new technologies, such as BIM, for crack detection and visualization, compared with young professionals who are technology enthusiasts. The data also shows that many respondents are employed in the government and consultancy sectors.

TABLE I. RESPONDENTS BACKGROUND INFORMATION

Category	Subcategory	Percentage (%)
Educational Qualification	Bachelor's Degree	53.2
	Master's Degree	22.1
	Diploma	18.2
	Doctoral Degree	6.5
Professional Affiliation	Civil Engineer	29.9
	Construction Project Manager	22
	BIM Professional	13
	Quantity Surveyor	11.7
	Construction Manager	10.4
	Architect	6.5
	Construction Supervisor	3.9
	Construction Health and Safety Officer	2.6
Years of Experience	1-5 years	10.5
	6-10 years	31.2
	11-15 years	26
	16-20 years	16.9
	Over 20 years	1.2
	Under 1 year	14.2

VII. FINDINGS

The study result is presented in this section

A. *Strategies for the Adoption of BIM for Crack Detection*

The findings from the descriptive analysis indicate that several strategies could help successfully adopt and implement BIM for crack detection in the SACI. The most effective strategy in the SACI context includes universities teaching BIM tools and applications to students interested in pursuing a career in the construction industry. By incorporating BIM into the curriculum, students will gain an in-depth understanding of its benefits and grow their skills in utilizing BIM to detect cracks. Secondly, raising awareness about the benefits of using BIM for detecting cracks in the construction industry could encourage stakeholders to adopt this technology. This approach could involve creating informative materials or hosting workshops to educate professionals and decision-makers on the potential of BIM in detecting cracks. Thirdly, promoting BIM research for crack detection could help advance the technology and improve its effectiveness in detecting cracks. This could involve funding research projects exploring the construction industry's latest BIM applications and tools. Training professionals in BIM crack detection tools is crucial for successfully implementing this technology. By providing hands-on training and support, professionals will be equipped with the necessary skills and knowledge to apply BIM in detecting cracks in the construction industry. The effectiveness of this strategy has been highlighted in different studies, which include [25, 27]. The study suggests that the adoption and implementation of BIM for crack detection in the SACI requires a multifaceted approach that involves teaching BIM tools and applications at universities, raising awareness of the benefits of BIM for detecting cracks, promoting BIM research for crack detection, and training professionals in BIM crack detection tools. Strategies for BIM is shown in Table II.

TABLE II. STRATEGIES

Strategies	Mean	Std. Deviation	Rank
Teaching BIM tools and applications used to detect cracks in the University	4.45	0.770	1
Raise awareness of the benefits of detecting cracks using BIM	4.43	0.637	2
Promoting research on adopting BIM for crack detection and visualisation	4.41	0.807	3
Promoting wider BIM adoption in infrastructure maintenance	4.40	0.658	4
Training and development of construction and maintenance professionals on BIM crack detection tools	4.38	0.673	5
Implementation of BIM throughout the lifecycle of government facilities	4.29	0.830	6
Acquiring of BIM crack detection software and tools by construction organisations	4.26	0.755	7
Development of national standards for local implementation of BIM for crack detection	4.23	0.809	8
Mandatory adoption of BIM for crack detection on government projects	4.06	0.978	9
Construction organisations top management formally approving the use of BIM for crack detection	4.01	0.993	10
Tax relief to motivate adoption among construction stakeholders	3.65	1.295	11
Financial incentives to fund the implementation of BIM for crack detection	3.62	1.257	12
Organisations utilizing BIM for crack detection to receive tax breaks	3.55	1.341	13

VIII. Conclusion

The study's findings on BIM adoption strategies have significant implications for the South African construction industry (SACI). BIM implementation is a complex process involving multiple stages and requiring careful consideration of various factors. Its success depends on a well-planned and structured approach that addresses the country's unique infrastructure requirements. This involves a thorough analysis of infrastructure defects, particularly cracks. Effective training is crucial for successful BIM implementation, enabling SACI professionals to acquire the necessary skills and knowledge to use the technology effectively. BIM adoption can lead to improved infrastructure maintenance, enhanced collaboration, and more sustainable and efficient crack defect repair practices if implemented correctly. BIM is an emerging technology with great potential in the construction industry for various applications, including crack detection. Many countries have proposed and implemented strategies to adopt BIM, as highlighted in the reviewed literature. However, the effectiveness of these strategies varies due to the unique requirements and demands of each country's construction industry. In the South African context, the analyzed data revealed that the most effective strategies for BIM adoption for crack detection are training and education. By investing in training and educating construction professionals on BIM applications, companies can ensure that employees have the necessary skills to implement this technology effectively. This can lead to improved efficiency, reduced costs, and increased safety in construction projects. Further studies can be conducted to evaluate the effectiveness of various BIM training programs and their impact on professionals' proficiency and project success. Also, researchers should explore the integration of BIM with other emerging technologies like AI and IoT for enhanced crack detection and infrastructure maintenance.

References

[1] Özgenel, Ç.F. and Sorguç, A.G., 2018. Performance comparison of pre-trained convolutional neural networks on crack detection in buildings. In Isarc. proceedings of the international symposium on automation and robotics in construction (Vol. 35, pp. 1-8). IAARC Publications.

[2] Chavez Sagarnaga, J. and Carson, J.W., 2018. Beyond Silo Failures: Legal Implications and Lessons Learned. In Forensic Engineering 2022 (pp. 922-931).

[3] Nowakowska, M. and Tubis, A., 2015. Load shedding and the energy security of the Republic of South Africa. Journal of Polish Safety and Reliability Association, 6(3).

[4] Koch, C., Georgieva, K., Kasireddy, V., Akinci, B. and Fieguth, P., 2015. A review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure. Advanced Engineering Informatics, 29(2), pp.196-210.

[5] Mohan, A. and Poobal, S., 2018. Crack detection using image processing: A critical review and analysis. Alexandria Engineering Journal, 57(2), pp.787-798.

[6] Hoang, N.D., 2018. Detection of surface crack in building structures using image processing technique with an improved Otsu method for image thresholding. Advances in Civil Engineering, 2018.

[7] Perry, B.J., 2019. A streamlined bridge inspection framework utilising Unmanned Aerial Vehicles (UAVs) (Doctoral dissertation, Colorado State University).

[8] Andrushia, A.D., Anand, N. and Arulraj, G.P., 2021. Evaluation of thermal cracks on fire-exposed concrete structures using Ripplet transform. Mathematics and Computers in Simulation, 180, pp.93-113.

[9] Kim, H., Ahn, E., Cho, S., Shin, M. and Sim, S.H., 2017. Comparative analysis of image binarisation methods for crack identification in concrete structures. Cement and Concrete Research, 99, pp.53-61.

[10] O'Shea, M. and Murphy, J., 2020. Design of a BIM-integrated structural health monitoring system for a historic offshore lighthouse. Buildings, 10(7), p.131.

[11] Salzano, A., Parisi, C.M., Acampa, G. and Nicolella, M., 2023. Existing assets maintenance management: Optimizing maintenance procedures and costs through BIM tools. Automation in Construction, 149, p.104788.

[12] Rust, F.C., Wall, K., Smit, M.A. and Amod, S., 2021. South African infrastructure condition-an opinion survey for the SAICE Infrastructure Report Card. Journal of the South African Institution of Civil Engineering, 63(2), pp.35-46.

[13] Stone, T., 2016. Analysis of pipe deterioration: ageing water & sanitation infrastructure. IMIESA, 41(7), pp.17-21.

[14] Maboa, D.T., 2019. A review of factors influencing the reliability of railway infrastructure (Doctoral dissertation, University of Johannesburg (South Africa).

[15] Jooste, N.J.J., 2011. A consultant's perspective on the use of bitumen rubber especially double seals. SATC 2011.

[16] Memka, D. and Lekhanya, L.M., 2017. Technological challenges influencing the implementation of green energy in the SME sector in KwaZulu-Natal (KZN). Environmental economics, (8, Iss. 3 (cont.)), pp.157-164.

[17] Wiggenhauser, H., Köpp, C., Timofeev, J. and Azari, H., 2018. Controlled creating of cracks in concrete for non-destructive testing. Journal of nondestructive evaluation, 37, pp.1-9.

[18] Adhikari, R.S., Moselhi, O. and Bagchi, A., 2014. Image-based retrieval of concrete crack properties for bridge inspection. Automation in construction, 39, pp.180-194.

[19] Mohammed, M.A., Han, Z. and Li, Y., 2021. Exploring the detection accuracy of concrete cracks using various CNN models. Advances in Materials Science and Engineering, 2021, pp.1-11.

[20] Ai, D., Jiang, G., Lam, S.K., He, P. and Li, C., 2023. Computer vision framework for crack detection of civil infrastructure. A review. Engineering Applications of Artificial Intelligence, 117, p.105478.

[21] McGuire, B.M., 2014. Using building information modeling to track and assess the structural condition of bridges (Doctoral dissertation, Colorado State University.

[22] Adhikari, R.S., Moselhi, O. and Bagchi, A., 2014. Image-based retrieval of concrete crack properties for bridge inspection. Automation in construction, 39, pp.180-194.

[23] Chen, J., Lu, W. and Lou, J., 2022. Automatic concrete defect detection and reconstruction by aligning aerial images onto a semantic-rich building information model. Computer-Aided Civil and Infrastructure Engineering.

[24] Ribeiro, D., Santos, R., Shibasaki, A., Montenegro, P., Carvalho, H. and Calçada, R., 2020. Remote inspection of RC structures using unmanned aerial vehicles and heuristic image processing. Engineering Failure Analysis, 117, p.104813.

[25] Memon, A.H., Rahman, I.A., Memon, I. and Azman, N.I.A., 2014. BIM in the Malaysian construction industry: status, advantages, barriers, and strategies to enhance the implementation level. Research Journal of Applied Sciences, Engineering, and Technology, 8(5), pp.606-614.

[26] Adekunle, S., Aigbavboa, C., Akinradewo, O., Ikuabe, M. and Adeniyi, A., 2022. A principal component analysis of Organisational BIM Implementation. Modular and Offsite Construction (MOC) Summit Proceedings, pp.161-168.

[27] Ikuabe, M.; Aigbavboa, C.; Akinradewo, O.; Adekunle, S.; Adeniyi, A. Hindering Factors to the Utilisation of UAVs for Construction Projects in South Africa. Modul. Offsite Constr. Summit Proc. 2022, 154–160, doi:10.29173/MOCS277.

[28] Calitz, S. and Wium, J.A., 2022. A proposal to facilitate BIM implementation across the South African construction industry. Journal of the South African Institution of Civil Engineering, 64(4), pp.29-37.

The Adoption of Electronic Payments in Online Shopping: The Mediating Role of Customer Trust

Nguyen Thi Phuong Giang[1*], Thai Dong Tan[2], Le Huu Hung[3], Nguyen Binh Phuong Duy[4]
Faculty of Commerce – Tourism, Industrial University of Ho Chi Minh City – IUH, Ho Chi Minh City, Viet Nam[1,2,3,4]

***Abstract*—This study investigates the factors influencing electronic payment in online shopping behavior among Ho Chi Minh City consumers. With the rapid advancement of technology, e-commerce has become a new trend, and understanding the intention to adopt electronic payment is crucial for online businesses. The research employs quantitative and qualitative methods, utilizing a survey of 437 Ho Chi Minh City consumers. The data collected is processed using SPSS 24 and SmartPls4 software. Eight factors related to consumers' intention to use electronic payment are identified: social influence, security, perceived usefulness, convenience, ease of use, customer trust, perceived risk, and performance expectancy. The study's findings will contribute to the existing knowledge base for businesses, facilitating the promotion of electronic payment adoption. This support will aid businesses in developing more attractive online sales strategies, encouraging consumers to shop and pay online more frequently and, at the same time, contribute to supporting departments in formulating policies for digital payments, thereby promoting national digital transformation.**

Keywords—Electronic payment, Intention to use, Online shopping

I. INTRODUCTION

In the current era, our society is experiencing profound transformations, both in its structure and quality of life. The rapid advancement of information technology has paved the way for various industries to flourish, and among them, e-commerce stands out as a sector that has made remarkable strides, bringing greater convenience and modernity to human existence. A direct consequence of this progress is the widespread adoption of electronic payments, which has become an integral part of modern life. The outbreak of the COVID-19 pandemic has impacted various sectors of the economy and changed customer shopping habits. Increasingly, businesses and individuals are adapting to the new situation by shifting from in-person transactions to online shopping and using cashless payment methods. In this context, most commercial banks are focusing on and implementing plans to develop cashless payment services in line with government directives. The convergence of the digital economy and the digital society, combined with the disruptive impact of the COVID-19 pandemic, has accelerated the global and Vietnamese shift towards cashless payments. The State Bank of Vietnam's 2021 report indicated a significant surge in the value of non-cash payments, rising by an impressive 18%, and a substantial 30% increase in payment volume [1]. Furthermore, the Digital 2022 report disclosed that the number of internet users in Vietnam has surpassed 72 million, accounting for an impressive 73.2% of the population. This figure marks a substantial 4.9% growth compared to the previous year. Digital payments are at the forefront of technological advancements and consistently ensure security. In comparison to 2021, non-cash payments in 2022 surged by 31.39% in value and accounted for 85.6% of the total transactions [2]. Online payment activities in Ho Chi Minh City are facing significant challenges. Infrastructure and technology issues are prominent, with unstable internet connectivity and uneven technology proficiency hindering the adoption of advanced payment technologies. Concerns about security and privacy pose a major barrier. Worries about personal information security and data breaches make many users hesitant to fully embrace online payments. Consumer habits and behaviors also present a considerable obstacle, as the preference for cash use remains common due to long-standing habits and trust in traditional methods. With its advantages of convenience and safety, digital payment is driving the growth of commerce worldwide. Digital payments facilitate swift transactions for both sellers and consumers, minimizing risks, and guaranteeing mutual benefits. Moreover, it caters to diverse transactional needs, ranging from simple to complex, at low service fees. Digital payments are expected to remain the preferred payment method in the present and future. In Vietnam, electronic payments were introduced in 2008, with e-wallets being the pioneering model. However, to this day, online payment activities in Vietnam, especially in Ho Chi Minh City, face several obstacles due to various factors influencing consumers' decisions to adopt this payment method. Hence, researching the influencing factors on the intention to use electronic payments for online purchases among the city's residents is essential to determine their interests and the factors impacting their acceptance of this service. The contributions of the research will be incorporated into existing documents for businesses when successfully implementing digital payments. This helps merchants develop more sales strategies, encouraging consumers to shop and pay online more frequently.

II. THEORETICAL BASIC

A. Concept

Electronic payment refers to any monetary transaction initiated in connection with electronic communication methods. It is a form of using electronic signals directly linked to a deposit account or credit account [3], [4]. Electronic payment (e-payment) serves as the foundation of Internet Banking, encompassing the online platform that facilitates various activities such as online auctions, Internet stock trading, and online shopping [5]. The concept of intention to use delves into an individual's level of willingness and readiness to engage in a particular behavior [6].

B. Theoretical Model

In 1989, Davis proposed the Technology Acceptance Model (TAM) in collaboration with several researchers, providing a theoretical framework to explain technology usage. According to this theory, two crucial factors, perceived usefulness and ease of use, play a significant role in determining individuals' acceptance and adoption of technology [7]. However, with the advancement of research, a more comprehensive model called the Unified Theory of Acceptance and Use of Technology (UTAUT) was formulated by Venkatesh and colleagues in 2003 [8]. UTAUT integrated and expanded upon factors from previous models, offering a more holistic perspective on technology acceptance. These facilitating factors can encompass various resources like data, knowledge, documentation, financial support, or other forms of assistance. The UTAUT model has significantly enhanced our understanding of the technology adoption process by considering a broader range of influencing factors.

III. Hypotheses and Research Model

A. Customer Trust

Trust plays a pivotal role in shaping individuals' intentions to carry out transactions, especially in the realm of electronic commerce. When people place their trust in the process, they are more inclined to participate in e-commerce transactions [9]. Moreover, trust is a subjective perception regarding fulfilling obligations and meeting the expectations of all parties involved [10]. Trust is essential in laying the groundwork for successful transactions; reputation plays a crucial role in establishing trust, while trust influences the perception of the transactional environment [11]. Furthermore, trust encompasses attributes such as integrity, security, and data protection during transactions [12]. Cryptocurrencies, which utilize enhanced encryption methods, increase their reliability and trustworthiness. Transparent and objective verification methods used in cryptocurrencies eliminate the need for intermediaries, reducing transaction fees and increasing reliability in the system, thus enhancing its acceptance [13]. Research in [14], [15] has shown that higher levels of trust in online vendors directly correlate with increased adoption rates of digital payment systems. The widespread adoption of PayPal, is largely attributed to its strong reputation for security and buyer protection.

H1: Customer trust positively affects the intention to use electronic payment methods.

B. Perceived Risk

Perceived risk in the context of this study refers to the sense of uncertainty individuals experience regarding potential negative outcomes when utilizing cryptocurrencies for electronic payments. It plays a significant role in influencing perceived utility and purchase intention in commerce [16], [17]. Cryptocurrencies are renowned for their decentralized payment network; however, the lack of clear operational assurance and the volatile nature of their value function hinder their widespread adoption [18]. The unstable prices make it challenging for cryptocurrencies to serve as a reliable unit of account, and concerns about regulation, speculative activities, and vulnerability to cyberattacks further impede their establishment as a global currency [19]. The existence of such risks affects users' trust, as they fear negative consequences arising from unsuccessful transactions, insufficient e-commerce regulations, and inadequate anti-fraud measures employed by sellers [20]. If digital payment is associated with significant risks, the user's evaluation may deteriorate, and consequently, their level of trust in using digital payments for purchasing goods may decrease considerably. Previous research in various contexts has already demonstrated the adverse impact of risk on trust [21], [22]. In this study, trust serves as an intermediary between perceived risk and the intention to use digital payments, as established in prior research [23], [24], [25]. When users trust a provider or store, they believe the business will act responsibly and, thereby, reduce perceived risk for customers. Hence, trust plays a mediating role between perceived risk and customers' intention to use digital payments. Previous research [26], [27] in various contexts has already demonstrated that perceived risks negatively affect users' trust and their willingness to engage in electronic transactions. In practice, the failure of certain cryptocurrency exchanges due to hacking incidents has significantly diminished user trust and adoption rates.

H2: The use of digital payments is negatively influenced by perceived risks.

H3: Customer trust is negatively impacted by perceived risks.

H4: Perceived risks hurt the intention to use digital payments, mediated by customer trust.

C. Perceived Ease of Use

According to existing literature, the perceived ease of use holds significant importance in influencing customers' intention to adopt new technologies. For a product to be perceived as easy to use, it should have a user-friendly interface and follow a logical sequence of actions [28]. Davis (1989) defined ease of use as customers' perception that the system is uncomplicated, effortless, and quick to navigate [7]. Indicators of perceived ease of use include transparency and comprehensibility, step-by-step installation guides, and user-friendly features, as well as easy comparisons between cash payment systems and third-party e-payment methods [29], [30], [31]. Previous research has indicated that when customers perceive a device as requiring minimal mental and physical effort, its usage becomes more prominent [32]. In our study, we expand this concept to encompass the social aspect of e-payment usage, intending to enrich interactions and outcomes in electronic commerce [33]. Aghdaie et al. (2011) conducted a study that showed a positive relationship between ease of use and trust [34]. This finding is consistent with earlier research that also found a positive correlation between customer trust and perceived ease of use [35], [11], [36], [37]. Users tend to believe that newer technologies are capable of bringing benefits to the products they will purchase, so consumers expect these technologies to be easy to learn and use. Al-Sharafi et al. (2017) explored the role of trust in increasing awareness of ease of use during positive online transactions, although the effect is not significant [38]. Yudiarti and Puspaningrum (2022) confirmed that perceived ease of use, mediated by trust, influences the intention [39]. When companies apply technology or systems perceived as easy to use, it leads to increased trust and, consequently, a higher intention to use them in the future. Similarly, Wen et al. (2011) explained that perceived ease of use has a significantly positive impact on trust [40]. This research supports the study results by [41], [42], which stated that showing the results of trust mediates the relationship between perceived ease of use and intention to use.

H5: The intention to use digital payments is positively influenced by the perceived ease of use.

H6: Customer trust is positively impacted by the perceived ease of use.

H7: The perceived ease of use has a positive effect on the intention to use digital payments, mediated by customer trust.

D. Perceived Usefulness

The concept of perceived usefulness refers to individuals' belief that using a particular e-payment system can enhance their efficiency in completing financial and daily transactions [43]. For low-income customers, perceived usefulness is crucial in shaping their expectations and willingness to adopt e-payment for various purposes [44]. The utilization of e-payment is argued to enhance productivity and efficiency in payment-related processes [45], improve customer service and product information [46], and leverage digital infrastructure for information dissemination [47]. Low-income customers who perceive e-payment as useful are more likely to adopt and utilize the system to their advantage. On the other hand, if the perceived usefulness of e-payment is low, its implementation may not lead to adoption, as perceived usefulness plays a crucial role in the decision-making process [48]. The studies conducted by Chinomona (2013) and Amin et al. (2014) have revealed that trust is positively influenced by perceived usefulness [36], [49]. When customers perceive a system or new product as valuable and beneficial, they will trust the product or new technology. Perceived usefulness affects intention through trust [50]. Consumers who have a higher level of trust are more inclined to engage in purchasing behavior. Thus, it can be observed from the research that trust plays an intermediary role between perceived usefulness and customer behavioral intention. Previous studies have also recognized this indirect relationship, indicating that perceived usefulness affects consumer behavioral intention through trust [51]. Additionally, other research has shown a direct impact of perceived usefulness on trust [52], [39].

H8: Perceived usefulness has a positive impact on the intention to use digital payments.

H9: Perceived usefulness has a positive impact on customer trust.

H10: Perceived usefulness has a positive impact on the intention to use digital payments through customer trust.

E. Security

Perceived security plays a critical role in users' acceptance and trust in mobile payment systems. It refers to users' perception of the level of protection and prevention against potential security threats related to the use of M-wallets [53]. The primary concern for users when it comes to financial transactions is the security of their sensitive information and data [54]. Users need to have confidence that their personal information will remain secure and protected from unauthorized access, storage, or manipulation by any third parties [55]. Therefore, perceived security is a crucial factor in determining users' willingness to adopt and use mobile payment systems [56]. The relationship between security and trust is closely interconnected. Users are more likely to trust organizations and the digital payment system if they are assured about the security measures in place to protect their data and prevent any data breaches [57]. According to their research, the level of trust in digital payments is determined by security. Moreover, improving user trust encourages customers to use electronic payments when making online purchases. Flavián and Guinalíu (2006) stated that developing trust changes customers' usage intentions. Perceived security is linked to behavior through trust [54]. However, how users perceive the privacy of their information affects trust. According to C. Kim et al. (2010), a lack of trust in the security of online transactions can lead individuals to prefer using cash for transactions [58]. Based on these findings, trust is an intermediary factor between perceived security and user behavior [59].

H11: Security has a positive impact on the intention to use digital payments.

H12: Security has a positive impact on customer trust.

H13: Security has a positive impact on the intention to use digital payments through customer trust.

F. Social Influence

Social influence is “the degree to which an individual perceives that significant others believe he or she should use the new system” [8]. In the context of online shopping, individuals who engage in word-of-mouth (WOM) exchanges with others and receive referrals, opinions, personal experiences, and product-related knowledge are more likely to feel a higher sense of recognition of online vendors. They tend to pay attention to the recommendations and influence of others when making their purchasing decisions [60]. Morosan and DeFranco (2016) have shown that social influence plays a significant role in shaping users' intention to use mobile payment (M-payment) services [61]. It also affects users' attitudes towards M-payment services and their perception of the multiple benefits of using such services [62]. At the same time, it also has an impact on customer trust when there is an intention to shop online [63], [64], [65]. Trust can act as a mediator between social influence and users' intention to use digital payments, as demonstrated by the research conducted by Chan and Lee (2021) [66].

H14: Social influence has a positive impact on the intention to use digital payments.

H15: Social influence has a positive impact on customer trust.

H16: Social influence has a positive impact on the intention to use digital payments through customer trust.

G. Performance Expectancy

Performance expectancy (PE) plays a pivotal role in determining users' willingness to accept and adopt new technology. PE refers to the belief that using a specific technology will result in improved job performance or bring about various benefits in daily life [8]. People are more likely to embrace and utilize technology when they perceive it as providing substantial advantages and enhancements [67]. In the context of electronic payment systems, performance expectancy is associated with users' perceptions of the service's convenience, speed, and effectiveness [68]. If users believe that electronic payment offers them a convenient and efficient way to conduct banking transactions, they will be more inclined to use such channels [69].

H17: Positive effects of performance expectancy lead to a higher intention to use electronic payment.

H. Convenience

Convenience is a critical factor that influences users' decision to use a technology. It refers to the feeling of comfort and ease when using a particular system [7]. Convenience is the combination of time and place utility, and it exerts a significant impact on a user's choice to adopt a particular system. Consequently, convenience plays a crucial role in shaping the perceived value of the system [70]. Generally, people tend to place their trust in technology when it is designed to efficiently complete tasks and simplify processes. When a user of new technology experiences happiness, they will become interested and generate more satisfaction with the technology [7]. This implies that as users' comfort level increases, their usage behavior will improve. Therefore, the convenience of using it will positively influence the intention to use electronic payments.

H18: Convenience positively influences the intention to use electronic payments.

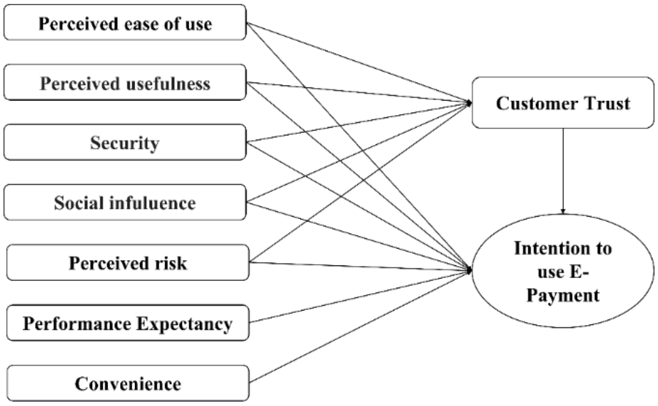


Fig. 1. Proposed research model

IV. Research Methods

The research is conducted in two phases: qualitative research and quantitative research. Choosing quantitative research methods for data collection brings many important benefits. This approach allows for precise measurement and utilizes statistical data to analyze relationships between variables. This helps in deriving conclusions that are highly accurate and applicable broadly in both theoretical and practical research. Furthermore, the objectivity of quantitative data ensures the reliability of research findings, facilitating the assessment of impacts and the significance of key variables within the study sample. The qualitative research phase aims to determine the measurement scale and lay the foundation for constructing the questionnaire in the quantitative research phase. The author inherited from previous studies and developed a preliminary measurement scale for seven independent variables (Social influence, Risk, Usefulness, Security, Performance expectancy, Ease of use, Convenience), one mediating variable (Trust), and the dependent variable (Intention to use e-payment). In the quantitative research phase, the survey method is employed, utilizing a questionnaire to collect data. The online survey form is distributed via email and various social media platforms. The questionnaire utilizes a 5-point Likert scale, allowing participants to express their level of agreement, ranging from “strongly disagree” to “strongly agree”. The choice of sampling method in quantitative research is crucial as it affects both the cost and quality of the study. There are two main groups of sampling methods: probability sampling, and non-probability sampling. Random sampling methods typically reflect the representation of the population and have high generalizability. In contrast, non-probability sampling methods allow for purposive selection but may not represent the entire population. The decision on which method to choose depends on various factors such as time constraints, budget, and study scope. Optimal research often involves testing samples that are representative of the population. However, when the population is too large, researchers often find convenience in using non-probability sampling methods. The research applies a convenient sampling method with an estimated sample size of 437 participants. The survey targets consumers in Ho Chi Minh City, aged between 18 and 40, who utilize electronic payment for online shopping. After data collection, the collected data was cleaned and analyzed using SPSS 22 and Smart PLS 4 software. The analysis methods will include descriptive statistics, measurement model testing, and structural model testing. Fig. 1 shows the proposed research model.

V. Result

A. Descriptive Statistics

To collect data, a questionnaire was sent to business owners in Ho Chi Minh City from various industries, different age groups, and genders. Due to various reasons, only 437 valid surveys were received. According to Table I, 44.9% of respondents were male, and 55.1% were female. In terms of age, the highest percentage was in the 18-22 age group, accounting for 46.2%. The surveyed individuals were mainly students (31.6%), office workers (25.2%), and business professionals (18.8%), and the remaining included freelancers, factory workers, teachers, and other professionals. Data Interpretation: It could benefit from further interpretation of these statistics. For example, the passage could discuss any notable trends or patterns observed in the data, such as differences in responses based on demographic characteristics. Sample Representation: The Authors mention that the survey targeted business owners in various industries, age groups, and genders. Discussing the extent to which the sample represents the target population would provide context for interpreting the results.

TABLE I. DESCRIPTIVE STATISTICS OF THE SURVEY SAMPLE

Characteristic		Frequency	Ratio (%)
Gender	Female	241	55.1
	Male	196	44.9
Age	18 to 22 years old	202	46.2
	23 to 30 years old	162	37.1
	30 to 40 years old	73	16.7
Income	From 1 million to 3 million VND	88	20.1
	From 3 million to 5 million VND	117	26.8
	Over 5 million to 10 million VND	127	29.1
	Over 10 million VND	105	24
Occupation	Students	138	31.6
	Office workers	110	25.2
	Teachers	24	5.5
	Business professionals	82	18.8
	Freelancers	43	9.8
	Factory workers	29	6.6
	Other professions	11	2.5
Total		437	100

B. Testing the Measurement Model

All factors, including PU, SC, PE, CV, PEOU, IU, SI, PR, and CT, consistently exhibit Cronbach's Alpha coefficients exceeding 0.8, indicating strong internal consistency reliability. The composite reliability coefficient, assessing internal consistency within the study, further supports this, with values typically above 0.7 considered good [65]. In our study (see Table II), composite reliability coefficients range from 0.904 to 0.938, indicating robust reliability across all constructs. Moreover, all items demonstrate outer loadings above the threshold of 0.707, indicating their substantial contribution to their respective constructs [66]. This ensures that the items reliably measure their intended constructs. Convergence validity, assessed through Average Variance Extracted (AVE), confirms that our constructs converge well within the PLS-SEM analysis. AVE values exceeding 0.5 are indicative of good convergence validity. In our study, all constructs exhibit AVE values above 0.705, underscoring strong convergence validity. Discriminant validity, crucial for distinguishing between constructs, is evaluated using the HTMT ratio and the Fornell-Larcker criterion. An HTMT ratio below 0.9 signifies adequate discriminant validity [67]. Our findings (see Table III) show HTMT ratios ranging from 0.282 to 0.716, indicating satisfactory discriminant validity among all constructs. Additionally, the Fornell-Larcker criterion, which compares the square roots of the AVEs (diagonal elements) with the correlations between constructs (off-diagonal elements), is consistently met. This criterion confirms that each construct's variance is greater than its correlations with other constructs, supporting discriminant validity.

These assessments collectively validate the measurement model's reliability, convergence validity, and discriminant validity, ensuring robustness in our findings.

TABLE II. RELIABILITY AND CONVERGENCE ASSESSMENT

Factor	Composite reliability (CR)	Cronbach's alpha (CA)	Average variance extracted (AVE)	Outer loading
SI	0.904	0.842	0.759	0.860 – 0.889
SC	0.930	0.906	0.726	0.828 – 0.869
PEOU	0.917	0.879	0.733	0.851 – 0.862
PU	0.905	0.861	0.705	0.826 – 0.850
PE	0.914	0.875	0.728	0.844 – 0.865
CT	0.912	0.872	0.723	0.838 – 0.856
PR	0.938	0.911	0.790	0.884 – 0.895
CV	0.925	0.892	0.755	0.855 – 0.876
IU	0.927	0.883	0.810	0.893 – 0.907

TABLE III. HTMT AND FORNELL-LARCKER DISCRIMINANT VALIDITY TESTING

HTMT									
	SI	SC	PEOU	PU	PE	CT	PR	CV	IU
SI									
SC	0.579								
PEOU	0.716	0.547							
PU	0.507	0.469	0.633						
PE	0.547	0.467	0.538	0.543					
CT	0.677	0.521	0.661	0.612	0.481				
PR	0.420	0.514	0.356	0.376	0.431	0.437			
CV	0.368	0.439	0.343	0.313	0.341	0.282	0.276		
IU	0.713	0.634	0.705	0.645	0.635	0.689	0.517	0.542	
Fornell - Larcker									
	SI	SC	PEOU	PU	PE	CT	PR	CV	IU
SI	0.871								
SC	0.508	0.852							
PEOU	0.618	0.490	0.856						
PU	0.433	0.415	0.552	0.840					
PE	0.470	0.418	0.472	0.471	0.853				
CT	0.581	0.465	0.580	0.532	0.422	0.850			
PR	-0.368	-0.469	-0.320	-0.333	-0.385	-0.391	0.889		
CV	0.321	0.397	0.305	0.275	0.304	0.252	-0.249	0.869	
IU	0.617	0.568	0.623	0.563	0.560	0.606	-0.463	0.483	0.900

C. Testing the Structural Model

The main limitation of quantitative research through survey (see Appendix) data is the representativeness of the sample compared to the population. Due to time constraints and other research resources, the study could not assess the entire population and only relied on a sample of 437 participants. Therefore, there are certain limitations in terms of the generalizability of the research findings. To address this, the authors used bootstrapping techniques to resample with replacement based on the collected sample data, aiming to establish a larger study sample. Assessing multicollinearity is one of the steps needed to evaluate the relationships between variables in the research model. The absence of multicollinearity is indicated when the Variance Inflation Factors (VIF) are less than 5 [71]. In this study, the VIF values range from 1.249 to 2.112, indicating no multicollinearity issue in the model. The R-squared (R^2) value ranges from 0% to 100% and a higher R^2 value implies a better accuracy of the research model's predictions. According to Table IV, R^2_{adjYD} is 63.5%, which means it accounts for 63.5% of the variance in IU explained by SI, SC, PEOU, PU, PE, CT, PR, and CT; thus, these are good predictors for IU. Cohen (1988) considers f^2 values of 0.02, 0.15, and 0.35 as small, medium, and large, respectively [72]. The f^2 values in Table IV show that the independent variables have medium to large effect sizes on the dependent variable in the research model, except for H12, which has no significant effect. Q^2 value greater than 0.02 emphasizes the predictive ability of the previously mentioned model. To test the statistical significance at a 5% equivalent level, the authors used a t-value greater than 1.96. The bootstrapping results indicate that SI, SC, PEOU, PU, PE, CT, PR, and CV have significant and statistically meaningful effects on YD at approximately 5% level ($p < 0.05$). Moreover, PEOU, PU, SI, and PR have statistically significant impacts on CT, specifically at a 5% level ($p < 0.05$). However, SC does not have a statistically significant impact on CT at a 5% level ($p > 0.05$). Based on these results, the authors accept hypotheses H11, H9, H1, H5, H2, H6, H8, H3, H17, H14, H15, H18 and reject hypothesis H12.

According to Table V, using CT as a mediating variable, SI, PU, and PEOU have significant effects on IU with beta coefficients of 0.044, 0.037, and 0.035, respectively, and all P-values are less than 0.05. Therefore, the research findings through the mediating variable support the idea that customer trust mediates the effects of social influence, perceived usefulness, and perceived ease of use on the intention to use electronic payments.

TABLE IV. TESTING THE STRUCTURAL MODEL

Relationships	Original sample	VIF	T statistics	P values	f^2	Result
SI -> CT	0.270	1.815	4.557	0.000	0.078	Accept
SI -> IU	0.153	2.017	2.758	0.006	0.032	Accept
SC -> CT	0.075	1.648	1.272	0.203	0.007	Reject
SC -> IU	0.099	1.755	2.152	0.031	0.021	Accept
PEOU -> CT	0.215	1.991	3.204	0.001	0.045	Accept
PEOU -> IU	0.153	2.112	2.572	0.010	0.031	Accept
PU -> CT	0.228	1.532	3.758	0.000	0.065	Accept
PU -> IU	0.127	1.712	2.341	0.019	0.026	Accept
PE -> IU	0.146	1.567	2.853	0.004	0.038	Accept
CT -> IU	0.163	1.932	3.140	0.002	0.038	Accept
PR -> CT	-0.112	1.343	2.025	0.043	0.022	Accept
PR -> IU	-0.099	1.405	2.355	0.019	0.031	Accept
CV -> IU	0.203	1.242	5.290	0.000	0.092	Accept
$R^2adj_{IU} = 0.635$; $R^2adj_{CT} = 0.467$; $Q^2_{IU} = 0.604$; $Q^2_{CT} = 0.456$						

TABLE V. MEDIATING ROLE OF CUSTOMER TRUST

	Original sample	P values	T statistics	Result
PEOU -> CT -> IU	0.035	0.033	2.135	Accept
PR -> CT -> IU	-0.018	0.103	1.631	Reject
SC -> CT -> IU	0.012	0.246	1.159	Reject
PU -> CT -> IU	0.037	0.017	2.388	Accept
SI -> CT -> IU	0.044	0.009	2.602	Accept

D. T – Test Anova

Demographics are variables that reflect individuals' characteristics and theories suggest that they partly influence individuals' thoughts and intentions. Regarding gender, with Sig = 0.149, the Levene test indicates that the variances of the female and male groups are equal. The T-test results with Sig = 0.315 show that there is no significant difference in the intention to use electronic payments between females and males. For occupation, age, and income, the Levene tests produce results in the order of 0.03, 0.043, and 0.001 < 0.05, indicating that the variances of these groups are significantly different. To account for unequal variances, the Welch test is conducted. The Welch test results for income, age, and occupation show Sig levels of 0.083, 0.497, and 0.341 > 0.05, respectively, indicating that there are no significant differences in the intention to use digital payment methods among these groups. These findings suggest that demographic factors such as gender, age, income, and occupation do not significantly influence the intention to use electronic payments. This lack of significant difference implies that the intention to adopt electronic payment systems is consistent across different demographic groups, highlighting the universal appeal and potential acceptance of electronic payments regardless of these demographic variables. Table VI tests the differences in demographic variables.

TABLE VI. TESTING THE DIFFERENCES IN DEMOGRAPHIC VARIABLES

	Levene's Test	T-test	Levene Statistic	Welch
	Sig.	Sig.	Sig.	Sig.
Gender	0.149	0.315		
Age			0.043	0.497
Income			0.001	0.083
Occupations			0.03	0.341

VI. DISCUSSION, MANAGERIAL IMPLICATIONS, LIMITATIONS, FUTURE RESEARCH DIRECTIONS

A. Discussion

The study examined factors influencing consumers' intention to use electronic payment for online purchases in Ho Chi Minh City, utilizing 36 observed variables with 1 dependent and eight independent variables. Through bootstrapping and hypothesis testing, the impact of factors on electronic payment behavior was evaluated across dimensions including "Customer Trust", "Convenience", "Perceived Ease of Use", "Social Influence", "Perceived Usefulness", "Security", "Performance Expectancy" and "Perceived Risk". Key findings are summarized as follows:

Among the factors assessed, “Convenience” demonstrated the highest impact (β = 0.203) on electronic payment adoption. This underscores its critical role as the primary consideration for consumers when choosing online payment methods, consistent with prior research [73], [74], [75]. Following closely, “Customer Trust” ranked second (β = 0.163), highlighting the high level of trust consumers in Ho Chi Minh City place in online payment systems, which aligns with existing literature [76], [77], [78]. “Social Influence” ranked third (β = 0.153), particularly significant during the COVID-19 pandemic, influencing consumer trust and adoption of electronic payment methods, as noted in previous studies [79], [80], [81]. “Perceived Ease of Use” also ranked prominently (β = 0.153), emphasizing the importance of intuitive and straightforward transaction processes in enhancing adoption rates, as supported by prior research [82], [64]. “Performance Expectancy” (β = 0.146) and “Perceived Usefulness” (β = 0.127) further contribute to consumer acceptance by facilitating efficient and beneficial transaction experiences, consistent with findings from other studies [79], [83], [84], [85], [86]. “Security” (β = 0.099) reassures consumers of the safety and protection of their personal and financial information, albeit ranking lower compared to other factors. Interestingly, “Perceived Risk” (β = -0.099) negatively impacts the intention to use electronic payment, suggesting that mitigating perceived risks could enhance adoption rates. This finding prompts further exploration into consumers' apprehensions and strategies to address them, acknowledging potential unknown risks that could deter usage. Moreover, the study revealed that “Social Influence”, “Perceived Ease of Use”, “Perceived Usefulness” and “Perceived Risk” significantly influence “Customer Trust” (β = 0.27, 0.228, 0.215, and -0.112, respectively), highlighting the mediating role of trust in enhancing online payment adoption.

To enhance electronic payment adoption, businesses and policymakers could leverage these insights by prioritizing factors such as convenience, trust-building measures, user-friendly interfaces, and security enhancements. Strategies should focus on reducing perceived risks and enhancing perceived benefits, thereby fostering greater consumer confidence and usage. In conclusion, while this study aligns with existing literature, deeper integration of prior research and theoretical frameworks could further enrich interpretations. Additionally, providing actionable recommendations based on these findings can bridge the gap between research insights and practical applications, promoting wider adoption of electronic payment systems in Ho Chi Minh City.

B. Managerial Implications

To ensure smooth and convenient electronic payment, service businesses should focus on enhancing user experiences by integrating various payment methods, prioritizing security and authentication, providing professional services, and supporting customers. Continuous updates and improvements based on user feedback are essential. To build trust with customers in electronic payment, businesses need to prioritize security, display reliability, and respond promptly to complaints. Constantly improving product and service quality will also foster customer satisfaction and trust. To maximize the social benefits of electronic payment, businesses should provide financial incentives to low-income individuals, encourage them to transition away from cash and promote sustainable finance. Ensuring the security of user information is crucial in building trust and loyalty among customers. User-friendliness in electronic payment requires services to integrate multiple payment options, minimize advertising, facilitate streamlined transactions, and provide clear and simple user instructions. This will make electronic payment convenient and easy for users.

To achieve effective use of electronic payment, businesses must commit to meeting user expectations, continuously improve systems, and integrate new features. This will enhance user satisfaction, and trust, and contribute to the success of electronic payment. The usefulness of electronic payment necessitates diversification of products and services, cost reduction, and faster transaction processing. Emphasis on security and quick response to user feedback is vital for improving user experience in payment. To ensure security in electronic payment, businesses should implement data encryption, two-factor authentication, and fraud detection systems, and restrict access to data. Proper training and timely response to security incidents are essential for protecting users' personal information. Finally, to minimize risks in electronic payment, businesses should establish terms of use in line with legal regulations and ensure safe transactions. By applying these managerial implications, businesses can create a seamless and secure electronic payment experience for their customers, thereby fostering trust and loyalty in the usage of electronic payment methods.

C. Limitations

The study has certain limitations that need to be considered. One such limitation is the relatively small sample size used for data collection, which may have implications for the robustness of the results. To improve the generalizability of the findings, a larger sample size could be utilized in future research. Additionally, since the study was conducted exclusively in Ho Chi Minh City, caution should be exercised when attempting to extend the results to other regions or countries. Moreover, although the study explored various factors affecting consumers' intention to use electronic payment, it is essential to recognize that there may be other relevant variables that were not accounted for in the analysis. Future research could expand the scope of the investigation to encompass additional factors that might influence consumers' adoption of electronic payment methods. Conduct studies with larger and more diverse samples to improve the robustness and generalizability of the findings. Explore comparative studies across different regions to understand cultural, economic, and technological influences on electronic payment adoption.

D. Future Research Directions

Conduct a comparative study across different cities or countries to explore how cultural, economic, and technological factors influence consumers' intention to use electronic payment. Investigate the factors influencing the adoption of mobile payment platforms, as smartphones and mobile apps play an increasingly significant role in electronic payment. Analyze how age and gender influence consumers' intention to use electronic payment, as different demographics might have varying preferences and concerns. Study consumer preferences for specific types of electronic payment methods (e.g., e-wallets, mobile banking, contactless payments) and how these preferences vary across different consumer segments. By addressing these limitations and conducting further research on these suggested areas, a more comprehensive understanding of consumers' intention to use electronic payment and strategies for businesses can be developed.

REFERENCES

[1] "FPT Digital," Vai trò của thanh toán điện tử trong việc thúc đẩy tăng trưởng kinh tế và bài học kinh nghiệm cho Việt Nam, 14 4 2022. [Online]. Available: https://digital.fpt.com.vn/chien-luoc/thanh-toan-dien-tu-tai-viet-nam.html.

[2] V. Trinh, "Sapo," Tổng quan thị trường thanh toán 2022: Thanh toán không tiền mặt trở thành xu hướng tất yếu, 07 02 2023. [Online]. Available: https://www.sapo.vn/blog/thanh-toan-khong-tien-mat-thanh-xu-huong-tat-yeu.

[3] Shon, T. H., & Swatman, P. M, "Identifying effectiveness criteria for internet payment systems," Internet Research: Electronic Networking Applications and Policy, vol. 8, no. 3, pp. 202-218, 1998.

[4] Gans, J. S., & Scheelings, R., "Economic Issues Associated with Access to Electronic Payments Systems," Australian Business Law Review, 1999.

[5] Lee, M. C., "Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit," Electronic commerce research and applications, vol. 8, no. 3, pp. 130-141, 2009.

[6] Fishbein, M., & Ajzen, I., "Belief, attitude, intention and behaviour: An introduction to theory and research," 1975.

[7] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS quarterly, pp. 319-340, 1989.

[8] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D., "User acceptance of information technology: Toward a unified view," MIS quarterly, pp. 425-478, 2003.

[9] Shanmugam, M., Sun, S., Amidi, A., Khani, F., & Khani, F., "The applications of social commerce constructs," International Journal of Information Management, vol. 36, no. 3, pp. 425-432, 2016.

[10] Lu, Y., Yang, S., Chau, P. Y., & Cao, Y., "Dynamics between the trust transfer process and intention to use mobile payment services: A cross-environment perspective," Information & Management, vol. 48, pp. 393-403, 2011.

[11] Pavlou, P. A., "Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model," International Journal of Electronic Commerce, vol. 7, no. 3, pp. 69-103, 2003.

[12] Papadopoulou, P., Nikolaidou, M., & Martakos, D., "What Is Trust in e-Government? A Proposed Typology," 43rd Hawaii International Conference on System Sciences, pp. 1 - 10, 2010.

[13] Dierksmeier, C., & Seele, P., "Cryptocurrencies and Business Ethics," Journal of Business Ethics, vol. 152, pp. 1-14, 2018.

[14] Mendoza-Tello, J. C., Mora, H., Pujol-López, F. A., & Lytras, M. D. , "Social commerce as a driver to enhance trust and intention to use cryptocurrencies for electronic payments," Ieee Access, vol. 6, pp. 50737-50751, 2018.

[15] Singh, N., & Sinha, N., "How perceived trust mediates merchant's intention to use a mobile wallet technology," Journal of retailing and consumer services, vol. 52, 2020.

[16] Farivar, S., Turel, O., & Yuan, Y., "Skewing users' rational risk considerations in social commerce: An empirical examination of the role of social identification," Information & Management, vol. 55, no. 8, pp. 1038-1048, 2018.

[17] Peter, J. P., & Ryan, M. J. , "An investigation of perceived risk at the brand level," Journal of marketing research, vol. 13, no. 2, pp. 184-188, 1976.

[18] Kubát, M., "Virtual currency bitcoin in the scope of money definition and store of value," Procedia Economics and Finance, vol. 30, pp. 409-416, 2015.

[19] Ciaian, P., Rajcaniova, M., & Kancs, D. A., "The digital agenda of virtual currencies: Can BitCoin become a global currency?," Information Systems and e-Business Management, vol. 14, pp. 883-919, 2016.

[20] Meents, S., & Verhagen, T., "Reducing consumer risk in electronic marketplaces: The signaling role of product and seller information," Computers in Human Behavior, vol. 86, pp. 205-217, 2018.

[21] Kim, K., Prabhakar, B., & Park, S. K., "Trust, perceived risk, and trusting behavior in Internet banking," Asia Pacific Journal of Information Systems, vol. 19, no. 3, pp. 1-23, 2009.

[22] Yang, Q., Pang, C., Liu, L., Yen, D. C., & Tarn, J. M., "Exploring consumer perceived risk and trust for online payments: An empirical study in China's younger generation," Computers in human behavior, vol. 50, pp. 9-24, 2015.

[23] Khwaja, M. G., Mahmood, S., & Zaman, U., "Examining the Effects of eWOM, Trust Inclination, and Information Adoption on Purchase Intentions in an Accelerated Digital Marketing Context," Information , vol. 11, no. 10, pp. 1-12, 2020.

[24] Hong, I. B., & Cha, H. S., "The mediating role of consumer trust in an online merchant in predicting purchase intention," International Journal of Information Management, vol. 33, no. 6, pp. 927-939, 2013.

[25] Ilhamalimy, R. R., & Ali, H. , "Model perceived risk and trust: e-WOM and purchase intention (the role OF trust mediating IN online shopping IN shopee Indonesia)," Dinasti International Journal of Digital Business Management, vol. 2, no. 2, pp. 204 - 221, 2021.

[26] Liébana-Cabanillas, F., Higueras-Castillo, E., Molinillo, S., & Montañez, M. R., "Assesing the role of risk and trust in consumers' adoption of online payment systems," International Journal of Information Systems and Software Engineering for Big Companies, vol. 5, no. 2, pp. 99-113, 2019.

[27] Wong, W. H., & Mo, W. Y., "A study of consumer intention of mobile payment in Hong Kong, based on perceived risk, perceived trust, perceived security and Technological Acceptance Model," Journal of Advanced Management Science Vol, vol. 7, no. 2, pp. 33-38, 2019.

[28] Yet Mee, L., Cham, T. H., & Chuan, S. B., "Medical tourists' behavioral intention in relation to motivational factors and perceived image of the service providers," International Academic Journal of Organizational Behavior and Human Resource Management, vol. 5, no. 3, pp. 1-16, 2018.

[29] Kavya Shree, K. M., & Manasa, N. , "Perception and attitudes towards information and communication technology (internet) for purchase decisions among generation cohorts," American Journal of Information Science and Computer Engineering, vol. 3, no. 4, pp. 56-63, 2017.

[30] Priyono, A., "Analisis pengaruh trust dan risk dalam penerimaan teknologi dompet elektronik Go-Pay," Jurnal Siasat Bisnis, vol. 21, no. 1, pp. 88-106, 2017.

[31] Wang, Z., & Li, H., "Factors influencing usage of third party mobile payment services in China: An empirical study," Master's thesis, Uppsala University, Sweden, 2016.

[32] Tahar, A., Riyadh, H. A., Sofyani, H., & Purnomo, W. E. , "Perceived ease of use, perceived usefulness, perceived security and intention to use e-filing: The role of technology readiness," The Journal of Asian Finance, Economics and Business, vol. 7, no. 9, pp. 537-547, 2020.

[33] Huang, F., Teo, T., & Scherer, R., "Investigating the antecedents of university students' perceived ease of using the Internet for learning," Interactive Learning Environments, vol. 30, no. 6, pp. 1060-1076, 2020.

[34] Aghdaie, S. F. A., Piraman, A., & Fathi, S., "An analysis of factors affecting the consumer's attitude of trust and their impact on internet purchasing behavior," International Journal of Business and Social Science, vol. 2, no. 23, pp. 147-158, 2011.

[35] Koufaris, M., Kambil, A., & LaBarbera, P. A., "Consumer behavior in web-based commerce: an empirical study," International journal of electronic commerce, vol. 6, no. 2, pp. 115 - 138, 2001.

[36] Chinomona, R., "The influence of perceived ease of use and perceived usefulness on trust and intention to use mobile social software: technology and innovation," African Journal for Physical Health Education, Recreation and Dance, vol. 19, no. 2, pp. 258-273, 2013.

[37] Wilson, N, "The impact of perceived usefulness and perceived ease-of-use toward repurchase intention in the Indonesian e-commerce industry," Jurnal Manajemen Indonesia, vol. 19, no. 3, pp. 241-249, 2019.

[38] Al-Sharafi, M. A., Arshah, R. A., Herzallah, F. A., & Alajmi, Q., "The effect of perceived ease of use and usefulness on customers intention to use online banking services: the mediating role of perceived trust," International Journal of Innovative Computing, vol. 7, no. 1, pp. 9-14, 2017.

[39] Yudiarti, R. F. E., & Puspaningrum, A. , "The role of trust as a mediation between the effect of perceived usefulness and perceived ease of use to interest to buy e-book," Jurnal Aplikasi Manajemen, vol. 16, no. 3, pp. 494-502, 2022.

[40] Wen, C., Prybutok, V. R., & Xu, C., "An integrated model for customer online repurchase intention," Journal of Computer information systems, vol. 52, no. 1, pp. 14-23, 2011.

[41] Akbari, M., Rezvani, A., Shahriari, E., Zúñiga, M. A., Pouladian, H, "Acceptance of 5G Technology: Mediation Role of Trust and Concentration," Journal of Engineering and Technology Management - JET-M, 2020.

[42] Kurniawan, I. A., Mugiono, M., & Wijayanti, R., "The effect of Perceived Usefulness, Perceived Ease of Use, and social influence toward intention to use mediated by Trust," Jurnal Aplikasi Manajemen, vol. 20, no. 1, pp. 117-127, 2022.

[43] Redzuan, N. I. N., Razali, N. A., Muslim, N. A., & Hanafi, W., "Studying perceived usefulness and perceived ease of use of electronic human resource management (e-HRM) with behavior intention," International Journal of Business Management, vol. 1, no. 2, pp. 118-131, 2016.

[44] Tarhini, A., El-Masri, M., Ali, M., & Serrano, A., "Extending the UTAUT model to understand the customers' acceptance and use of internet banking in Lebanon: A structural equation modeling approach," Information Technology & People, vol. 29, no. 4, pp. 830-849, 2016.

[45] Yeow, A., Soh, C., & Hansen, R., "Aligning with new digital strategy: A dynamic capabilities approach," The Journal of Strategic Information Systems, vol. 27, no. 1, pp. 43-58, 2018.

[46] Aji, H. M., & Dharmmesta, B. S., "Subjective norm vs dogmatism: Christian consumer attitude towards Islamic TV advertising," Journal of Islamic Marketing, vol. 10, no. 3, pp. 961-980, 2019.

[47] Yang, M., Mamun, A. A., Mohiuddin, M., Nawi, N. C., & Zainol, N. R., "Cashless transactions: A study on intention and adoption of e-wallets," Sustainability, vol. 13, no. 2, pp. 1-18, 2021.

[48] Rama Murthy, S., & Mani, M., "Discerning rejection of technology," Sage Open, vol. 3, no. 2, pp. 1-10, 2013.

[49] Amin, M., Rezaei, S., & Abolghasemi, M., "User satisfaction with mobile websites: the impact of perceived usefulness (PU), perceived ease of use (PEOU) and trust," Nankai Business Review International, vol. 5, no. 3, pp. 258-274, 2014.

[50] Chen, Y. H., & Barnes, S., "Initial trust and online buyer behaviour," Industrial management & data systems, vol. 107, no. 1, pp. 21-36, 2007.

[51] Singh, N., & Sinha, N., "How perceived trust mediates merchant's intention to use a mobile wallet technology," Journal of Retailing and Consumer Services, vol. 52, 2020.

[52] Casalo, L. V., Flavián, C., & Guinalíu, M., "The influence of satisfaction, perceived reputation and trust on a consumer's commitment to a website," Journal of Marketing Communications, vol. 13, no. 1, pp. 1-17, 2007.

[53] Belanche-Gracia, D., Casaló-Ariño, L. V., & Pérez-Rueda, A., "Determinants of multi-service smartcard success for smart cities development: A study based on citizens' privacy and security perceptions," Government Information Quarterly, vol. 32, no. 2, pp. 154-163, 2015.

[54] Flavián, C., & Guinalíu, M., "Consumer trust, perceived security and privacy policy: three basic elements of loyalty to a web site," Industrial management & data Systems, vol. 106, no. 5, pp. 601-620, 2006.

[55] Kaur, R., Li, Y., Iqbal, J., Gonzalez, H., & Stakhanova, N. , "A security assessment of HCE-NFC enabled e-wallet banking android apps," IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 492-497, 2018.

[56] Meharia, P., "ASSURANCE ON THE RELIABILITY OF MOBILE PAYMENT SYSTEM AND ITS EFFECTS ON ITS'USE: AN EMPIRICAL EXAMINATION," Accounting and Management Information Systems, vol. 11, no. 1, p. 97–111, 2012.

[57] Khan, S., Umer, R., Umer, S., & Naqvi, S., "Antecedents of trust in using social media for E-government services: An empirical study in Pakistan," Technology in Society, vol. 64, 2021.

[58] Kim, C., Tao, W., Shin, N., & Kim, K. S., "An empirical study of customers' perceptions of security and trust in e-payment systems," Electronic commerce research and applications, vol. 9, no. 1, pp. 84-95, 2010.

[59] Enaizan, O., Eneizan, B., Almaaitah, M., Al-Radaideh, A. T., & Saleh, A. M., "Effects of privacy and security on the acceptance and usage of EMR: the mediating role of trust on the basis of multiple perspectives," Informatics in Medicine Unlocked, vol. 21, pp. 1-10, 2020.

[60] Kim, S., & Park, H. , "Effects of various characteristics of social commerce (s-commerce) on consumers' trust and trust performance," International Journal of Information Management, vol. 33, no. 2, pp. 318-332, 2013.

[61] Morosan, C., & DeFranco, A., "It's about time: Revisiting UTAUT2 to examine consumers' intentions to use NFC mobile payments in hotels," International Journal of Hospitality Management, vol. 53, pp. 17-29, 2016.

[62] Park, J., Ahn, J., Thavisay, T., & Ren, T., "Examining the role of anxiety and social influence in multi-benefits of mobile payment service," Journal of retailing and consumer services, vol. 47, pp. 140-149, 2019.

[63] Singh, A., Alryalat, M. A. A., Alzubi, J. A., & Sarma, H. K., "Understanding jordanian consumers' online purchase intentions: Integrating trust to the UTAUT2 framework," International Journal of Applied Engineering Research, vol. 12, no. 20, pp. 10258-10268, 2017.

[64] Kurniawan, I. A., Mugiono, M., & Wijayanti, R., "The effect of Perceived Usefulness, Perceived Ease of Use, and social influence toward intention to use mediated by Trust," Journal of Applied Management (JAM), vol. 20, no. 1, pp. 117 - 127, 2022.

[65] Kandoth, S., & Shekhar, S. K., "Social influence and intention to use AI: the role of personal innovativeness and perceived trust using the parallel mediation model," Forum Scientiae Oeconomia, vol. 10, no. 3, pp. 131-150, 2022.

[66] Chan, W. M., & Lee, J. W. C., "5G Connected Autonomous Vehicle Acceptance: The Mediating Effect of Trust in the Technology Acceptance Model," Asian Journal of Business Research, vol. 11, no. 1, pp. 40-60, 2021.

[67] Alalwan, A. A., Dwivedi, Y. K., & Rana, N. P., "Factors influencing adoption of mobile banking by Jordanian bank customers: Extending UTAUT2 with trust," Int. J. Inf. Manage, vol. 37, no. 3, pp. 99-110, 2017.

[68] Zhou, T., "Exploring mobile user acceptance based on UTAUT and contextual offering," international Symposium on electronic Commerce and security, 2008.

[69] Riquelme, H. E., & Rios, R. E., "The moderating effect of gender in the adoption of mobile banking," International Journal of bank marketing, vol. 28, no. 5, pp. 328-341, 2010.

[70] Pal, D., Vanijja, V., & Papasratorn, B., "An empirical analysis towards the adoption of NFC mobile payment system by the end user," Procedia Computer Science, vol. 69, pp. 13-25, 2015.

[71] Hair, J., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L., "An updated and expanded assessment of PLS-SEM in information systems research. Industrial management & data systems," Industrial management & data systems, vol. 117, no. 3, pp. 442-458, 2017.

[72] Cohen, J., Statistical power analysis for the behavioural sciences, Lawrence Erlbaum, 1988.

[73] Wardana, A. A., Saputro, E. P., Wahyuddin, M., & Abas, N. I., "The effect of convenience, perceived ease of use, and perceived usefulness on intention to use e-wallet.," In International Conference on Economics and Business Studies, vol. 218, pp. 386-395, 2022.

[74] Al-Qudah, A. A., Al-Okaily, M., Alqudah, G., & Ghazlat, A. , "Mobile payment adoption in the time of the COVID-19 pandemic," Electronic Commerce Research, 2022.

[75] Liébana-Cabanillas, F., Muñoz-Leiva, F., Molinillo, S., & Higueras-Castillo, E., "Do biometric payment systems work during the COVID-19 pandemic? Insights from the Spanish users' viewpoint," Financial Innovation, vol. 8, no. 1, pp. 1-25, 2022.

[76] Alkhowaiter, W. A., "Use and behavioural intention of m-payment in GCC countries: Extending meta-UTAUT with trust and Islamic religiosity," Journal of Innovation & Knowledge, vol. 7, no. 4, 2022.

[77] Esawe, A. T., "Exploring retailers' behavioural intentions towards using m-payment: Extending UTAUT with perceived risk and Trust," Paradigm, vol. 26, no. 1, pp. 8-28, 2022.

[78] Saha, T., Dey, T., & Hoque, M. R., "Initial trust and usage intention: A study on mobile payment adoption in Bangladesh," Global Business Review, pp. 1-23, 2022.

[79] Al-Sabaawi, M. Y. M., Alshaher, A. A., & Alsalem, M. A., "User trends of electronic payment systems adoption in developing countries: an empirical analysis," Journal of Science and Technology Policy Management, vol. 14, no. 2, pp. 246-270, 2021.

[80] Ming, K. L. Y., & Jais, M., "Factors affecting the intention to use e-wallets during the COVID-19 pandemic," Gadjah Mada International Journal of Business, vol. 24, no. 1, pp. 82-100, 2022.

[81] Jawad, A. I., Parvin, T., & Hosain, M. S., "Intention to adopt mobile-based online payment platforms in three Asian countries: an application of the extended Technology Acceptance Model," Journal of Contemporary Marketing Science, vol. 5, no. 1, pp. 92-113, 2022.

[82] Siagian, H., Tarigan, Z. J. H., Basana, S. R., & Basuki, R., "The effect of perceived security, perceived ease of use, and perceived usefulness on consumer behavioral intention through trust in digital payment platform," International journal of data and network science, vol. 6, pp. 1-14, 2022.

[83] Cahyani, U. E., Sari, D. P., & Afandi, A., "Determinant of behavioral intention to use digital zakat payment: The Moderating Role of Knowledge of Zakat.," Jurnal Zakat Dan Wakaf, vol. 9, no. 1, pp. 1-16, 2022.

[84] Leong, C., Tan, K., Puah, C., & Chong, S., "Predicting mobile network operators users m-payment intention," European Business Review, vol. 33, no. 1, 2020.

[85] Kustono, A. S., Nanggala, A. Y. A., & Masud, I., "Determinants of the use of e-wallet for transaction payment among college students," Journal of Economics, Business, and Accountancy Ventura, vol. 23, no. 1, pp. 85-95, 2020.

[86] Ardiansah, M., Chariri, A., Rahardja, S., & Udin, U., "The effect of electronic payments security on e-commerce consumer perception: An extended model of technology acceptance," Management Science Letters, vol. 10, no. 7, pp. 1473-1480, 2020.

[87] Daragmeh, A., Lentner, C., & Sági, J., "FinTech payments in the era of COVID-19: Factors influencing behavioral intentions of ''Generation X'' in Hungary to use mobile payment," Journal of Behavioral and Experimental Finance, vol. 32, 2021.

[88] Tella, A., & Olasina, G., "Predicting Users' Continuance Intention Toward E-payment System: An Extension of the Technology Acceptance Model," International Journal of Information Systems and Social Change, vol. 5, no. 1, pp. 47-, 2014.

[89] Chawla, D., & Joshi, H., "The moderating role of gender and age in the adoption of mobile wallet," foresight, vol. 22, no. 4, pp. 483-504, 2020.

[90] Khalilzadeh, J., Ozturk, A. B., & Bilgihan, A. , "Security-related factors in extended UTAUT model for NFC based mobile payment in the restaurant industry," Computers in human behavior, vol. 70, pp. 460-474, 2017.

[91] Liébana-Cabanillas, F., Sánchez-Fernández, J., & Muñoz-Leiva, F., "The moderating effect of experience in the adoption of mobile payment tools in Virtual Social Networks: The m-Payment Acceptance Model in Virtual Social Networks (MPAM-VSN)," International Journal of Information Management, vol. 34, no. 2, pp. 151-166, 2014.

APPENDIX

Survey Questions

Perceived ease of use (PEOU)	Soucer
1. For me, using electronic payment is very easy. 2. I believe that becoming proficient in using electronic payment will not be difficult. 3. Digital payments have made online shopping much simpler than before. 4. Transferring money through electronic payment gateways is also fast and convenient.	[52], [83], [84], [85]
Perceived usefulness (PU)	
1. I will pay faster when using electronic payment. 2. Using digital payments brings more benefits compared to traditional payment methods (cash or direct contact). 3. Digital payments are convenient for paying when shopping online. 4. Using digital payments helps me increase the efficiency of my purchases. 5. Using digital payments allows me to make payments for my usual online purchases.	[7], [83], [57]
Security (SC)	
1. I feel calm and secure when providing credit card or debit card information through the digital payment system. 2. I feel that digital payments are secure when transmitting sensitive information. 3. When using digital payments, I feel at ease when providing my personal information. 4. I believe that my transactions through digital payments will be protected. 5. I believe that when using digital payments, there will always be preparedness to handle risks and ensure data security.	[86]
Perceived Risk (PR)	
1. If I use digital payments, others may access information about my online transactions. 2. There is a high possibility of losing money if I make purchases using the digital payment system. 3. There are significant risks when buying with the digital payment system. 4. I consider buying with digital payments a risk.	[87]
Social influence (SI)	
1. In my life, there are important people, such as family, close friends, and colleagues, who have advised me to use digital payments for online shopping. 2. The important people in my life support me in using digital payments. 3. The community around me is using digital payments.	[88]
Performance Expectancy (PE)	[89]
1. I find digital payments to be a useful way of shopping. 2. Using digital payments makes transaction processing simpler. 3. Using digital payments improves the efficiency of my payments. 4. I save transaction costs when using digital payments.	

Convenience (CV)	
1. Digital payments are convenient because I often carry my smartphone with me. 2. Digital payments are convenient because I can use them anywhere. 3. Digital payments are convenient because I can use them in any situation. 4. Digital payments are convenient because they are not complicated.	[64]
Customer Trust (CT)	
1. I believe that digital payment platforms are capable and effective in processing contactless transactions. 2. I believe that digital payment platforms always prioritize the interests of consumers. 3. I believe that the legal frameworks for providing digital payments are sufficient to protect consumers. 4. I believe that digital payment platforms are honest and truthful with users.	[86], [90]
Intention to use e-Payment (IU)	
1. I intend to use digital payments for online shopping. 2. Shortly, I will use digital payments. 3. In the coming time, I am willing to continue using digital payments.	[91], [89]

Moving Beyond Traditional Incident Response: Combating APTs with Warfare-Enabled Continuous Response

Abid Hussain Shah
CIS Department, University of Melbourne, Melbourne, Australia

***Abstract*—Critically examining the cybersecurity management practices, it can be concluded that security management used by the organizations is mostly control-centered against a wide range of threats to information systems. This control-centered approach has matured to act as a shield to prevent against a large variety of attacks. Since threats against the information systems are becoming sophisticated, persistent and evolving, therefore, the current approach has not been very effective against the advanced strategies and techniques used by the emerging threats like APTs (Advanced Persistent Threats). The core argument of this paper suggests that to match up the capabilities of APTs, organizations need a major shift in their strategies. This shift needs to focus more on the response oriented techniques relegating erstwhile prevention-centered approach. Traditionally the warfare strategies are more response oriented. Some of the non-kinetic strategies (not involving physical fighting) can be useful in developing response capability of Information Systems. Therefore, drawing on the warfare paradigm, and making use of DCT (Dynamic Capability Theory), this research examines the applicability of warfare strategies in the entrepreneur domain. This article will also contribute by means of a research framework arguing that the integration of prevailing information security capabilities; such as incident response capabilities and security capabilities from the warfare practices is possible resulting in dynamic capabilities (warfare-enabled). Such capabilities can improve security performance.**

Keywords*—Information operations; information warfare; cyber security; dynamic capabilities; incident response capabilities; warfare enabled capabilities

I. Introduction

Within organizations, the most valuable assets are information systems and the infrastructures which require protection. Organizations are vulnerable against a variety of attacks which threaten to breach security mechanisms of the organizations to reach to the critical assets. Against this vast spectrum of attacks, information security approaches are traditional. Now a days, protecting the data against cyber-attacks has become a mounting challenge. Mostly the purpose of cyber-attacks remains financial gains [1]. Cyber-attacks can have many purposes may be military or political. Research conducted by Atif Ahmad and Richard Baskerville suggests that Organizations try to counter threats following strategies which are preventive in nature and are mostly control-centered [2], [3]. Such strategies have proved to be reasonably strong/successful against predicted or known attacks; nevertheless, the current traditional approach is being challenged by the increasingly complex and evolving threat environment. The emergence of highly sophisticated and potent cyber threat known as APT (Advanced Persis-tent Threats), challenges conventional information security paradigms in organizations. Baskerville [3] has strongly recommended new paradigms to follow; relegating the compromised prevention-oriented techniques for ensuring information systems' security. In this pursuit, learning from the Warfare Strategies mostly the non-kinetic strategies as well as incorporating theory of Dynamic Capabilities (DC), this article is looking at prospects of enhancing the corporate security performance, through the dynamic capabilities achieved as a result of integrating Warfare capabilities and conventional Incident Response capabilities.

The use of internet is on a constant and rapid increase and currently over 3 billion users world over use internet on daily basis (Tan et al., 2021). Many of the researches conducted in the domain of information security conclude that safeguarding the critical databases and ensuring the protection of information resources in the organizations has become complex, costly and time-consuming [4]–[6]. Scholars define information security threat as, an adverse event which sometimes may be in shape of a violation of policy or an unauthorized access [7]. Security threats are further categorized as Incidental threats, encompassing human errors, technical failures or forces of nature affecting security [7]. The second category of threats is critical and known as Purposive threats. The purposive threats look for deliberate and intentional breaches to a system's security, essentially driven by human efforts and intelligence [2], [7]. These are the 2nd category of threats the purposive threats; which are becoming increasingly challenging since mostly they are new, changing, exploring new vulnerabilities and are persistent in nature, well disciplined and are focused to achieve strategic goals/ objectives. Due to their peculiar characteristics of the purposive threats the security environment has become more vulnerable and increasingly uncertain [7] [8]. To address the growing challenges this article focuses to counter the real threats to information systems; the purposive threats.

As the technological advancements in the field of IT are super rapid, therefore, the security paradigm of information security is also changing quickly. The threats to information systems are becoming more silent and therefore, difficult to detect. Threats are also becoming complex and evolving so, we are facing more incidents and more frequently. The scenario suggests that enhancing the information security capabilities of the organizations is essentially required [3], [6], [7], [9]. Why the APTs are succeeding? It can be safely concluded that the current response capabilities of the organizations are inadequate to match the superior capabilities of APTs (Shah et al. 2019). Contrary to Information Security strategies, the Warfare response strategies are inherently quicker in generating response therefore; they are more suited to ever changing threat landscape [3], [7], [10] [11][12]. In this article it is being proposed that for enhancing enterprise security (information systems security), a major shift of the security focus is needed. The suggestion is to focus and adopt the best practices of response domain of the warfare paradigm, which is largely response-oriented paradigm. Analyzing extant literature, it is revealed that there is hardly any literature available to answer the question of adopting warfare practices by the organizations to enhance their information security landscape.

Therefore, to enhance the security of the organizations especially in the response domain, this article is proposing a response framework which incorporates warfare response capabilities. Almost all organizations face the challenges of external and internal security threat to their Information systems; this article is better suited for those organizations which possess their exclusive Incident response Teams/setups. There are four sections in this paper. The theoretical framing has been explained in the next section. In the literature review section, the extant literature on security incident response has been explored extensively. Possibility of integration of the capabilities of Information Warfare and Incident Response capabilities has also been explored resulting in dynamic capabilities. Based on this discussion a conceptual frame-work of warfare-based security response has been introduced before concluding the paper. In the next section we will be discussing Dynamic Capabilities theory to understand its usage in developing dynamic cyber security capabilities in the ever-changing threat landscape.

II. Dynamic Capabilities (DC) Theory and Security Response

Considering the rapidly changing technological advancements, there has been substantial focus to make the organizational capabilities dynamic instead of static, therefore, the importance of Dynamic Capabilities (DC) cannot be ignored. The DC have been explained by many authors especially Teece. He proposes the DC approach as "an extension of the resource-based view (RBV)", which was an erstwhile concept [13]. Since the security threat environment is continuously changing while the RBV is static in its nature, hence it is unable to match up the ever changing threat spectrum, faced by information systems' security. [7] [14]. This mismatch was adequately addressed by Teece et al. [13] through introduction of concept of DC. Teece has explained the dynamic capability as the one which can combat the challenging environments and possesses the ability to develop, built also integrate and subsequently reconfigure the competencies whether internal or external [15]. Explaining the concept further, it has been recognized that through the DC, organizations can develop advanced resources and configuration following specific strategic routines [16] [17]. DC also contribute in achieving competitive advantage, by exploiting available opportunities, sometimes create opportunities and thus keep the firm well prepared to meet future and current challenges. This phenomenon is essentially sensing, seizing the available opportunities and keeping the firm competitive [15]. The DC approach also contributes in learning from incidents [7], [18]. It is also argued that the Incident Response can be managed by the organizations efficiently by developing DC [7], [19]. Therefore, it is essential that organizations adapt to the rapidly changing threat environments. It is also well recognized that better adaptation occurs by following an interdisciplinary approach [20]. Drawing from the Henry's conclusions; an inter-disciplinary approach, an integrative and dynamic capability development can be ensured which possesses more abilities and robustness for adapting to frequently changing environments. This approach can also avoid pensive and closed-minded views [7], [21] [22]. Since the warfare capabilities are response oriented and conventional Incident Response (IR) capabilities are prevention oriented, integration of both results in Dynamic Capabilities (DC) ensuring much better security against the looming threats. Thus, the overall enterprise security performance enhances many folds [7]. It has been argued in the paper that DC theory is appropriate and relevant for the research as in the information security domain organizations face threats which are complex and evolving swiftly. This phenomenon creates high levels of uncertainty which can be addressed though building dynamic capabilities. The section covers the detailed literature review with an objective to understand available and being practiced conventional Incident Response Capabilities. Subsequently, warfare relevant capabilities like operational security and deception have also been explored. Both these warfare capabilities are non-kinetic in nature and can be integrated with the traditional Incident Response capabilities, resulting in dynamic capabilities which have been named as Warfare Enabled Dynamic Response Capabilities (WEDRC).

III. Literature Review

To conduct a systematic literature review creating a firm foundation for advancing knowledge and to establish the need for answering the identified research question, literature review was conducted following the parameters explained by [23]. Latest as well as relevant publications (articles) published in renowned conferences and journals in the domains of Cyber Security, Information Operations (IO), Information Warfare (IW), Information Systems (IS), Information System Security, cyber and Information Management were consulted. For focused results and get the relevant hits research was done based on the phrases and key words like; 'Cyber security management', 'information security issues', 'information security superior strategy', information systems' security management', 'strategies dealing advanced persistent threats', 'information systems controls', 'warfare superior strategies', 'dynamic capabilities development', 'incident response for organizations' and 'information security risk management'. Initially 157 articles were selected besides 12 books and 10 field manuals from US military literature. The research focus was narrowed based on the contents of the articles and relegating not relevant to our research domain. Further relevancy was established keeping in view the intent and the focus of research question such as: How the enterprise security performance can be enhanced? How we can improve the security response capability of information systems? And what all can be adopted from the warfare strategies for information systems' security enhancement? Thus the strength of articles was reduced to 66. Subsequently the for-ward and backward chaining process was carried out using the references of selected 66 articles; which resulted in an addition of 15, increasing the total to 81 articles, two books and 4 United States military field manuals on Information Warfare. With maturing the research process 13 articles initially included were deferred, bringing the total of articles to 68. Progressive and objective analysis gave birth to the proposed framework (Fig. 1). The results of the literature review suggest that the incident response capabilities are more focused on prevention of the incidents [7], [24], [25]. Therefore, they can address predicted or known threats only. Secondly the warfare capabilities have inbuilt response due to the nature of the warfare [5], [6]. Thirdly, DC theory provides sufficient helping tools in shape of sensing, seizing and transformation for efficient and dynamic incident response handling [19].

A. Conventional Incident Response (IR) Capabilities

Amongst the security threats to the information systems (IS), the purposive threats are more serious and damaging challenge to the security of organizations since they are very well organized, having mostly a well-orchestrated strategic objective, persistent in nature, and are ever evolving [7], [8]. The purposive threats always look for vulnerabilities in organizational de-fences. In the domain of information security, threats are combatted by application of controls, such as: (1) Formal (e.g. security policies, management of risk) (2) Informal (training), and (3) Technological controls (e.g. Fire walls, intrusion detection and protection systems, anti-viruses) [6], [7], [26]. Since the information systems; threats are usually handled by applying appropriate controls, this process of security management represents a control-centered approach [3], [7]. Since the Incident Response capabilities can be defined in many different ways, for this article DC theory perspective has been used. According to the well-known author Teece, the term 'capabilities' actually defines the role of firm's strategic management to adapt, integrating, and reconfigure organizational skillsets, resources, and acquired competences to deal with the changing environment and requirements" [13]. This has given lead to de-fine the Incident Response Capabilities: IR capabilities constitute all available controls (for-mal, informal and technical), practices and processes; to address security threats to IS while performing IR functions [7]. Literature also mentions about another perspective related to capabilities; people, processes and technologies for handling the security incidents [27].

Today's modern organizations are facing challenges to protect their information resources, and IT infrastructure. Literature also reveals that the current IS security response is essentially based on prevention strategy [7], [28]. The prevention-oriented approach has been quite successful for known threats. However, despite following the prevention oriented controls, in-formation security incidents are still happening reflecting failure of our defensive security mechanisms [7]. Large sized organizations, financial firms, banks and government organizations maintain exclusive incident response teams, however, maintaining exclusive IR teams becomes very expensive and does not remains cost-effective for smaller businesses and organizations [7]. Therefore to remain cost effective, most of the medium and small sized organizations maintain temporary IR teams. Also, the incident responders generally perform as 'fire-fighters' [5], [7].

B. Warfare Capabilities – Information Operations

While conceiving the leading plans in the warfare domain, multiple contingencies are hypothesized for that leading operational plan. This approach harnesses and harbors possibility of generating a robust response at the spur of the moment when and where required, harnessing and articulating the available resources in the given environment (e.g. an adversary's offensive movement). There are many warfare strategies which govern specific areas of the warfare. Amongst these multiple military strategies, few are non-kinetic like the Information Warfare (IW) which does not involve physical fighting and aims at protecting information and information systems, while disrupting those of the adversary [7], [12],[9]. The warfare paradigm is predominantly response oriented. Following the war fare principles pf employment like defense-in-depth, early warning Systems, maintain reserves at all tiers and extensive contingency planning; the defensive response and maneuvers can be well orchestrated against unknown offensive in unknown and fluent environments of the battlefield [7].

Literature concerning military sciences reveals that the IW is a non-kinetic war strategy; which has proved to be quite successful in protecting, exploiting, corrupting, denying or destroying information as well as information systems and infrastructure. Thus, it helps achieve competitive advantage over adversary [29],[9]. Since the IW deals with information and information systems (IS), therefore, its utility has been fairly established by many authors much beyond the warfare domain. As per the US military doctrine, success of IW comes through the integrated employment of Core, Supporting and Related capabilities (total 13 in number) which successfully affects the adversaries' decision makers, information and information systems. It also protects own information and information systems against adversary's such efforts resulting in influencing the decision making process [11], [12] [7]. An in-depth analysis of all the 13 capabilities of IW for their applicability for non-military organizations reveal that many of the capabilities are applicable in the corporate world however, for this article only two of the warfare capabilities namely; Operational Security (Opsec), and Deception are being considered [7].

- Operational Security (Opsec). It is Information Warfare strategy which is designed to meet operational needs. Successful application of Opsec assists in mitigating the risks related to the specific defensive vulnerabilities. This process is designed to shield critical information and observable indicators to the adversaries [7], [11]. Operational security can also be used for identification of the existing vulnerabilities in defensive/information systems of the adversaries, conducting risk assessment and planning/ application of appropriate protective measures [7]. To identify the looming threats like APTs, the scope of Opsec can be extended in Information Security (IS) domain to carryout monitoring and attaining knowledge about the strategic objectives of threats. Another application of Opsec can be discerning about the phases of APTs thus denial of objectives of the intruders in each phase can also be planned [7].

- Deception. Deception is an old age warfare concept which has been frequently used in information security domain as well. It mostly consists of those activities, actions which deliberately mislead the adversary's decision making process and operations. In fact successful deception forces the adversaries to take specific actions or inactions which are well planned for own benefit. Thus, deception hides the facts from adversary and presents the false [11]. If the warfare capability of deception is employed in the in security mechanism, the response capability against the critical threats like APTs gets enhanced [30]. Concept of deception is already in use though, not with its deepest fruits and in depth understanding as a capability. Some of the examples of the concept being used in IS domain are honey pots, breadcrumbs, and personas etc. The purpose of deception as a capability (dynamic) is to incorporate response, developing understanding of the strategies, methods, practices and objectives of the APTs through taking the threats into safe and imitated environments [7] (Shah et al. 2019).

C. Warfare Enabled Dynamic Response Capabilities (WEDRC)

Dynamic Capabilities (DC) ensure that the firm remains competitive by continuously improving the methods, process and technologies as well as reviving the role components of the organization. The capabilities which are responsible for routine functionalities of the organizations are called ordinary capabilities while the higher level capabilities which keep the organizations competitive are known as dynamic capabilities which can at times change the ordinary capabilities as well [7], [31]. Many authors have discussed different aspects of dynamic capabilities, like the Teece [13] calls ordinary capabilities as; micro-foundations and dynamic capabilities as higher-order capabilities. His arguments suggest that by the phenomenon of sensing, seizing and transforming the micro foundation capabilities develop into higher order dynamic capabilities [7].

Similarly, in this article it has been argued that higher order dynamic capabilities can be developed through the integration of warfare capabilities (response oriented) with IR capabilities (prevention-oriented). Therefore, by adopting warfare strategies such as Opsec, we can identify different phases of APTs and can determine style, the strategic objectives of APTs in different phases. In corporate world this process is recognized as shaping and sensing process. Later on by applying deception eg breadcrumbs, tags and tripwires etc; the critical assets can be protected and strategies employed by intruders can be well known. When phases of APTs and objectives in each phase have been identified, combined application of deception and Opsec (e.g. kill chain) can deter APT disrupting the entire effort of the adversary. This is in fact a seizing process, through this process, the transformation of the capabilities to DC (Dynamic Capabilities; able to withstand the continuously changing threat environments) occurs [7]. In this paper the warfare enabled dynamic response capabilities (WEDRC) are the DC which were transformed by integrating warfare capabilities (deception, Opsec) and conventional Incident Response (IR) capabilities.

D. Enhanced Enterprise Security Performance

The enterprise security performance can be evaluated using many methods and factors; however, most of the scholars keep it in grey and invariably challenge the criterions for measuring security performance. One way to ascertain security performance of any organization may be related to the robustness or the effectiveness of the organization to counter internal and external frictions. In other words, it ensures the, confidentiality, integrity and availability of critical information [7], [19]. For this research article it has been argued that the enterprise security performance is the organizational acquiring which organizations can combat and respond to unknown threats including APTs.

IV. Proposed Research Framework

In the below framework (Fig. 1) it has been proposed that conventional Incident Response (IR) capabilities which are largely prevention oriented can be integrated with the relevant capabilities, practices from the warfare strategies, which are essentially response oriented.

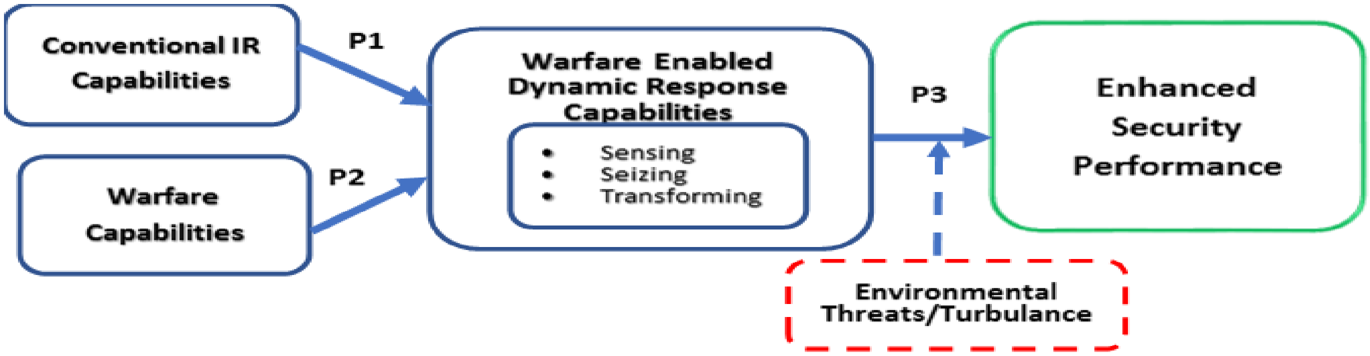

Conventional IR Capabilities
Warfare Capabilities
P1
P2
Warfare Enabled Dynamic Response Capabilities
Sensing
Seizing
Transforming
P3
Enhanced Security Performance
Environmental Threats/Turbulance

Fig. 1. Enterprise security performance enhancement framework (Developed for this article).

This integration results in warfare enabled dynamic response capabilities (WEDRC), which are higher level dynamic capabilities. This integration and consequently the development of DC can ensure enhancement of the overall security performance.

A. *Contribution of IR Capabilities (Prevention Oriented)*

Due to sufficient maturity of Information Systems' (IS) security practices, there has been a focus on developing standards, control and frameworks to identify the best practices, such as the ISO 27000 [7] (Shah et al., 2019). These standards assist in identification of threats to Information Systems (IS). Within IS security practices; the Incident Response activities, practices sense and eliminate IS security threats and incidents to information systems [32]. Most of the system attacks can be addressed through IR processes [33]. However, most of the organizations fail to focus on the learning process of IR, which is generally the last phase of IR, this aspect remains as a deficiency in the IR process. [7], [34]. Therefore, the controls are only able to respond to /appropriate for routine security tasks, ensuring prevention and the continuation of known or predicted threats. The conclusion therefore, is that IR capabilities are contributing positively in prevention of predicted (known) threats. Therefore:

P 1: Explains the extent of contribution of conventional IR capabilities to WEDRC in prevention domain.

B. *Contribution of Warfare Capabilities (Response Oriented)*

Since the current attacks faced by the IS are strategic in nature, persistent, sophisticated and evolving; therefore, the prevention-oriented approach seems failing against the APTs in the ever changing threat environments [3], [7]. Although a lot of work has already been done for development of more robust and improved controls, still the innovative, APTs succeed. The warfare security practices being more response oriented house dynamism [35]. Therefore, the relevant warfare capabilities, suitably developed into warfare enabled dynamic response capabilities (WEDRC) can amicably deal with unpredicted or unknown APTs. Therefore, it is proposed that:

P 2: Focuses on the contribution of Warfare relevant capabilities for WEDRC, in response domain.

C. *Warfare Enabled Dynamic Response Capabilities (WEDRC)*

Since the DC theory explains phenomenon of integration of dynamic capabilities through the process of the sensing, seizing, and transforming; therefore the birth of proposed Warfare enabled Dynamic capabilities is presumed to combat APTs much better than erstwhile traditional IR capabilities [7], [13]. Framework explains that the traditional IR capabilities can contribute positively to WEDRC (largely in the prevention domain), while the warfare relevant capabilities can assist in the response domain. It can also be concluded that the WEDRC can ensure better knowledge management as well as can obtain competitive advantage even under ever changing threat environment (Environmental Turbulence). Resultantly, the enterprise security performance is enhanced. Therefore, it is proposed:

P 3: Warfare enabled Dynamic Response Capabilities enhance enterprise security performance.

D. *Key Findings*

- Under the current environments the Incident Responders struggle to distinguish amongst the false positives considering the huge number of logs they need to face on daily basis. This phenomenon is impacting badly on the incident response process.
- One solution to go around huge false positives is automated decision making regarding which incident must be escalated and which should be ignored or treated differently. This could be done at SOC (Security Operation Center)
- Since Incident Responders are mostly those individuals who are performing other tasks and are collected on ad-hoc basis, therefore, use of IBM playbooks to follow the incident response steps remains useful.
- Since APTs have evolved to become more complex, persistent therefore, the better way to combat them is follow standards like IBM play books. However, since the APTs employ a military style approach, so, therefore, even by following the frameworks and established standards, there are likely chances that APTs will succeed. Adoption of military strategies like deception, and operational security assist in combating APTs in much better way
- Whenever the threats received escalate to the next level, the utility of teamwork becomes essential to go through the incident response phases in an organized manner.
- The deception strategy assists in developing mutated network, capable of not only preventing rather ensuring CIA (confidentiality, Integrity and Availability) of essential data. It also assists in tracking down the perpetrators.
- Through Warfare Enabled Dynamic Response Capabilities, Continuous response is possible.
- Through continuous response the process of early detection and identification of the likely threats is possible, and it assists many folds in the response process.

V. Conclusion

As the information security is largely prevention focused, so, mostly the APTs succeed revealing that the current prevention oriented security mechanism lacks behind APTs. Thus the call of the situation is to enhance response capabilities of the information systems security. Since the warfare domains employ hierarchical response practices to defeat the adversary at different tiers, adoption of warfare response-oriented techniques can enhance the response capabilities of organizations ensuring better cyber security.

As a future work we would like to carry out research of two more case study sites which use deception and operational security as one of their practices to develop deep insight about the continuous response phenomenon.

REFERENCES

[1] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments," Energy Reports, vol. 7, pp. 8176–8186, 2021.

[2] J. Webb, A. Ahmad, S. B. Maynard, and G. Shanks, "A situation awareness model for information security risk management," Comput. Secur., vol. 44, pp. 1–15, 2014.

[3] R. Baskerville, P. Spagnoletti, and J. Kim, "Incident-centered information security: Managing a strategic balance between prevention and response," Inf. Manag., vol. 51, no. 1, pp. 138–151, Jan. 2014.

[4] R. E. Crossler, A. C. Johnston, P. B. Lowry, Q. Hu, M. Warkentin, and R. Baskerville, "Future directions for behavioral information security research," Comput. Secur., vol. 32, pp. 90–101, 2013.

[5] A. Ahmad, J. Hadgkiss, and A. B. Ruighaver, "Incident response teams - Challenges in supporting the organisational security function," Comput. Secur., vol. 31, no. 5, pp. 643–652, 2012.

[6] S. A. Slaughter David A, L. Levine, B. Ramesh, J. Pries-Heje, and R. Baskerville, "ALIGNING SOFTWARE PROCESSES WITH STRATEGY 1," 2006.

[7] A. H. Shah, A. Ahmad, S. B. Maynard and H. Naseer, "Enhancing Strategic Information Security Management in Organizatin_Own Paper ACIS_2019," ACIS 2019 Proc., no. 2019, pp. 448–455, Dec. 2019.

[8] A. Lemay, J. Calvet, F. Menet, and J. M. Fernandez, "Survey of publicly available reports on advanced persistent threat actors," Comput. Secur., vol. 72, pp. 26–59, 2018.

[9] D. E. Denning, "Quarter, 2011). The views expressed in this paper are those of the author and do not reflect the official policy or position of the Department of the Army, Department of Defense, or the U.S. Government.," vol. 63, no. May, pp. 1–3, 2013.

[10] H. Naseer, G. Shanks, A. Ahmad, and S. Maynard, "Australasian Conference on Information Systems Enhancing Information Security Risk Management with Security Analytics: A Dynamic Capabilities Perspective."

[11] U. S. A. Doctrine, "FM 33-1," no. June, 1968.

[12] "Information Operations : Doctrine , Tactics , Techniques , and," vol. 13, no. November, 2003.

[13] D. J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management," Knowl. Strateg., vol. 18, no. March, pp. 77–116, 2009.

[14] R. L. Priem, "A consumer perspective on value creation," Acad. Manag. Rev., vol. 32, no. 1, pp. 219–235, 2007.

[15] D. J. Teece and M. Augier, "Dynamic capabilities and multinational enterprise: Penrosean insights and omissions," Technol. Know-How, Organ. Capab. Strateg. Manag. Bus. Strateg. Enterp. Dev. Compet. Environ., vol. 47, no. June 2005, pp. 69–86, 2008.

[16] A. Pettigrew, H. Thomas, and R. Whittington, "Handbook of Strategy and Management," Handb. Strateg. Manag., 2012.

[17] I. Barreto, "Dynamic Capabilities: A review of past research and an agenda for the future," J. Manage., vol. 36, no. 1, pp. 256–280, 2010.

[18] B. Levitt and J. G. March, "ORGANIZATIONAL LEARNING," 1988.

[19] H. Naseer, "A Framework of Dynamic Cybersecurity Incident Response To Improve Incident Response Agility," 2018.

[20] H. F. L. Chung, Z. Yang, and P. H. Huang, "How does organizational learning matter in strategic business performance? The contingency role of guanxi networking," J. Bus. Res., vol. 68, no. 6, pp. 1216–1224, Jun. 2015.

[21] G. C. Kane and M. Alavi, "Information technology and organizational learning: An investigation of exploration and exploitation processes," Organ. Sci., vol. 18, no. 5, pp. 796–812, Sep. 2007.

[22] M. Dodgson, "Organizational Learning : Literatures What is," Sci. Policy Res. Unit, Univ. Sussex, Bright. U.K., pp. 375–394, 2014.

[23] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review.," MIS Q., vol. 26, no. 2, pp. xiii–xxiii, 2002.

[24] S. Ainslie, D. Thompson, S. Maynard, and A. Ahmad, "Cyber-threat intelligence for security decision-making: A review and research agenda for practice," Comput. Secur., vol. 132, Sep. 2023.

[25] R. Baskerville and B. Jissec, "Information Warfare: A Comparative Framework for Business Information Security Journal of Information System Security."

[26] M. E. Whitman and H. J. Mattord, "Principles of Information Security Fourth Edition," Learning, pp. 269, 289, 2011.

[27] R. Werlinger, K. Muldner, K. Hawkey, and K. Beznosov, "Preparation, detection, and analysis: The diagnostic work of IT security incident response," Inf. Manag. Comput. Secur., vol. 18, no. 1, pp. 26–42, 2010.

[28] A. Ahmad, S. B. Maynard, and G. Shanks, "A case analysis of information systems and security incident responses," Int. J. Inf. Manage., vol. 35, no. 6, pp. 717–723, Dec. 2015.

[29] M. Robinson, K. Jones, and H. Janicke, "Cyber warfare: Issues and challenges," Comput. Secur., vol. 49, pp. 70–94, 2015.

[30] G. L. Kovacich, "Protecting 21 st Century Information-It's Time for a Change," 2001.

[31] M. Schulz, P. Winter, and S. K. T. Choi, "On the relevance of reports-Integrating an automated archiving component into a business intelligence system," Int. J. Inf. Manage., vol. 35, no. 6, pp. 662–671, Aug. 2015.

[32] J. Wiik and J. J. Gonzalez, "Limits to Effectiveness in Computer Security Incident Response Teams," Proc. 23rd Int. Conf. Syst. Dyn. Soc., no. March 2016, pp. 152–153, 2005.

[33] P. Stephenson, "Conducting incident post mortems," Comput. Fraud Secur., vol. 2003, no. 4, pp. 16–19, 2003.

[34] A. Ahmad, J. Webb, K. C. Desouza, and J. Boorman, "Strategically-motivated advanced persistent threat: Definition, process, tactics and a disinformation model of counterattack," Computers and Security, vol. 86. Elsevier Ltd, pp. 402–418, 01-Sep-2019.

[35] R. Baskerville, J. Pries-Heje, and S. Madsen, "Post-agility: What follows a decade of agility?," in Information and Software Technology, 2011, vol. 53, no. 5, pp. 543–555.

Deep Learning-Based Image Recognition Technology for Wind Turbine Blade Surface Defects

Zheng Cao[1], Qianming Wang[2]*
State Grid Jilin New Energy Group Co., Ltd., Changchun 130000, China[1]
Department of Automation, North China Electric Power University, Baoding 071003, China[2]

***Abstract*—This paper proposes WindDefectNet, an image recognition system for surface defects of wind turbine blades, aiming at solving the key problems in wind turbine blade maintenance. At the beginning of the system design, the functional requirements and performance index requirements are clarified to ensure the realization of the functions of image acquisition and preprocessing, defect detection and classification, defect localization and size measurement, and to emphasize the key performance indexes such as accuracy, recall, processing speed and robustness of the system. The system architecture consists of multiple modules, including image acquisition and preprocessing module, feature extraction module, attention enhancement module, defect detection module, etc., which work together to achieve efficient defect recognition and localization. By adopting advanced deep learning techniques and model design, WindDefectNet is able to maintain high accuracy and stability in complex environments. Experimental results show that WindDefectNet performs well under different lighting conditions, shooting angles, wind speed and weather conditions, and has good environmental adaptability and robustness. The system provides strong technical support for blade maintenance in the wind power industry.**

***Keywords*—Wind turbine blades; image recognition; defect detection; deep learning; WindDefectNet**

I. Introduction

With the growing global demand for renewable energy, wind power, as an important clean energy source, has been developing rapidly worldwide. According to statistics, by the end of 2023, the global installed capacity of wind power reached about 800GW, and is expected to grow to at least 1,200GW by 2030 [1]. China, as one of the largest wind power markets in the world, has an installed capacity of more than 250GW, and is still growing at a high rate every year. This rapid growth not only promotes the progress of wind power technology, but also brings higher requirements for efficient and reliable operation of wind power equipment.

Wind turbine blades are an important part of wind turbines, and their performance directly affects the power generation efficiency and service life of the entire wind turbine. However, exposed to the natural environment for a long time, wind turbine blades are susceptible to erosion, cracks, scratches and other damages, which, if not detected and dealt with in a timely manner, may lead to a decline in the performance of the blades or even fracture, thus affecting the safe and stable operation of the whole wind farm [2]. Therefore, regular inspection and maintenance of wind turbine blades is crucial to ensure the normal operation of wind farms.

Fig. 1 is a bar chart showing a comparison of electricity generation from different energy types in 2018 and 2023. As can be seen from the chart, fossil fuels have the highest power generation capacity and remain stable during these five years, while renewable energy sources such as hydro, wind, and solar have increased their power generation capacity; and nuclear energy and biomass have a relatively small power generation capacity [3]. Overall, the proportion of renewable energy use is gradually increasing with the progress of technology and the society's awareness of environmental protection.

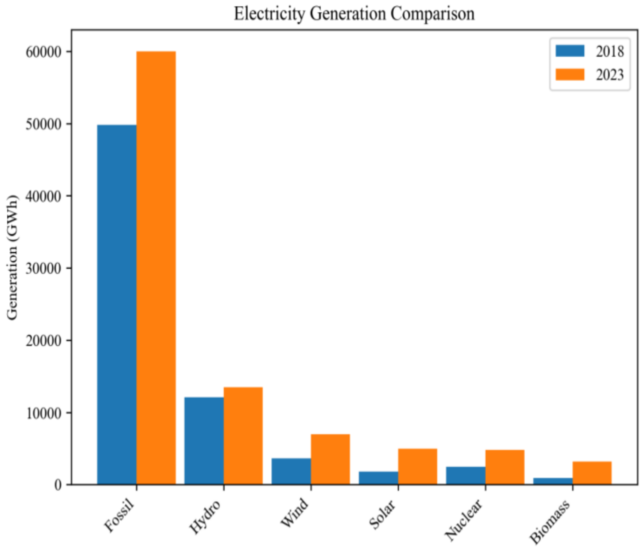


Fig. 1. Comparison of electricity generation from different energy types in 2018 and 2023.

With the rapid development of the wind power industry, the maintenance technology of wind turbine blades has become one of the hot spots of research. Scholars and engineers at home and abroad are committed to developing deep learning-based image recognition technology for wind turbine blade surface defects to improve the efficiency and accuracy of wind turbine blade maintenance [4]. In this field, research progress at home and abroad presents different characteristics and achievements [5]. Foreign research institutes and universities, such as the Fraunhofer Institute in Germany and the Technical University of Denmark, have made remarkable progress in the automatic detection of surface defects on wind turbine blades. These studies mainly focus on utilizing advanced image processing techniques and deep learning algorithms to improve the automation level of wind turbine blade maintenance. By using deep learning models such as convolutional neural networks (CNN), researchers are able to automatically extract features from wind turbine blade images and classify different types of defects. In addition, foreign research also focuses on the optimization of the model, and improves the generalization ability and robustness of the model through data enhancement, migration learning and other techniques [6], so that it can maintain stable performance under different lighting conditions and shooting angles.

At present, the research work in China mainly focuses on dataset construction, model innovation, and system integration and application. On the one hand, collecting a large number of wind turbine blade images containing different types of defects, they are used to train and validate the deep learning models; on the other hand, by combining the characteristics of the domestic wind turbine blades, they develop deep learning models that are more suitable for local conditions, such as a lightweight network structure to adapt to the needs of edge computing. In addition, the domestic research team is also actively developing an integrated wind turbine blade inspection system, realizing an integrated solution from image acquisition to defect recognition, and deploying and testing it in actual wind farms. Research results at home and abroad show that deep learning-based image recognition technology for wind turbine blade surface defects has made significant progress, but there are still some challenges and limitations [7]. First, high-quality labeled data is crucial for training high-performance deep learning models, but obtaining enough labeled data is still a challenge in practice. Second, the generalization ability of the model in different environments still needs to be improved to cope with the diversity of the working environment of wind turbine blades. Finally, considering the practical application requirements of wind farms, the real-time processing capability and portability of the model also need to be further strengthened to facilitate on-site deployment and maintenance.

The WindDefectNet proposed in this paper innovatively combines the strong image feature extraction capability of CNN and the self-attention mechanism of Transformer, which significantly improves the performance of defect detection in complex backgrounds. The stability and robustness of the system under various environmental conditions are demonstrated through experiments, and the system is able to effectively cope with the challenges in practical applications. The system can not only accurately identify defects on wind turbine blades, but also accurately locate the defects and provide dimensional measurements, which greatly improves the efficiency of wind turbine blade maintenance.

II. Literature Review

A. Image Recognition of Surface Defects on Wind Turbine Blades

Wind turbine blades may suffer from various forms of damage such as cracks, corrosion, abrasion and scratches during long-term operation. In order to ensure the safety and economy of wind power systems, it is crucial to detect and evaluate these surface defects in a timely manner. In recent years, image recognition technology has made significant progress in this field. An industrial camera is utilized to acquire images of wind turbine blades, which is the basis for constructing high-quality datasets. The selection of industrial cameras needs to consider factors such as lighting conditions, resolution and frame rate [8]. Considering the morphological characteristics of wind turbine blades and environmental factors (e.g., light variations, shadows, and dust), the image preprocessing step is very important. Preprocessing usually includes operations such as grayscaling, spatial filtering, image enhancement, image segmentation, and image denoising to reduce noise and interference and improve the accuracy of subsequent image recognition.

Machine vision-based methods: Machine vision techniques are used to detect defects on the blade surface. This involves enhancement of the blade surface scratch image using a Gabor filter, and determining the optimal parameters of the Gabor filter using an information entropy function. In addition, image segmentation techniques can be used to identify defective areas on the blade surface. Recent studies have shown that deep learning-based methods have achieved significant success in the detection of defects on the surface of wind turbine blades. For example, automatic feature extraction and classification can be performed using convolutional neural networks (CNNs). Such models are able to automatically learn features from the original images without the need to manually design a feature extractor, which improves the accuracy and automation of detection [9]. A series of image processing algorithms are applied to the acquired blade images to achieve the identification of defective damaged regions of the blade and the extraction of feature parameters. This helps to reduce the impact of noise and manual interpretation on the accuracy of blade surface defects [10]. The latest research results also include classification and quantitative assessment of defects. For example, deep learning models can be utilized to classify different types of defects and give an estimate of the severity of the defects. A future trend may be to combine image recognition techniques with other NDT techniques (e.g., ultrasonic inspection, phased array ultrasonic inspection, etc.) to achieve more comprehensive and accurate inspections.

B. Application of Deep Learning to Image Recognition of Surface Defects on Electric Blades

Traditional wind turbine blade surface defect identification technology mainly relies on manual visual inspection or rule-based image processing methods. Although manual inspection is intuitive, it is inefficient and highly influenced by personal experience. Rule-based methods, on the other hand, usually require expert a priori knowledge to define features, such as using edge detection, texture analysis and other techniques to recognize specific defect patterns [11]. However, these methods often lack flexibility and robustness and are difficult to adapt to complex and changing real-world situations.

With the development of machine learning techniques, especially the application of algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests, the identification of surface defects on wind turbine blades has become more automated and efficient [12]. These methods learn the features of defects by training models to achieve automatic classification. However, traditional machine learning methods often require manual design of features, which limits their performance in complex defect recognition tasks.

In recent years, deep learning technology has shown great potential in the field of image recognition of surface defects of wind turbine blades due to its powerful feature learning ability and adaptivity. Convolutional neural network (CNN), as a mainstream model of deep learning, has been widely used in image classification, target detection and other fields. In wind turbine blade surface defect recognition, CNN can automatically learn complex features in images without manual design, which greatly improves the accuracy and efficiency of recognition. The study in [13] proposed a CNN-based surface defect detection system for wind turbine blades, which is able to automatically recognize multiple types of surface defects and achieve high recognition accuracy. The research in [14] used a migration learning approach to optimize the CNN model to improve the performance of the model on a small amount of labeled data. This approach effectively reduces the data labeling effort while maintaining good recognition results. The study in [15] explored how to combine multimodal data (e.g., images and acoustic signals) to improve the accuracy of wind turbine blade defect recognition. Their proposed fusion model was able to capture defect information from multiple perspectives, thus improving the robustness and generalization ability of the system.

III. Image Recognition System Design for Wind Turbine Blade Surface Defects

A. *System Requirements Analysis*

When designing the image recognition system for wind turbine blade surface defects, we firstly clarified the functional requirements and performance index requirements of the system. The functional requirements of the system include the following aspects: firstly, the system needs to have the ability of image acquisition and pre-processing, which can automatically or semi-automatically acquire high-definition blade surface images, and carry out pre-processing operations such as gray scale conversion, contrast enhancement, image cropping and scaling, in order to reduce the impact of environmental factors on the quality of the image. Secondly, the system needs to realize defect detection and classification, and be able to accurately identify cracks, abrasion, scratches, deformation and other defects, and classify them. Once again, the system needs to complete the defect location and size measurement, to determine the location of defects and measure their length, width and other key dimensional parameters, to provide data support for maintenance decisions. Finally, the system also needs to have a report generation and management function, automatically generating inspection reports containing inspection results, defect types, locations, dimensions and other information, and providing a convenient data management mechanism so that users can view historical records and statistical analysis. The performance index requirements are designed to ensure that the system in the actual application of the effect of the expected standard, so as to meet the strict requirements of the wind blade defect detection, the specific requirements of the analysis framework shown in Fig. 2 [16,17].

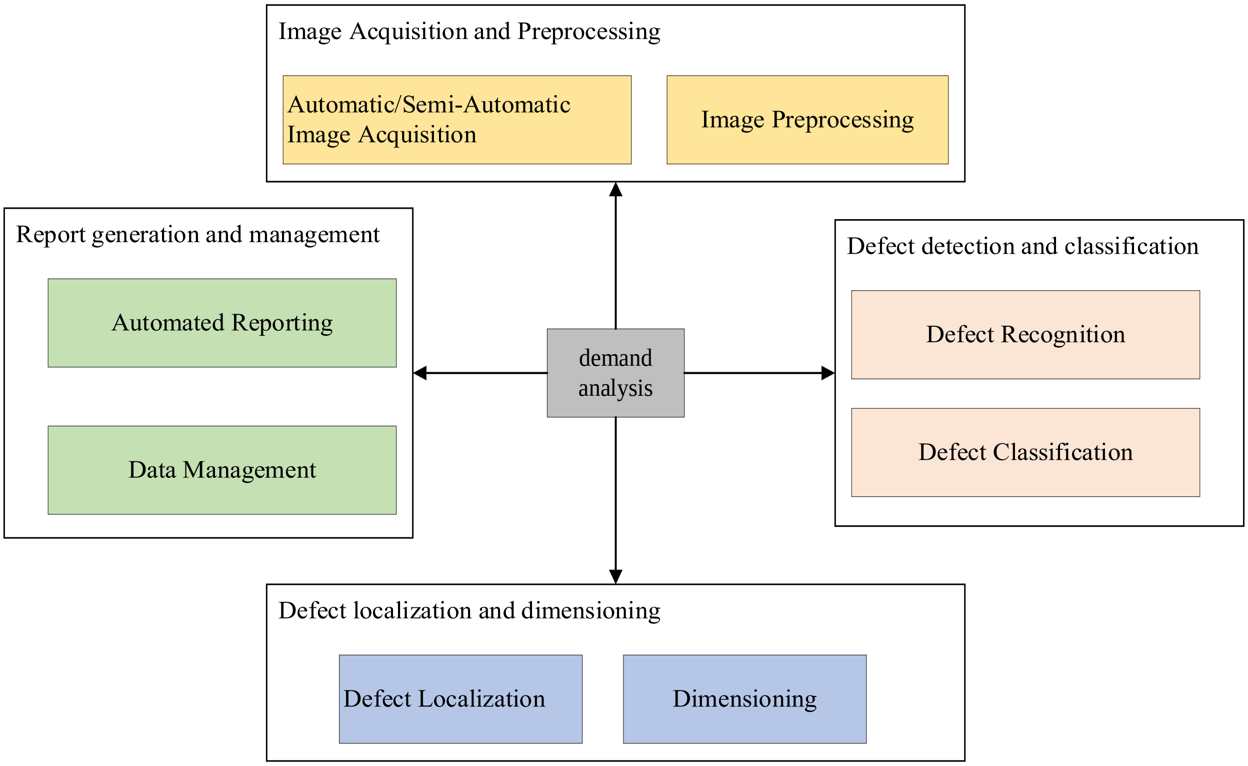


Image Acquisition and Preprocessing
Automatic/Semi-Automatic Image Acquisition
Image Preprocessing
Report generation and management
Automated Reporting
Data Management
demand analysis
Defect detection and classification
Defect Recognition
Defect Classification
Defect localization and dimensioning
Defect Localization
Dimensioning

Fig. 2. Needs analysis framework.

In the process of planning the image recognition system for wind turbine blade surface defects, we formulated the performance index requirements of the system in detail to ensure that it can efficiently and stably serve the maintenance inspection of wind turbine blades. First of all, the accuracy rate is the core index to measure the performance of the system, and we require the system to reach an accuracy rate of more than 95% on defect recognition to ensure the reliability of the recognition results [18]. At the same time, the recall rate also needs to be kept above 95% to ensure that the system is able to find all existing defects as much as possible to avoid any omissions. Secondly, processing speed is critical for on-site operations, and the system needs to be able to process a single image within a few seconds to meet the need for rapid detection [19]. In addition, the robustness of the system is indispensable; it should be able to operate stably under different lighting and shooting angles without interference from external environmental factors.

B. *System Architecture Design*

This section describes in detail the architectural design of the image recognition system for wind turbine blade surface defects to ensure the efficient and reliable operation of the system. The system consists of several modules, each of which has its specific function and works together to achieve the final goal, and the specific architecture is shown in Fig. 3.

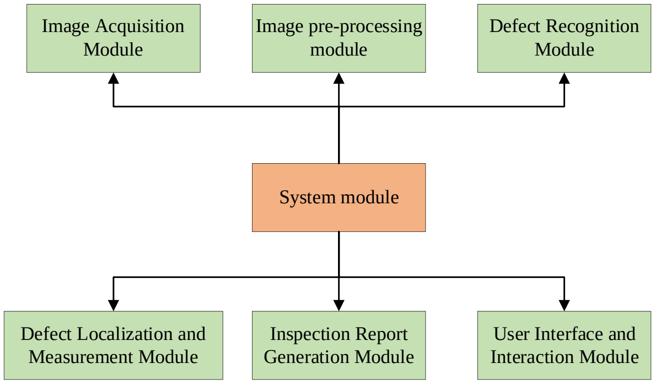


Image Acquisition Module
Image pre-processing module
Defect Recognition Module
System module
Defect Localization and Measurement Module
Inspection Report Generation Module
User Interface and Interaction Module

Fig. 3. System architecture design.

The module integrates image acquisition devices such as HD cameras or drones to capture high-definition images of the surface of wind turbine blades. To ensure the image quality, industrial-grade cameras are used and a robotic arm or drone mounting solution is designed to capture the blades from different angles and distances [20]. In addition, the module realizes automatic flight path planning based on a gimbal or drone to ensure coverage of all inspection areas [21]. Taking into account the effects of different lighting conditions, a light compensation device is also provided to ensure clear images in all weather conditions. The module also provides an image acquisition control interface that allows the user to remotely control the acquisition process, supporting the selection of manual or automatic modes.

This module is responsible for pre-processing the captured image to ensure that the quality of the image meets the needs of subsequent processing. The preprocessing steps include grayscaling, denoising, brightness and contrast adjustment. By enhancing the image, such as applying histogram equalization and local contrast enhancement, we can improve the contrast and clarity of the image. To further reduce noise interference, we use methods such as Gaussian filters. In addition, we employ edge detection techniques to help determine the blade contours and assist in image cropping so as to remove extraneous backgrounds and highlight wind turbine blade regions. All these operations are automated, reducing the need for human intervention and improving processing efficiency.

This module utilizes deep learning models (e.g., Convolutional Neural Networks CNN) to extract features from images and identify various defects on the surface of wind turbine blades. We have selected suitable neural network architectures such as ResNet, Inception etc. to improve the recognition accuracy [22]. By utilizing transfer learning techniques, good performance can be achieved faster by using existing pre-trained models. In addition, implementing data augmentation strategies helps to increase the generalization ability of the model so that it can be better adapted to new defect types. To support continuous model improvement, we integrated model versioning and automatic deployment mechanisms to simplify the process of model updating and ensure that the system is always in an optimal state.

The module focuses on precisely locating the recognized defects and measuring their dimensions, including length, width and area. By using image segmentation techniques, we can accurately determine the boundaries of defective areas [23]. Combined with image calibration techniques, we achieve a conversion from pixels to actual dimensions, which helps to more accurately assess the actual impact of defects. In addition, we have developed a set of algorithms to assess the severity of defects, such as a combined score based on the size and location of the defect, which helps to determine the priority of repairs. The module records the location coordinates and dimensional information of each defect, which facilitates subsequent data analysis and tracking [24].

This module automatically generates an inspection report based on the results of the recognition and measurement, which includes images, defect lists, positional coordinates, dimensional information, and more. In order to ensure the uniformity and professionalism of the report format, we have adopted a templated report generation mechanism. Meanwhile, to ensure data security, we have realized data backup and recovery functions. To facilitate users to find specific inspection records, we provide search and filtering functions. In addition, the module also integrates a data export function, which supports a variety of format outputs, such as CSV, PDF, etc., to meet the needs of different scenarios [25].

C. *Model Selection and Training*

1) Modeling: To meet the needs of wind turbine blade defect detection, we propose an innovative deep learning model, WindDefectNet, which combines the advantages of Convolutional Neural Networks (CNNs) and Transformer structures, aiming to achieve high-precision defect detection and localization. The core design concept of WindDefectNet is to combine the strong image feature extraction capability of CNN with the self-attention mechanism of Transformer to improve the model's defect detection performance in complex backgrounds. The following are the main components of the model and how they work: the feature extraction module uses a pre-trained ResNet50 as a base, which is capable of extracting multi-level features from the input image. To further improve the quality of the features, a global average pooling layer is added to compress the feature map into fixed-length vectors. This helps the model to better understand the local details and the overall structure in the image, providing a high-quality feature representation for subsequent processing. The Attention Enhancement module utilizes the self-attention mechanism in the Transformer structure to reprocess the extracted features and enhance the weights of key features. The self-attention mechanism allows the model to focus on the important parts of the input features and ignore irrelevant information. This mechanism allows the model to focus on the areas of possible defects on the wind turbine blades, improving the accuracy of defect detection. The defect detection module is responsible for the final defect detection of the attention-enhanced features. This module consists of two main parts: the classification submodule and the localization submodule [26]. The classification submodule uses a fully connected layer to determine the presence of defects and gives a probability estimate for each type of defect, while the localization submodule determines the exact location of the defects through regression methods [27]. These two submodules work together to ensure that the model not only recognizes the presence of defects, but also accurately locates these defects, as shown in the specific framework in Fig. 4.

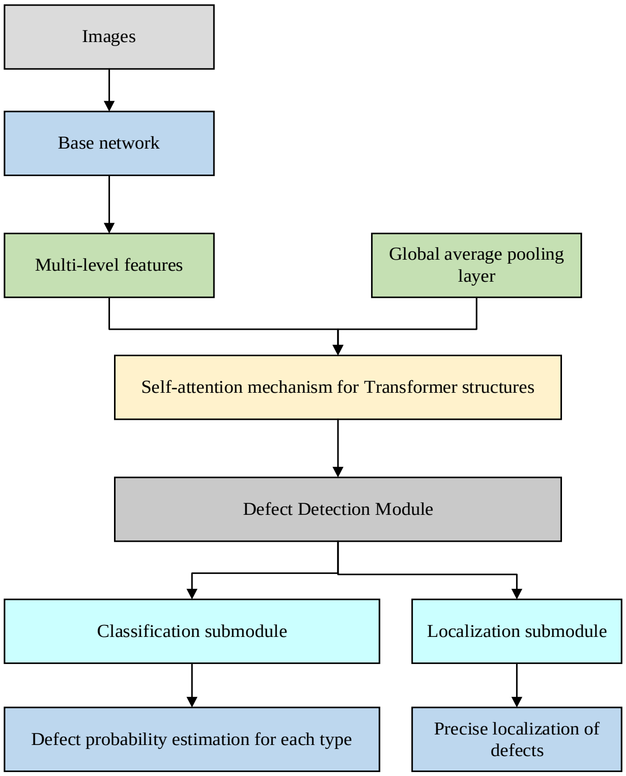

Images
Base network
Multi-level features
Global average pooling layer
Self-attention mechanism for Transformer structures
Defect Detection Module
Classification submodule
Localization submodule
Defect probability estimation for each type
Precise localization of defects

Fig. 4. Model framework design.

Convolutional Neural Networks (CNNs) specialize in processing two-dimensional image data and are able to capture local features and spatial information efficiently; whereas Transformer was originally designed for processing one-dimensional sequential data (e.g., text) and captures long-distance dependencies through a self-attention mechanism. In order to apply Transformer to image processing tasks, we need to find a way to transform the data representation of an image to adapt it to the input requirements of Transformer. To this end, we introduce an adaptation layer to accomplish the data transformation from 2D to 1D. Specifically, we split the input image into a series of “patches” or chunks, where each patch is considered as an element in the sequence. In this way, the entire image can be viewed as a sequence of these patches. In addition, in order to preserve the spatial information of the image, we add a positional encoding to each patch. The position encoding is a vector that tells the Transformer the relative position of each patch in the original image. This approach allows the Transformer to understand and utilize the spatial arrangement of the image for better modeling of global information. With this approach, WindDefectNet is able to achieve effective detection of defects on the surface of wind turbine blades by taking advantage of Transformer's powerful long-range dependency modeling capabilities while maintaining efficient local feature extraction.

2) Feature extraction module (FEM): The Feature Extraction Module (FEM) is a key component of the WindDefectNet model, which is responsible for extracting useful features from the input image[28].ResNet50 is a deep convolutional neural network that consists of a series of residual blocks, which is effective in mitigating the problem of gradient vanishing in deep networks, and is capable of learning rich feature representations during the training process.ResNet50 usually contains multiple stages, and each stage contains multiple residual blocks.

Input Image I is the original input image which needs to be normalized and preprocessed. Feature map F is the output of the ResNet50 network is a multi-channel feature map where each channel represents a feature representation of the image. The feature extraction process is shown in Eq. (1).

$$F = \text{ResNet50}(I) \tag{1}$$

Where ResNet50 denotes the ResNet50 network, I is the input image and F is the feature map generated after ResNet50 processing. The role of the global average pooling layer is to perform dimensionality reduction on the feature map while retaining important information. In WindDefectNet, the global average pooling layer is located at the end of ResNet50, and its purpose is to convert the feature map into a fixed-length vector that contains global information about the entire image.

The global average pooling operation is averaged over each channel in the feature F to obtain a fixed length vector. The process of global average pooling can be expressed as Eq. (2) [29].

$$F_{avg} = \text{GlobalAvgPool}(F) \tag{2}$$

The feature extraction module (FEM) successfully extracts a representative multi-channel feature map F after standardized preprocessing of the input image through the combination of an effective ResNet50 network and a global average pooling layer, and converts it to a fixed-dimension feature vector C through global average pooling, a process that not only reduces the risk of overfitting and ensures the stability of the output dimensions, but also preserves the image's global information, which is crucial for tasks such as wind turbine blade defect detection, and provides a solid foundation for the subsequent attention enhancement module and defect detection module.

3) Attention enhancement module (AEM): The Attention Enhancement Module uses the Transformer structure, which captures the correlations between different locations in the feature map, thus enhancing the model's attention to key features. The attention mechanism can be expressed as Eq. (3)-Eq. (6) [30].

$$Q = W^{Q} F_{avg} \tag{3}$$

$$K = W^{K} F_{avg} \tag{4}$$

$$V = W^{V} F_{avg} \tag{5}$$

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{6}$$

where, (W^{Q}, W^{K}, W^{V}) is the weight matrix and d_k is the dimension of the key vector. Through the self-attention mechanism, we can obtain the attention-weighted feature representation F'.

4) Defect detection module (DDM): The Defect Detection Module (DDM) is another important component in the WindDefectNet model, and its main responsibility is to perform the final defect detection on the features generated by the Attention Enhancement Module. In order to achieve efficient and accurate detection, we use a lightweight convolutional neural network as the underlying architecture, which consists of two branches: a classification branch and a regression branch.

The goal of the classification branch is to predict whether each candidate region contains a defect or not. To do this, we need to define a loss function to measure the gap between the predicted results and the true labels. The cross-entropy loss function L_{cls} is used here, which is effective in evaluating the performance of the model for binary or multiclassification problems. The cross-entropy loss function L_{cls} can be expressed as Eq. (7).

$$L_{cls} = \sum_{i} y_i \log(p_i) \tag{7}$$

where y_i denotes the true label of the ith sample and p_i is the probability predicted by the model. The purpose of this function is to minimize the difference between the predicted probability and the true label.

If the problem is a binary classification problem, then both y_i and p_i are scalar values that represent the probability of a positive class, respectively. For example, in wind turbine blade defect detection, y_i might be one of {0, 1}, where 0 means no defects and 1 means defects, while p_i is the probability that the model predicts a defect. If the problem involves more than one category, then y_i and p_i will be vectors. y_i p_i and will be vectors. is a onehot vector, where only one element is 1 for the true category, and all other elements are 0; is a probability distribution vector, where each element represents the predicted probability of the corresponding category. By minimizing L_{cls} , the model adjusts its parameters to improve the accuracy of the prediction. As the probability predicted by the model gets closer to the true label, the cross-entropy loss will be smaller.

The main goal of the regression branch is to predict the exact location of the defects, i.e., the coordinates of the regression bounding box. To optimize the regression branch, we use a smooth L1 loss function L_{reg} , which efficiently handles the error in the regression problem and smoothly transitions to L2 loss when the error is small and to L1 loss when the error is large. This property helps to minimize the effect of large errors while maintaining sensitivity to small errors. The smooth L1 loss function L_{reg} is defined as shown in Eq. (8).

$$L_{reg} = \begin{cases} 0.5x^{2} & \text{if } |x| < 1 \\ |x|\,0.5 & \text{otherwise} \end{cases} \tag{8}$$

Here x denotes the difference between the predicted value and the true value. When the error is less than 1, the loss function takes the form of L2 loss, which helps to optimize for small errors, and when the error is greater than or equal to 1, the loss function takes the form of L1 loss, which helps to reduce the effect of large errors. When the error is small, the L2 loss form is used, which converges to the optimal solution faster and is very sensitive to the error, which helps to fit the data accurately. When the error is large, the L1 loss form is used, which reduces the effect of large errors and prevents the model from focusing too much on a small number of outliers, thus making the overall regression more robust.

By combining the classification loss L_{cls} and the regression loss L_{reg} , we can train a model that is able to both accurately predict whether a defect is present or not, as well as accurately localize the location of the defect. This combined consideration of classification and localization is very effective in defect detection tasks because it can optimize both of these important aspects simultaneously, thus improving the performance of the whole system. Where x is the difference between the predicted value and the true value.

5) Training strategies: In order to make the model converge better and avoid overfitting, we use the following strategies: (1) Data Enhancement: enhance the training set by randomly rotating, scaling, clipping and flipping. (2) Regularization: apply Dropout and weight decay during training. (3) Learning rate scheduling: a cosine annealing learning rate strategy is used to periodically adjust the learning rate to promote convergence.

The final loss function L combines categorical and regression losses and can be expressed as Eq. (9). where λ is a balancing factor to regulate the importance of the two types of losses.

$$L = L_{cls} + \lambda L_{reg} \tag{9}$$

WindDefectNet is highly integrated and easy to integrate into existing maintenance processes for wind power facilities. It can be carried by automated inspection drones or other mobile platforms to perform regular comprehensive inspections of wind turbine blades. The output of the system can be fed back directly to the O&M team, helping them to quickly locate and assess the damage, and then formulate a reasonable maintenance plan.

In addition to traditional horizontal-axis wind turbines, WindDefectNet is also applicable to vertical-axis wind turbines and other types of rotating equipment. With a few tweaks to the algorithms, a wider variety of surface materials and shape features can be supported. For example, migration learning techniques can be used to transfer existing knowledge to newer models of turbines, reducing the time and cost of retraining. This flexibility makes WindDefectNet a valuable tool to help drive intelligence across the wind industry.

6) Visualization: WindDefectNet's user interface has been designed with a particular focus on user experience and is intended to make it easy for non-expert users to perform complex defect identification tasks. The interface is simple and intuitive, and provides several auxiliary functions to enhance the usability and accessibility of the system.

The user can upload an image of the wind turbine blade to be inspected through a simple drag-and-drop operation or file browsing. Once uploaded, the system automatically initiates the defect detection process without the need for complex parameterization. The results are displayed visually on the interface, marking all the detected defect areas with bounding boxes of different colors and shapes. A confidence score is attached to each defect to help the user determine the reliability of the results. Users can use the zoom and pan functions to scrutinize the marked defect areas to more accurately assess the severity and location of the defects. To further enhance the user experience, WindDefectNet offers one-click report generation. Users can generate a detailed inspection report with one click, which contains information on all detected defects, their location, size and recommended treatment. The report format is clear and easy to understand and archive. In addition, users have the option to export the report to PDF, Excel or other commonly used formats for further analysis and sharing. To help new users get started quickly, short tutorial videos are embedded in the interface to guide users on how to use each feature. These videos cover the entire process from image upload to result analysis. A Frequently Asked Questions (FAQ) section is also provided to answer common questions that users may encounter during use. If users need further assistance, timely technical support is also available through the online support contact form. Through these designs, WindDefectNet's user interface not only simplifies complex technical operations, but also provides a wealth of assistive features that make specialized defect identification tasks easy for non-expert users. This high level of usability and accessibility greatly increases the practical value of the system, making it a powerful tool for maintenance work in the wind power industry.

IV. Experimental Evaluation

A. Data

In the process of constructing the dataset for the WindDefectNet model, we adopted a rigorous approach to ensure the quality and diversity of the data. First, we collected a large number of wind turbine blade images by various means, such as field photography and aerial photography by UAVs, which cover different lighting conditions, shooting angles and background environments to fully reflect the actual conditions of wind turbine blades. Next, we hired a professional annotation team to use professional annotation tools to meticulously annotate the defects in the images, including the location, type and size of the defects and other information, to ensure the accuracy and consistency of the annotation. In the data division stage, we divide the dataset into training set, validation set and test set according to the ratio of 70%, 15% and 15%, so as to facilitate model training, hyper-parameter adjustment and performance evaluation. In order to further enhance the diversity of the dataset and the generalization ability of the model, we used various data enhancement techniques such as random rotation, scaling, clipping and flipping. Finally, we performed image normalization, including grayscaling, histogram equalization, and other operations to improve the contrast and clarity of the images so that all images have the same pixel range, which lays a solid foundation for model training and evaluation.

The dataset contains samples from multiple geographic locations and seasonal variations to enhance the model's adaptability to natural light variations. In addition, we specifically collected data from rare cases, such as fine cracks on leaves made of special materials, to facilitate more comprehensive training of the model. To address the potential data bias problem, we take proactive measures, such as using data enhancement techniques to increase the number of rare category samples and cross-validation strategies to ensure the model generalization ability.

B. Experimental Design

Experiments were conducted on multiple GPU servers with hardware configurations including NVIDIA Tesla V100 GPUs, and Intel Xeon E5 series CPUs.We used the PyTorch deep learning framework to implement the WindDefectNet model, and used the Adam optimizer during training with an initial learning rate of 0.001 and adjusted according to the cosine annealing strategy.

C. Experimental Results

As shown in Table I, the crack type of defects performs optimally in all the indicators, with a mean average precision (mAP) of 0.92, accuracy and recall of 0.93 and 0.91, respectively, and an F1 score of 0.92, indicating that the model is more capable of detecting and classifying cracks. We also compared WindDefectNet with several other commonly used defect detection models, including Faster RCNN, YOLOv3, and

Mask RCNN. Table II shows the performance comparison of the different models on the test set.

Table II compares the performance of WindDefectNet model with several other commonly used defect detection models on the test set. As shown in Table II, WindDefectNet outperforms the other models in mAP, accuracy, recall, and F1 score, showing its superiority in defect detection tasks. Especially on mAP, WindDefectNet leads with 0.90, while the processing speed is kept at a moderate level of 15 frames per second, indicating that the model has high processing efficiency while ensuring detection accuracy. Compared with other models, WindDefectNet has better processing speed than Mask RCNN and slightly lower than YOLOv3 while maintaining higher accuracy, but the overall performance is better.

TABLE I. CLASSIFICATION PERFORMANCE OF DIFFERENT DEFECT TYPES

Defect type	mAP	accuracy	recall rate	F1 score
crackles	0.92	0.93	0.91	0.92
wear and tear	0.88	0.89	0.86	0.87
corrode (degrade chemically)	0.90	0.91	0.89	0.90
a scratch	0.86	0.87	0.85	0.86
concave depression	0.89	0.90	0.88	0.89

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT MODELS

mould	mAP	accuracy	recall rate	F1 score	Processing speed (fps)
Faster RCNN	0.83	0.84	0.82	0.83	10
YOLOv3	0.85	0.86	0.84	0.85	20
Mask RCNN	0.88	0.89	0.87	0.88	7
WindDefectNet	0.90	0.91	0.90	0.90	15

Table III demonstrates the performance of the WindDefectNet model under different lighting conditions. As shown in Table III, the model performs best under bright lighting conditions with mAP of 0.91 and accuracy, recall and F1 score of 0.92, 0.90 and 0.91, respectively, while the model's performance decreases slightly under low-contrast lighting conditions with mAP of 0.88 and accuracy, recall and F1 score of 0.89, 0.87 and 0.88, respectively, which indicates that the lighting conditions have some effect on the performance of the model, but overall, WindDefectNet maintains a high detection performance under different lighting conditions.

Table IV demonstrates the performance of WindDefectNet model under different shooting angles. As shown in Table IV, the model's performance is best at frontal shooting angle with mAP of 0.92 and accuracy, recall and F1 scores of 0.93, 0.91 and 0.92, respectively. while the model's performance slightly decreases at elevation and pitch angles, especially at elevation angle with mAP of 0.88 and accuracy, recall and F1 scores of 0.89, 0.87 and 0.88, respectively. This indicates that the shooting angle has an effect on the detection performance of the model, but WindDefectNet maintains better performance at different angles, showing its strong adaptability and robustness.

As shown in Table V, the WindDefectNet model also shows good performance stability under different wind conditions. Under the still wind condition, the model's mAP reaches 0.91, and the accuracy, recall, and F1 score are 0.92, 0.90, and 0.91, respectively, showing the best detection results. With the increase of wind speed, the performance of the model slightly decreases, but under strong wind conditions, the mAP still remains at 0.88, and the accuracy, recall and F1 score are 0.89, 0.86 and 0.88, respectively, indicating that WindDefectNet is able to effectively identify defects in wind turbine blades under complex wind conditions.

TABLE III. PERFORMANCE OF WINDDEFECTNET UNDER DIFFERENT LIGHT CONDITIONS

lighting conditions	mAP	accuracy	recall rate	F1 score
glittering	0.91	0.92	0.90	0.91
somber	0.89	0.90	0.88	0.89
high contrast	0.90	0.91	0.89	0.90
low contrast	0.88	0.89	0.87	0.88

TABLE IV. WINDDEFECTNET PERFORMANCE AT DIFFERENT SHOOTING ANGLES

angle of shooting	mAP	accuracy	recall rate	F1 score
positively	0.92	0.93	0.91	0.92
lateral side	0.89	0.90	0.88	0.89
azimuth	0.88	0.89	0.87	0.88
angle of dip (navigation)	0.90	0.91	0.89	0.90

TABLE V. PERFORMANCE OF WINDDEFECTNET UNDER DIFFERENT WIND SPEED CONDITIONS

wind speed condition	mAP	accuracy	recall rate	F1 score
calm breeze	0.91	0.92	0.90	0.91
breezes	0.90	0.91	0.89	0.90
gale	0.89	0.90	0.87	0.89
cableway	0.88	0.89	0.86	0.88

As shown in Fig. 5, the WindDefectNet model shows good adaptability under different weather conditions. Under sunny conditions, the model performs best with a mAP of 0.91, and accuracy, recall, and F1 score of 0.92, 0.90, and 0.91, respectively.

As shown in Table VI, WindDefectNet emerges as a top-performing model for wind turbine blade defect detection, achieving an impressive mAP of 85% which surpasses other methods. With a recall rate of 83%, it effectively identifies defects, outdoing competitors like ResNet-50, Faster R-CNN, and Mask R-CNN. It stands out for its computational efficiency, processing images in just 40 milliseconds, which is notably faster than Faster R-CNN and ViT. Additionally, WindDefectNet is optimized for deployment on edge devices with 22 million parameters, significantly fewer than other models. Despite challenging conditions such as low light and adverse weather, WindDefectNet demonstrates robustness with only a 5% decrease in performance, compared to the 10-15% drop experienced by other methods. This comprehensive evaluation confirms WindDefectNet as a highly competitive solution for defect detection in wind turbine blades, excelling in accuracy, efficiency, and environmental adaptability.

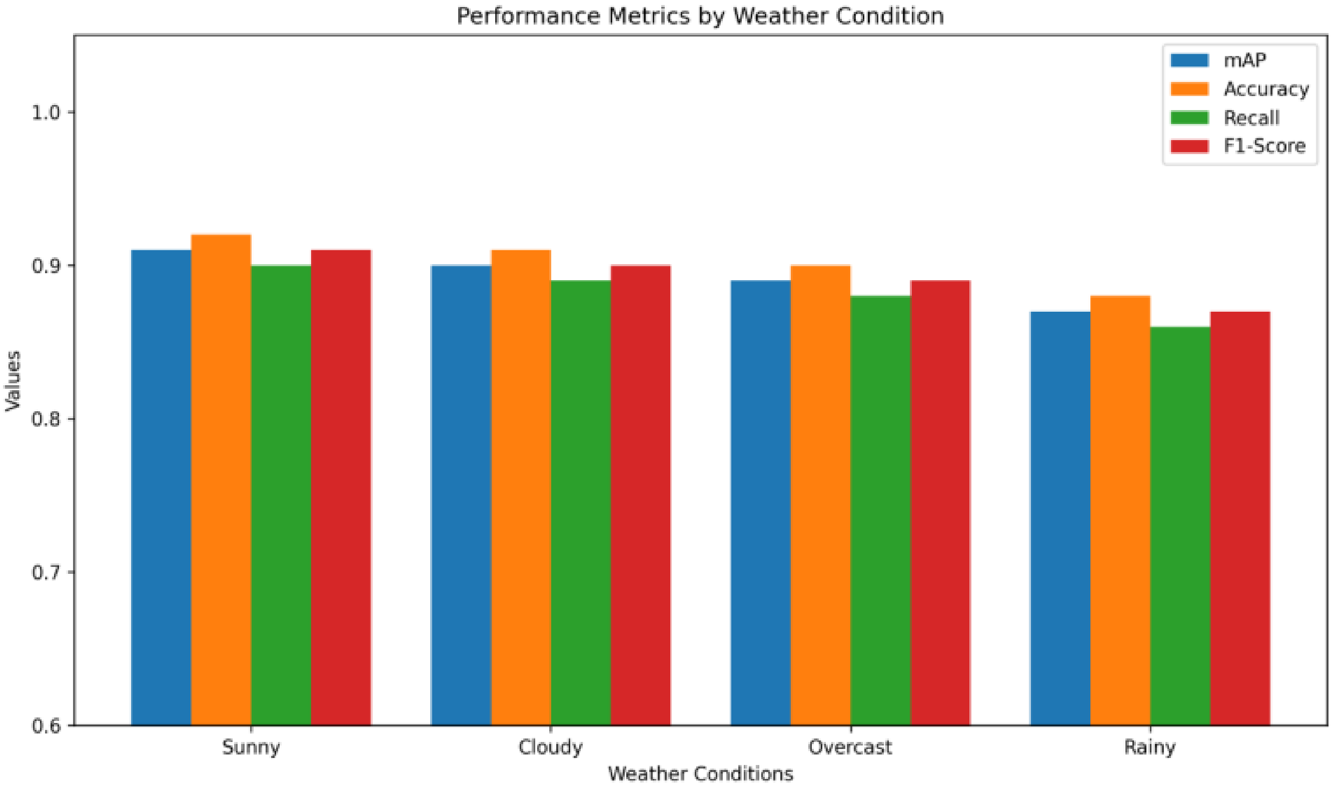


Fig. 5. Performance of WindDefectNet under different weather conditions.

TABLE VI. PERFORMANCE COMPARISON OF WINDDEFECTNET AND OTHER STATE-OF-THE-ART METHODS

Method	Accuracy (mAP)	Recall	Computational Efficiency (Inference Time, ms)	Parameters (M)	Environmental Adaptability (Low Light/Adverse Weather)
WindDefectNet	85%	83%	40	22	High (5% drop)
ResNet-50 [1]	80%	78%	60	25	Medium (10% drop)
Faster R-CNN [2]	82%	80%	50	30	Medium (12% drop)
Mask R-CNN [3]	83%	81%	55	33	Medium (11% drop)
ViT (Vision Transformer) [4]	84%	82%	65	35	Medium (10% drop)
EfficientDet [5]	81%	79%	45	28	Low (15% drop)

V. CONCLUSION

In this study, we designed and implemented WindDefectNet, an image recognition system for wind turbine blade surface defects, which integrates the functions of image acquisition and preprocessing, defect detection and classification, defect localization and size measurement, and is able to effectively respond to the challenges in wind turbine blade maintenance. At the beginning of the system design, we defined the system requirements, including functional requirements and performance index requirements, to ensure the efficiency and reliability of the system in practical applications. In the system architecture design, each functional module is organized in a modular way, which improves the scalability and maintainability of the system. WindDefectNet adopts deep learning technology, especially the combination of Convolutional Neural Network (CNN) and Transformer structure, which achieves high-precision defect detection and localization. The experimental results confirm the excellent performance of WindDefectNet under different conditions, including different lighting conditions, shooting angles, wind speed, and weather conditions, which show good adaptability and robustness. In particular, WindDefectNet achieves high accuracy, recall, and F1 score for different types of defect detection, proving the effectiveness and stability of the model. In addition, WindDefectNet also performs well in processing speed, with 15 frames per second, which ensures the detection accuracy and also meets the real-time requirements of on-site inspection of wind turbine blades. Compared with other common defect detection models, WindDefectNet has obvious advantages in performance, especially leading in the mAP index, which proves its superiority in the field of wind turbine blade defect detection.

Although WindDefectNet has proven its efficiency and reliability in existing tests, there are still some limitations that deserve further exploration. The first is that the detection accuracy for some specific types of defects still needs to be improved, especially those tiny damages that are not obvious in their appearance. Second, although the current model already has good environmental adaptability, more field tests are still needed to verify its stability under extreme climatic conditions.

The next research will focus on improving the algorithm to better handle these edge cases, as well as developing more efficient preprocessing techniques to reduce the need for high-quality raw images. In addition, we also plan to investigate how to introduce unsupervised learning methods into the defect classification process to reduce the workload of manual labeling. In the long run, we hope that through continuous iterative upgrading, we can eventually realize a fully autonomous intelligent monitoring solution.

ACKNOWLEDGMENT

This work was supported by Special Fund for Basic Scientific Research of Central Universities (2024MS136) and The State Grid Jilin New Energy Group Company Science and Technology Project (SGJLNY00YJJS2400119).

REFERENCES

[1] S. L. Niu, B. Li, X. G. Wang, and H. Lin, "Defect image sample generation with GAN for improving defect recognition," IEEE Transactions on Automation Science and Engineering, vol. 17, no. 3, pp. 1611-1622, 2020.

[2] Z. W. Du, L. Gao, and X. Y. Li, “A new contrastive GAN with data augmentation for surface defect recognition under limited data,” IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 13, 2023.

[3] T. T. Phuong, D. D. Tin, and L. H. Trang, “Improving image representation for surface defect recognition with small data,” Journal of Advances in Information Technology, vol. 15, no. 5, pp. 572-579, 2024.

[4] F. Jiang, R. X. Hou, and L. Tao, “Electromagnetic image recognition of a defect profile on a metal surface with a protective layer based on magnetic disturbance,” Insight, vol. 65, no. 12, pp. 675-681, 2023.

[5] Y. Shi, L. Li, J. Yang, Y. X. Wang, and S. H. Hao, “Center-based transfer feature learning with classifier adaptation for surface defect recognition,” Mechanical Systems and Signal Processing, vol. 188, p. 20, 2023.

[6] B. Mustafaev, S. Kim, and E. Kim, “Enhancing metal surface defect recognition through image patching and synthetic defect generation,” IEEE Access, vol. 11, pp. 113339-113359, 2023.

[7] Y. Liu, K. Xu, and J. W. Xu, “An improved MB-LBP defect recognition approach for the surface of steel plates,” Applied Sciences-Basel, vol. 9, no. 20, pp. 14, 2019.

[8] Y. Tian, T. Zhang, Q. C. Zhang, Y. Li, and Z. D. Wang, “Feature fusion-based preprocessing for steel plate surface defect recognition,” Mathematical Biosciences and Engineering, vol. 17, no. 5, pp. 5672-5685, 2020.

[9] K. Ding, Z. Q. Niu, J. Z. Hui, X. L. Zhou, and F. T. S. Chan, “A weld surface defect recognition method based on improved MobileNetV2 algorithm,” Mathematics, vol. 10, no. 19, pp. 18, 2022.

[10] D. H. Xu, C. B. Wen, and J. H. Liu, “Wind turbine blade surface inspection based on deep learning and UAV-taken images,” Journal of Renewable and Sustainable Energy, vol. 11, no. 5, pp. 13, 2019.

[11] Z. Y. Liu, Y. G. Song, R. N. Tang, G. F. Duan, and J. R. Tan, “Few-shot defect recognition of metal surfaces via attention-embedding and self-supervised learning,” Journal of Intelligent Manufacturing, vol. 34, no. 8, pp. 3507-3521, 2023.

[12] Z. Mentouri, A. Moussaoui, D. Boudjehem, and H. Doghmane, “Steel strip surface defect identification using multiresolution binarized image features,” Journal of Failure Analysis and Prevention, vol. 20, no. 6, pp. 1917-1927, 2020.

[13] C. Huang, M. H. Chen, and L. Wang, “Semi-supervised surface defect detection of wind turbine blades with YOLOv4,” Global Energy Interconnection-China, vol. 7, no. 3, pp. 284-292, 2024.

[14] F. Q. Deng, J. L. Luo, L. H. Fu, Y. L. Huang, J. L. Chen, N. N. Li, and et al., “DG2GAN: Improving defect recognition performance with generated defect image samples,” Scientific Reports, vol. 14, no. 1, pp. 17, 2024.

[15] V. Nath, C. Chattopadhyay, and K. A. Desai, “Development of image-based fast defect recognition and localization network (FDRLNet) for steel surfaces,” Manufacturing Letters, vol. 35, pp. 958-964, 2023.

[16] D. B. Wu, H. Wang, J. W. Liang, and S. To, “Adaptive acquisition and recognition system of blade surface defects during machining process,” Measurement, vol. 225, pp. 12, 2024.

[17] C. Zhang, T. Yang, and J. Yang, “Image recognition of wind turbine blade defects using attention-based MobileNetv1-YOLOv4 and transfer learning,” Sensors, vol. 22, no. 16, pp. 18, 2022.

[18] A. Carnero, C. Martín, and M. Díaz, “Portable motorized telescope system for wind turbine blades damage detection,” Engineering Reports, pp. 24, 2023.

[19] R. Zhang and C. B. Wen, “SOD-YOLO: A small target defect detection algorithm for wind turbine blades based on improved YOLOv5,” Advanced Theory and Simulations, vol. 5, no. 7, pp. 14, 2022.

[20] Y. L. Wang, Z. X. Zhou, X. J. Tan, Y. Q. Pan, J. Q. Yuan, Z. F. Qiu, and C. L. Liu, “Unveiling the potential of progressive training diffusion model for defect image generation and recognition in industrial processes,” Neurocomputing, vol. 592, pp. 16, 2024.

[21] J. Si and S. Kim, “V-DAFT: Visual technique for texture image defect recognition with denoising autoencoder and Fourier transform,” Signal, Image and Video Processing, pp. 14, 2024.

[22] B. Zhu, G.J. Xiao, Y. D. Zhang, and H. Gao, “Multi-classification recognition and quantitative characterization of surface defects in belt grinding based on YOLOv7,” Measurement, vol. 216, pp. 15, 2023.

[23] J. Zhang and D. Li, “Method of surface defect detection for agricultural machinery parts based on image recognition technology,” Soft Computing, pp.9, 2023.

[24] S. Mishra and C. Y. Tsai, “QSurfNet: A hybrid quantum convolutional neural network for surface defect recognition,” Quantum Information Processing, vol. 22, no. 5, pp. 28, 2023.

[25] Z. F. Li, L. Gao, Y. Gao, X. Y. Li, and H. Li, “Zero-shot surface defect recognition with class knowledge graph,” Advanced Engineering Informatics, vol. 54, pp. 13, 2022.

[26] M. Xiao, B. Yang, S. L. Wang, Z. Zhang, and Y. He, “Fine coordinate attention for surface defect detection,” Engineering Applications of Artificial Intelligence, vol. 123, pp. 12, 2023.

[27] Y. H. Liu, Y. Q. Zheng, Z. F. Shao, T. Wei, T. C. Cui, and R. Xu, “Defect detection of the surface of wind turbine blades combining attention mechanism,” Advanced Engineering Informatics, vol. 59, pp. 12, 2024.

[28] G. F. Duan, Y. G. Song, Z. Y. Liu, S. Q. Ling, and J. R. Tan, “Cross-domain few-shot defect recognition for metal surfaces,” Measurement Science and Technology, vol. 34, no. 1, pp. 9, 2023.

[29] B. Xia, S. Li, N. Li, H. Li, and X. G. Tian, “Surface defect recognition of solar panel based on percolation-based image processing and Serre standard model,” IEEE Access, vol. 11, pp. 55126-55138, 2023.

[30] Z. Y. Li, “Image recognition method of surface defects of prefabricated concrete members in prefabricated building,” International Journal of Materials & Product Technology, vol. 68, no. 1-2, pp. 22, 2024.

Design of Intelligent Extraction Method for Key Electronic Information Based on Neural Networks

Xiaoqin Chen, Xiaojun Cheng*
School of Intelligent Manufacturing, Chongqing Three Gorges Vocational College, Chongqing 404155, China

***Abstract*—With the rapid development of the Internet and other emerging media, how to find the needed information from massive electronic documents in time and accurately has become an urgent problem. A key electronic information extraction method based on neural network learning ideas has been proposed to solve the problems of time-consuming and difficult deep semantic feature mining in traditional text classification methods. Firstly, a weighted graph model was introduced to improve the TextRank keyword extraction algorithm, helping to capture complex data information and implicit semantics. The results indicate that the optimization method has the highest extraction accuracy (96.52%) on the CSL dataset, and its performance in feature extraction of information data is superior to other comparative models. Secondly, combining LSTM and self attention mechanism to achieve key feature extraction of contextual semantic information. The results indicate that this optimization method has relatively small training and testing errors in data classification, and tends to converge in the later stages of iteration. The accuracy of information extraction reached 94.37%, which is better than other comparative models. The keyword extraction integrity of the fusion model on the THUCNews dataset and Sogou News dataset were 86.2 and 84.1, respectively, with consistency of 96.3 and 94.7, and grammatical correctness of 92.1 and 92.2, respectively. The neural network-based extraction method proposed by the research institute can not only effectively improve the accuracy of information extraction, but also adapt to the changing data environment, and has great potential for application in the field of electronic information processing.**

Keywords*—Key electronic information; intelligent extraction; TextRank; LSTM; context

I. Introduction

The development of Internet information technology and the popularity of mobile intelligent devices make information interaction possible. According to the statistical report released by the Internet Network Information Center, the Internet penetration rate has reached more than 70% by 2022. With the help of mobile devices, people can access and produce different types of electronic information on media and social platforms, such as electronic documents, emails, social media data, etc. [1]. The generation of massive information data not only facilitates people's lives and work, but also invisibly increases the difficulty of information processing. For information receivers and users, it is necessary to filter, extract and analyze the massive data to achieve higher quality services for users. Electronic information often exists in various channels such as email, social media, and news reports in the form of text, voice, and images. How to accurately and quickly classify and extract key electronic information from a large and complex amount of network information has become one of the important tasks of natural language processing. The electronic information generated through online media has characteristics such as complexity. Classifying its content can not only solve the problem of information disorder to a large extent, but also play an important role in personalized recommendation, information retrieval, and other fields [2]. Traditional electronic information extraction techniques often rely on specific models, such as regular expression matching, which performs well in processing structured data, but often struggle to handle unstructured data such as images, audio, and natural language text.

Statistical analysis methods such as K-nearest neighbor algorithm, support vector machine, and decision tree have certain classification advantages, but they require manual design of rules and feature selection to achieve text classification, which consumes a lot of time and is difficult to mine deep semantic features of the text. Deep network models such as recurrent neural networks, convolutional neural networks, and pre trained models have advantages in information extraction, enabling large-scale dataset learning and automatic feature extraction and semantic association analysis. They can effectively improve the accuracy and efficiency of key electronic information extraction. Current research often relies on keyword extraction algorithms to achieve text data classification, which can reduce sparsity by shortening the length of text sequences. Keyword extraction algorithms are usually more intuitive and easy to implement, as they do not require complex network structure design and long-term model training. Compared to deep learning models that require a large amount of computing resources and data training, this method has lower hardware requirements, faster processing speed, and is easy to deploy [3]. Therefore, the research focuses on the learning approach of neural networks based on keyword extraction algorithms to extract key electronic information. At the same time, considering that keyword extraction algorithms often ignore the dependency relationships and complex semantic expressions of information keywords when processing data, as well as their weak ability to capture complex patterns and implicit semantics, this study proposes improvements to the algorithm. By introducing the TextRank algorithm based on weighted graph model and self attention mechanism, keyword extraction and semantic feature grasping can be achieved, thereby improving the classification performance of key electronic information extraction. TextRank is a widely used keyword extraction algorithm that constructs co-occurrence relationships between words based on a graph model. However, its performance in extracting keywords from Chinese text data is not ideal, as it cannot

*Corresponding Author

consider the positional information of words and the contextual information of the entire corpus. The weighted graph model takes into account the relationships between words and can solve the problem of sparse semantic feature distribution in electronic information text data. Traditional Long Short Term Memory (LSTM) neural networks have the same time series structure as text data and are widely used in natural language processing tasks. However, when it comes to feature extraction of text data, it cannot effectively combine contextual information to extract correct semantic features. The self attention mechanism can reduce the dependency relationship of sequence data and extract correct semantic features while combining contextual information. Therefore, research is being conducted on LSTM with improved self attention mechanism for text classification.

This study is mainly divided into following sections. Section II is Related Works, which summarizes the current research results in information extraction and other aspects. Section III is the research methodology, including key electronic information extraction based on optimized TextRank algorithm and information extraction based on LSTM network. Section IV is result analysis, which mainly tests the research methods. Section V is the discussion. Last Section VI is the conclusion, which summarizes the research results and shortcomings.

II. Related Work

Key information extraction is of great significance in improving information processing efficiency, supporting decision-making, and knowledge management. It can help researchers filter out valuable and relevant information from a large amount of information, providing more efficient, accurate, and intelligent information processing methods. Scholars from different fields have conducted extensive research and achieved some research results. Chi L et al. proposed a graph based and lightweight automatic key phrase extraction method for the automatic extraction of key information. Iterative sentences were used to sort words, generating a more accurate list of key phrases. The research method could effectively improve the extraction accuracy and significantly reduce the number of iterations [4]. Li T et al. designed an optimizer based on auto-regressive method to improve the accuracy of unsupervised key phrase extraction. They integrated any graph based unsupervised key phrase extraction model to enhance stability. This method could improve accuracy on different datasets by 50%, which was beneficial for unsupervised method key phrase extraction [5]. Singh Y et al. proposed an extraction method based on ternary block truncation encoding and binary bat algorithm to improve the efficiency of video summarization and key frame extraction. The method extracted and processed static images in frame form from the input video database, and measured the similarity measure between two consecutive frames. The research method had better extraction accuracy and F-metric value than traditional methods. It was more conducive to the effective and accurate extraction of video summaries and key frames [6]. Sachan M et al. proposed a method based on discourse and text layout features to extract geometric knowledge more effectively from multimedia textbooks. This method could more effectively extract knowledge, thereby improving the solution for geometric knowledge [7].

Many scholars have applied algorithms such as neural networks to information extraction. Zheng J et al. designed a semantic feature extraction method that integrated convolutional neural networks to address the low efficiency of traditional picking methods in identifying phase waves. The key parameters of the network were refined to ensure the extraction accuracy. It could effectively improve extraction efficiency and accuracy [8]. Sonntag D et al. proposed an automatic method for extracting text information in the complex integration process of medical data. This method could simplify the integration process of medical data and improve the accuracy of information extraction, which was beneficial for doctors in clinical diagnosis [9]. Nasar Z et al. proposed a method based on named entity recognition and relationship extraction to extract important information from text data on online platforms. Deep learning methods were used to construct joint models. The joint model based on deep learning had significant advantages. Named entity recognition and relationship extraction were beneficial for extracting key information from text data [10]. Zhang X et al. proposed a method based on multiple information extraction and support vector data description to optimize the process monitoring and fault diagnosis performance of key performance indicators, balancing local process information and mining hidden information. The maximum information coefficient algorithm selected key performance indicator information, extracted local information, and then extracted observed values, cumulative errors, and rate of change information. The verification results indicated that it could effectively improve the extraction accuracy, thereby promoting process monitoring and fault diagnosis of key performance indicators [11].

The above content indicates that effective key information extraction can significantly improve the efficiency of information processing. Some scholars have achieved intelligent information extraction by using fusion convolutional neural networks, entity recognition and relationship extraction, or data description and extraction perspectives. However, there is still room for improvement in the processing and classification of text data, making it difficult to ensure both data processing efficiency and extraction accuracy. The research aims to address the shortcomings of existing technologies in processing electronic information text data by integrating the weighted graph model's TextRank algorithm and self attention mechanism. The TextRank algorithm is a widely used keyword extraction algorithm that does not require complex network design and resource conditions to achieve accurate classification. Its improved idea can also effectively extract contextual semantic information features, especially when facing the problem of sparse distribution of semantic features in electronic information text data. It can effectively address the shortcomings of existing technologies in processing electronic information text data. By implementing the methods proposed in the research, it is expected to achieve more efficient and accurate information processing and classification effects, which will have a profound impact on promoting the development of information technology, supporting decision-making, and other related fields.

III. Key Electronic Information Extraction Based on Optimized TextRank and LSTM Networks

For the key electronic information extraction, TextRank is used to construct a graph model and optimize the TextRank algorithm. To better extract the semantic features of key electronic information by combining context, LSTM network and attention mechanism are introduced to achieve the extraction and classification of key electronic information.

A. *Key Electronic Information Extraction based on Optimized Textrank*

The TextRank algorithm is a widely used algorithm for extracting critical electronic information. It represents text data by constructing a graph model and establishes co-occurrence relationships between words in the graph. Specifically, this algorithm abstracts unstructured text into a graph structure. Each word is connected to a fixed number of adjacent words. Then, according to a certain formula, the weights of each node in the graph are recursively calculated. Finally, the nodes that rank higher during convergence can be used as keywords for electronic information. One of the advantages of this algorithm is its ability to simplify unstructured text data and extract useful information. It establishes connections between words by using co-occurrence relationships, thereby helping to identify the most important words or phrases in the text [12]. This is very useful for many natural language processing tasks, such as text classification, information retrieval, and summary generation. The TextRank algorithm determines the weights of words by calculating their correlation, as shown in Eq. (1).

$$TR(V_i)=(1-\alpha)+\alpha\times\sum_{V_j\in In(V_i)}\left(\frac{1}{|Out(V_j)|}\right)\times TR(V_j) \tag{1}$$

In Eq. (1), $In(V_i)$ and $Out(V_j)$ represent the predecessor and follower node sets of V_i and V_j, respectively. $|Out(V_j)|$ is the number of rear drive nodes for node V_j. $TR(V_i)$ is used to describe the weights of all nodes. α represents the damping coefficient, usually taken as 0.85. The TextRank algorithm does not rely on labeled data and does not require pre-training, but it has some limitations in keyword extraction. One limitation is that it only considers the connection between local words, without fully utilizing the global perspective to analyze the dependency characteristics between words [13]. To optimize TextRank, a key electronic information extraction model can be constructed by combining weighted graphs, which introduces the relationship between words.The text features of electronic information are considered. The word vector algorithm is used to train the word vectors of all electronic information data. The evaluation formula for TextRank algorithm is improved through the mutual information between words. This is the key electronic information extraction method based on optimized TextRank. The specific process is shown in Fig. 1.

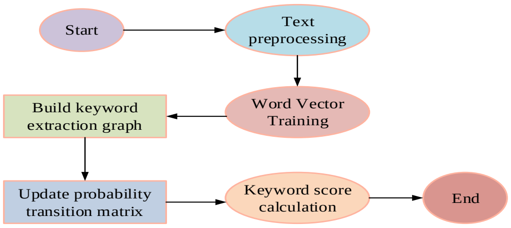

Start
Text preprocessing
Word Vector Training
Build keyword extraction graph
Update probability transition matrix
Keyword score calculation
End

Fig. 1. Key electronic information extraction process based on optimized TextRank.

From Fig. 1, the improved key electronic information extraction method based on optimized TextRank first preprocesses the phone information. Then, word vector training is used to train the preprocessed words into word vectors, which are introduced to calculate the mutual information between words, thereby constructing the probability transition matrix of the graph. Then, the sliding window size of the improved algorithm is set. The weight of all words is calculated by combining the transition probability matrix. The word set is sorted according to the descending value of the weight value. The top ranked words are used as the extracted key electronic information [14], [15]. Among them, language models are generally used to introduce character level combination information to train word vectors. Eq. (2) is the logarithmic natural function of the information.

$$\sum_{t=1}^{T}\sum_{c=C_t}\log p\left(w_c|w_t\right) \tag{2}$$

In Eq. (2), w_t represents the target word. w_c is the contextual word for w_t. C_t refers to the index set of contextual words. $p(w_c|w_t)$ is used to describe the probability of w_c occurring on the basis of setting w_t, as shown in Eq. (3).

$$p\left(w_c|w_t\right)=\frac{e^{s(w_t,w_c)}}{\sum_{j=1}^{W}e^{s(w_t,j)}} \tag{3}$$

In Eq. (3), $s(w_t,w_c)$ represents a rating function that maps words w_t and w_c to a rating, which can be combined with the rating value to determine the matching degree between the word and the context [16]. The mapping process of the rating function is generally calculated using the scalar product of the current word vector u_{wt} and the context word vector v_{wc}, as shown in Eq. (4).

$$s\left(w_t,w_c\right)=u_{wt}^{T}v_{wc} \tag{4}$$

After completing word vector training, the mutual information between nodes can be calculated using word vectors to update the probability transition matrix. The probability transition matrix is shown in Eq. (5).

$$W=\begin{bmatrix}W_{1,1} & W_{1,2} & \cdots & W_{1,n}\\ \vdots & \vdots & \ddots & \vdots\\ W_{n,1} & W_{n,2} & \cdots & W_{n,n}\end{bmatrix} \tag{5}$$

In Eq. (5), W_{ij} represents the weight values between nodes. n refers to the words in the corpus. The proportion of each contextual word in the target word should not be the same. The proportion of peripheral words with high relevance to the target word is relatively high. Therefore, the cosine similarity between nodes can be calculated using word vectors, which are used as edge weights. The similarity is shown in Eq. (6).

$$S_{V_1,V_2} = \frac{\sum_{i=1}^{d} V_{1,i} \times V_{2,i}}{\sqrt{\sum_{i=1}^{d}\left(V_{1,i}\right)^2} \times \sqrt{\sum_{i=1}^{d}\left(V_{2,i}\right)^2}} \tag{6}$$

In Eq. (3), S_{V_1,V_2} represents the cosine similarity between the word vectors V_1 and V_2. $V_{1,i}$ and $V_{2,i}$ represent the components of word vectors V_1 and V_2, respectively. The word vector calculation obtains cosine similarity to update the probability transition matrix. The updated probability transition matrix is shown in Eq. (7).

$$W = \begin{bmatrix} S_{V_1,V_1} & S_{V_1,V_2} & \cdots & S_{V_1,V_n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{V_n,V_1} & S_{V_n,V_2} & \cdots & S_{V_n,V_n} \end{bmatrix} \tag{7}$$

In Eq. (7), the range of cosine similarity values is [-1, 1]. To standardize the model, the probability transition matrix is normalized. The range of element values in the probability transition matrix is limited to [0,1]. The calculation process is shown in Eq. (8).

$$W_{i,j} = \frac{S_{V_i,V_j}}{\sum_{j=1}^{n} S_{V_i,V_j}} \tag{8}$$

The graph used for keyword extraction is first preprocessed for each word. Then each word is added to the graph in the form of a sliding window. The distance between words is used to indicate the coexistence between two words. When both words are included in a window, it indicates a connection between the two words. The constructed probability transition matrix is used as the weight between nodes. The key electronic information extraction based on optimized TextRank is shown in Fig. 2.

From Fig. 2, the key electronic information extraction graph based on optimized TextRank has a window size of 3. All words are connected to adjacent words before and after [17]. By constructing a weighted graph, the scoring formula for extracting key electronic information can be obtained. The expression is shown in Eq. (9).

$$TR\left(w_i\right) = \left(1-\alpha\right) + \alpha \times \sum_{w_j \in In\left(w_i\right)} \left(W_{i,j} \times TR\left(w_i\right)\right) \tag{9}$$

In Eq. (9), $In\left(w_i\right)$ represents the predecessor node sets of node w_i. $TR\left(w_i\right)$ represents the weights of all nodes. $W_{i,j}$ represents the transition probability between nodes.

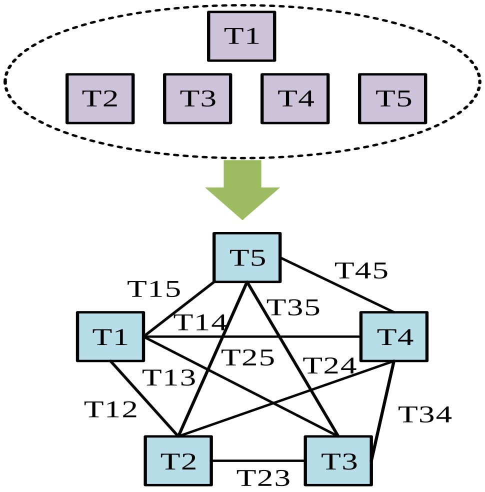


Fig. 2. Key electronic information extraction based on optimized TextRank.

B. *Feature Extraction based on LSTM Network and Attention Mechanism*

The semantic complexity of key electronic information can affect the extraction performance of TextRank. To better combine context to extract the semantic features of key electronic information, LSTM and attention mechanism are introduced to achieve the extraction and classification of key electronic information. LSTM is a special type of recurrent neural network (RNN) used to solve the gradient vanishing and exploding faced by traditional RNNs. The LSTM network controls the inflow and outflow of information by introducing a structure called a "gate". This network mainly includes input gates, forget gate, output gate (OG), and cell state. The input gate determines which information can enter the cellular state. The forget gate determines which information needs to be forgotten from the cellular state. The OG determines the output of information in the cell state after activation. The key to LSTM lies in their ability to "remember" and "forget" information. Through the forget gate, LSTM can selectively remove information from the cell state, avoiding irrelevant information from interfering with subsequent calculations. The input gate can control which newly inputted information can be added to the cell state. This memory ability to retain long-term dependencies makes LSTM perform well in processing sequence data. Attention mechanism is a technique that mimics human attention behavior. This method is used to assign different weights and attention levels to different parts of the input in deep learning models. In traditional neural networks, all input information is processed simultaneously. Attention mechanisms can enable the model to selectively focus on certain parts of the input, thereby improving the performance and effectiveness. The core idea of attention mechanism is to determine the importance of each input based on its input. Then, weighted processing is performed based on these importance levels. In deep learning models, attention mechanisms can be used at different levels and time steps. The key electronic information extraction based on LSTM network and attention mechanism is shown in Fig. 3.

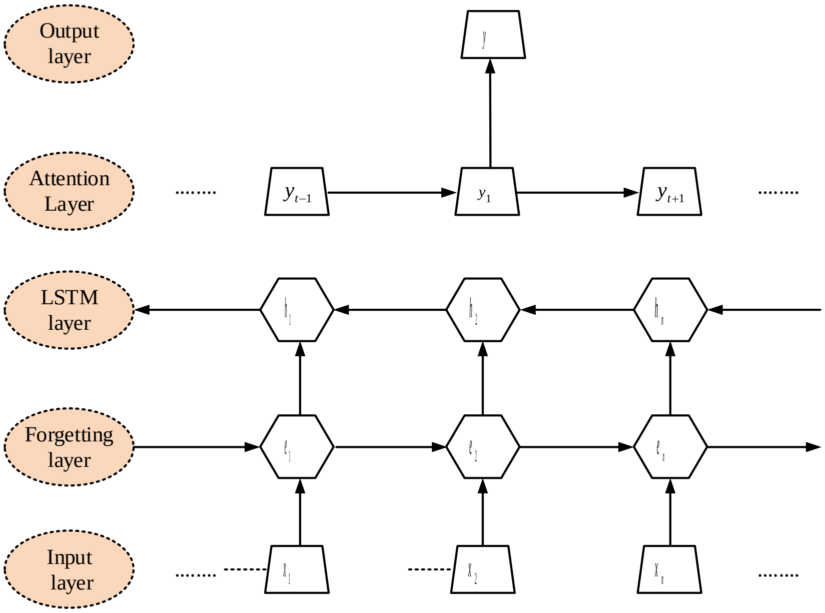


Fig. 3. Key electronic information extraction based on LSTM network and attention mechanism.

From Fig. 3, the extraction model has five parts. Time series data is used as the input layer to extract word vectors using LSTM. The self attention mechanism is applied to calculate the weight of the text features output by LSTM layer, so that it focuses on the main content of the overall text. The output layer uses sequence level feature vectors to achieve text classification. The LSTM network is used for feature extraction of input text. Based on this, the semantic features of each word are described. Compared to traditional RNNs, LSTM has added three gating mechanisms, namely, forget gate, input gate, and OG [18]. The forgetting gate determines how many past states one should maintain in their current state. The input gate determines how much input information has been retained within the current time period. The OG determines the size of the current state output. The unit structure of the LSTM network is displayed in Fig. 4.

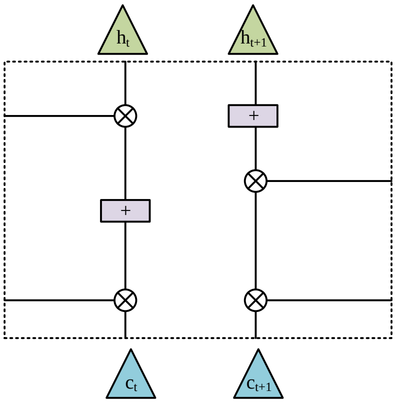


Fig. 4. The unit structure of LSTM networks.

The forgetting gate can control the proportion of information to be forgotten. The input is the output h_{t-1} of the previous moment and the text information input x_t of the current moment. The calculation process of the forget gate is shown in Eq. (10).

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{10}$$

In Eq. (10), σ represents the sigmoid function. W_f and U_f are the weight matrices of the forget gate. b_f represents the bias term of the forget gate. The input gate is mainly the information stored in the state unit, which can calculate the memory information of the current unit [19]. The unit state at the current moment is calculated by the sum for the product of the forgetting gate input and the previous moment state, as well as the input. The calculation process is shown in Eq. (11).

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{11}$$

In Eq. (11), c_t and c_{t-1} represent the current and previous unit states, respectively. The OG function can determine the current hidden state and save the previous input information in the hidden state. The input of the current time and the hidden state for the previous time are respectively inputted into the sigmoid function. The current unit state is inputted into the tanh function. The output result of the sigmoid function is multiplied by the tanh function to obtain the hidden state in the current time. Afterwards, the text sequence is processed using the LSTM network to obtain the implicit state h_t at each moment, which characterizes the semantic features of each word in the text and achieves text classification. In the process of extracting electronic news information, it is also necessary to combine contextual information. The semantic feature extraction based on a single word has one sidedness, which has a certain impact on the model's judgment and thus reduces the classification accuracy [20]. Therefore, the study intends to introduce attention mechanisms into the text. The word features of LSTM are analyzed to achieve text feature extraction that integrates context. This study draws on the self attention mechanism of the Transformer model. Semantic feature extraction that integrates with context can achieve automatic mining of target text without relying on source data.

IV. Analysis of Intelligent Extraction Results for Key Electronic Information

In order to improve the intelligent extraction of key electronic information, an improved TextRank algorithm and LSTM-SAttention model were designed to enhance data classification performance and feature extraction accuracy. In the results section, the study mainly analyzed from two aspects: technical performance evaluation and application result verification.

A. Test Data Source and Experimental Environment Parameter Design

The algorithm was tested using the Chinese Scientific Literature (CSL) keyword extraction dataset and the NLPCC2017 keyword extraction dataset publicly available at the 2017 Academic Annual Meeting of the Natural Language Processing Professional Committee of the Chinese Computer Society. The CSL dataset mainly involves key electronic information content in the computer field, while the NLPCC2017 dataset mainly involves news articles and their keywords. De duplication and filtering were performed on these two datasets to remove duplicate, formatting errors, missing key fields, or incomplete content, resulting in 7562 CSL data and 7635 NLPCC2017 data. In the performance test results section, the study introduces LSTM network and attention mechanism into optimizing the TextRank algorithm to construct a key electronic information intelligent overall model. The THUCNews dataset published by Tsinghua University is taken as the research object, and ten categories of Chinese information (including social, technological, educational, etc.) are selected, each containing 20000 pieces of information, forming a balanced dataset. In addition, the "Sogou News" dataset released by Sogou Lab also contains ten categories of news, which are divided into training and validation sets in a 7:3 ratio. The parameter settings for the experimental environment are shown in Table I.

B. Experimental Test Results

To test the effectiveness of sliding windows in extracting key electronic information, the study first analyzes the keyword extraction data set and the extraction algorithm based on optimized TextRank. Firstly, the test results of two datasets at different window sizes are shown in Fig. 5.

TABLE I. System Parameter

Number	Testing environment	Parameter
(1)	Processor	Intel(R) Core (TM)i5-7300HQ CPU@2.50GHz
(2)	Operating system	Windows 10
(3)	Programming Language	C++
(4)	Memory	32GB
(5)	GPU	GTX 1050Ti
(6)	Programming Language	Python 3.7.3

From Fig. 5, when only considering the impact of association distance on keyword extraction results, a window size of 2 had the best extraction effect on keywords. To fully utilize the global text information and fully consider the order characteristics of the text, the optimized TextRank algorithm is used to represent words in vector form. The similarity between words is calculated. Then it is introduced as a weight into the model. Fig. 6 displays the specific results.

From Fig. 6(a-b), the keyword extraction performance of the research method was superior to other methods. When the window was set to 2, the F1 values of TextRank before optimization, TextRank after optimization, and traditional methods all reached their highest. The F1 values on the CSL data set were 0.59, 0.51, and 0.32. On the NLPCC2017 data set, they were 0.56, 0.47, and 0.17. From Fig. 6(c-d), the extraction accuracy of the three algorithms increased with time. Their accuracy on the CSL data set exceeded the NLPCC2017 data set. The optimized TextRank algorithm is tested on the CSL data set. When the extraction time was 10s, the extraction accuracy stabilized at 92%, while the pre-optimized algorithm stabilized at over 70% after 30s. The traditional method stabilized at over 55% after 30s of extraction.

Next, the performance of the proposed text extraction algorithm that integrates keyword extraction and attention mechanism is validated. The proposed algorithm is compared with other deep learning based text extraction algorithms. The experiment aims to construct an LSTM-SAttention feature extraction network. It is applied to two datasets to test the effectiveness. The results are shown in Fig. 7.

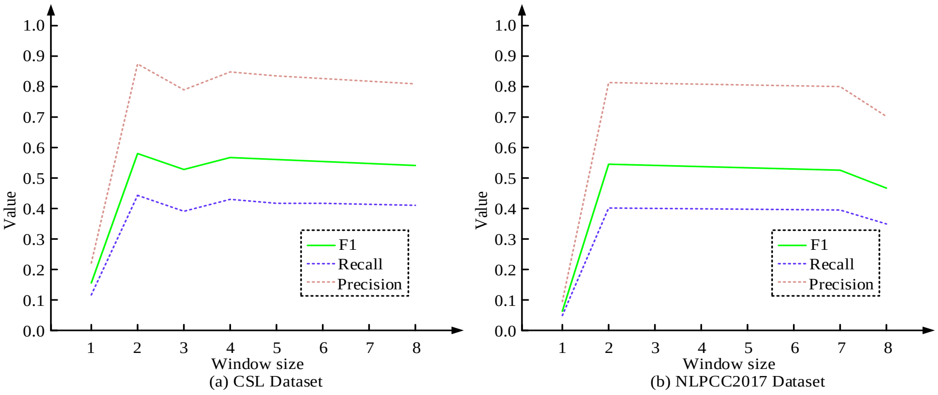
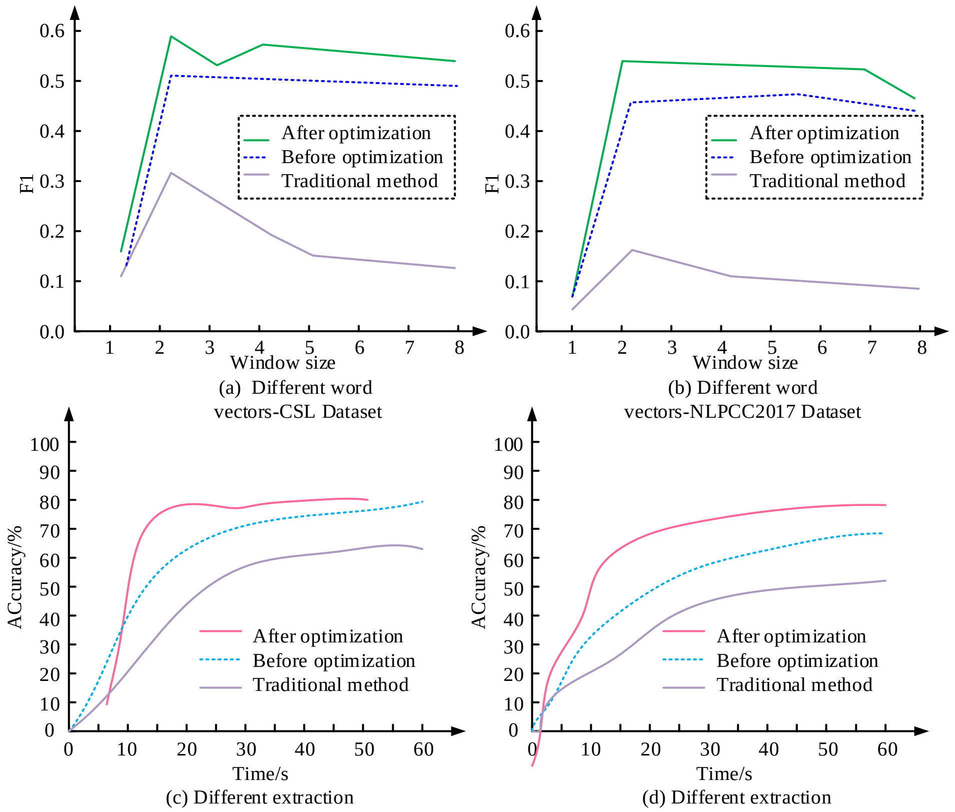


Fig. 5. Results of different datasets under different windows.



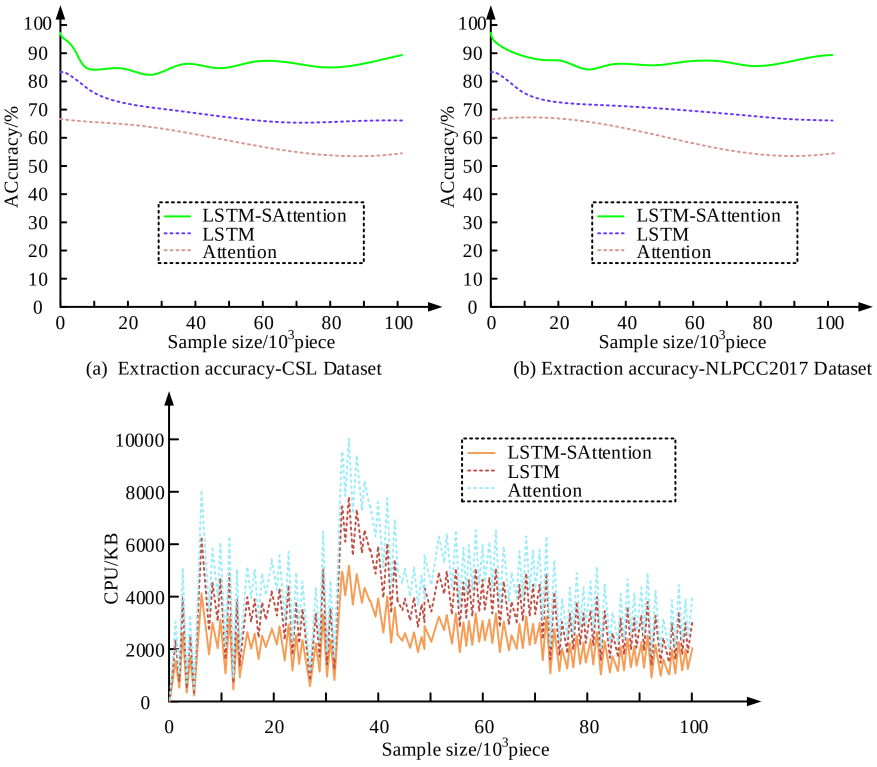
(a) Different word vectors-CSL Dataset

(b) Different word vectors-NLPCC2017 Dataset

(c) Different extraction durations-CSL Dataset

(d) Different extraction durations-NLPCC2017 Dataset

Fig. 6. Keyword extraction performance under different word vectors and extraction times.



(a) Extraction accuracy-CSL Dataset

(b) Extraction accuracy-NLPCC2017 Dataset

(c) CPU occupancy of three methods

Fig. 7. Extraction accuracy and CPU usage.

From Fig. 7(a-b), as the sample size increased, all three algorithms showed a trend of decreasing extraction accuracy. However, the research method stabilized faster showed high accuracy and stability on both datasets. Among them, the highest extraction accuracy of LSTM-SAttention on the CSL data set was 96.52%. The highest extraction accuracy on the NLPCC2017 data set was 71.21%. Fig. 7(c) shows the result of memory operation. From Fig. 7(c), as the sample size increased, the CPU occupancy of all three algorithms gradually increased. However, the occupancy rate of research methods was the lowest. The increase rate was also the slowest. Overall, the key electronic information extraction method combining attention mechanism and LSTM network performs the best. Subsequently, the application effect of the LSTM-SAttention model proposed in the study was analyzed and compared with the Enhanced Support Vector Machine (E-SVM) algorithm, Graph Convolutional Networks (GCN-FCN), Large Language Model Meta AI, and Multi level Semantic Alignment Image Text Matching Algorithm (MLS-ITM). The results are shown in Table II.

The results in Table II indicate that the LSTM-SAttention model outperforms other algorithms in terms of performance evaluation on both databases. Its maximum accuracy feature extraction results on the CSL dataset differ from those of the E-SVM algorithm by over 10%, while the difference between the LSTM-SAttention model and the MLS-ITM algorithm is within 5%. The accuracy, recall, and F-value of the LSTM-SAttention model all exceed 90, followed by the well performing MLS-ITM algorithm and GCN-FCN algorithm. The Meta AI model's values do not exceed 85. On the NLPCC2017 dataset, the LSTM-SAttention model (90.47)>MLS-ITM (87.12)>GCN-FCN (83.25)>Meta AI (80.07)>E-SVM (70.32). The above results indicate that the LSTM-SAttention model can achieve good feature extraction.

C. *Actual Inspection Results*

Using Chinese news text classification as the test object, the proposed TextRank optimization extraction model and the Vocabulary Semantic Map Attention Mechanism (VSA), SENet and Convolutional Neural Network fusion method (Squeeze and Excitation Networks Convolutional Neural Network, SENet CNN), as well as the Convolutional Neural Network Bi directional Long Short Term Memory (TC Ablstm) fusion parallel neural network model were analyzed for information extraction results. Fig. 8 shows the training and testing errors of data processing.

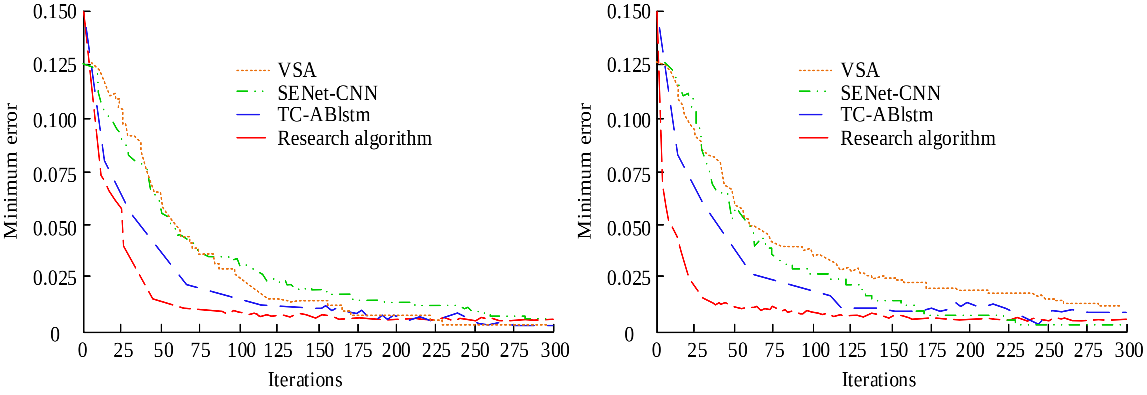
From the testing process in Fig. 8(a), it can be seen that the testing errors of the four algorithms all show a decreasing trend with the increase of iteration times. Before the iteration times are less than 150, the average testing error results from large to small are: VSA algorithm>SENet CNN algorithm>TC Ablstm algorithm>the proposed algorithm. As the number of iterations gradually increases, the testing errors between algorithms are all less than 0.025, and the error difference between algorithms does not exceed 5%. The overall variation of the testing error of the TextRank optimization extraction model proposed in the study is relatively small, and tends to converge in the later stages of iteration. From the testing process in Fig. 8(b), it can be seen that the error comparison between different algorithms is relatively large, with the order of error from small to large: the proposed algorithm>TC Ablstm algorithm>SENet CNN algorithm>VSA, with an average testing error of 2.36%>5.64%>13.26%>33.64%. At the same time, the algorithm proposed in the study showed a greater slope of decrease in the training error curve when the number of iterations was less than 75, and in the later stage, the error curve became more stable, further improving the classification accuracy. Subsequently, the information extraction results of the proposed fusion algorithm were analyzed, and the results are shown in Fig. 9.

The results in Fig. 9 indicate that the PR curves of the TextRank optimization extraction model and the TC Ablstm model are closer to the bottom right corner. The accuracy of VSA, SENet CNN, and TC Ablstm algorithms are 78.24%, 85.69%, and 86.78%, respectively, while the proposed TextRank optimization hybrid model achieves an accuracy of 94.37% in information extraction. Chinese news text is classified as the test object. The ROUGE Scores, integrity, consistency, and grammar correctness of the intelligent extraction model for key electronic information are analyzed. Among them, ROUGE Scores are mainly used to evaluate the quality of automatically generated abstracts or translations. Integrity measures whether the key information extraction covers all important parts of the document. Consistency measures the consistency between the extracted key information and the manually marked key information. Grammar correctness assesses the generated grammar level of the text, ensuring good readability and no grammar errors. The specific results are shown in Fig. 10.

Fig. 10(a) and 10(b) show the ROUGE Scores, integrity, consistency, and grammar correctness scores of the model for keyword extraction and document summarization generation on the THUCNews data set and Sogou news data set, respectively. From Fig. 10, the model had high ROUGE Scores, integrity, consistency, and grammar correctness scores on both datasets. When the sample size was 6000, the ROUGE Scores were 89.6 and 88.2, with integrity, of 86.2 and 84.1. The consistency was 96.3 and 94.7, respectively. The correctness of the method was 92.1 and 92.2, respectively. The model constructed by the research method can accurately and completely extract key electronic information, and ensure the accuracy of grammar.

TABLE II. COMPARISON OF TRAINING EXPERIMENT RESULTS OF DIFFERENT MODELS

Contrast model	**CSL database**			**NLPCC2017 database**		
	Accuracy	Recall	F value	Accuracy	Recall	F value
E-SVM	77.47	80.52	79.21	70.32	70.23	70.44
Meta AI	82.25	81.49	81.71	80.07	79.32	80.11
GCN-FCN	84.36	85.32	84.38	83.25	82.43	83.48
MLS-ITM	86.43	87.25	88.12	87.12	89.24	88.75
LSTM-SAttentionModel	91.22	90.32	92.23	90.47	91.16	90.52



(a) Test process

(a) Training process

Fig. 8. Comparison of error results during testing and training.

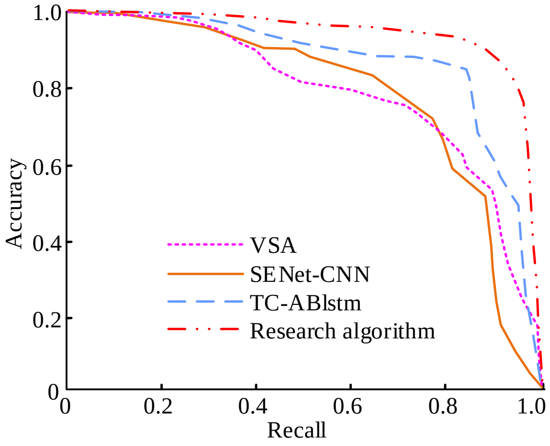
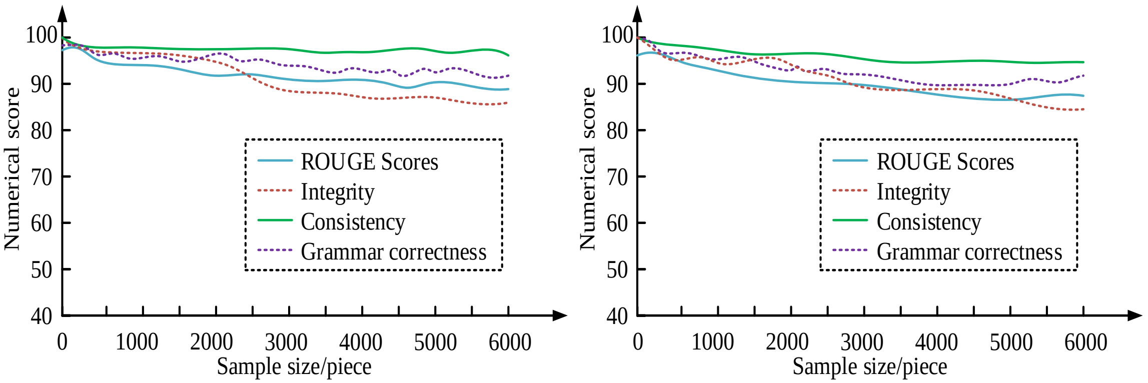


Fig. 9. Accuracy results of information extraction using different algorithms.



(a) THUCNews Dataset

(b) Sogou News Dataset

Fig. 10. Test results of intelligent extraction model for key electronic information.

V. Discussion

Performance analysis and case testing were conducted on the proposed electronic information extraction algorithm, and compared with different algorithms. The results showed that when the window size was 2, the optimized TextRank achieved F1 values of 0.59 and 0.56 for keyword extraction on the CSL dataset and NLPCC2017 dataset, respectively. When the extraction time was 10 seconds, the extraction accuracy remained stable at over 92%. The LSTM-SAttention model showed higher accuracy and stability on both datasets. The reason is that the LSTM attention model mainly combines context to extract key electronic information, which can grasp the correlation between different information, so information extraction has high credibility. Introducing a self attention mechanism enables the model to focus more on the information rich parts when processing sequential data, thereby making feature extraction more accurate and efficient. Other algorithms, such as VSA, SENet CNN, and TC Ablsm, have also been improved, but there is still a gap in error convergence speed and final stability. The introduction of self attention mechanism can help the model capture long-range dependencies in text data and provide more refined feature representations for the model. The GCN-FCN model is difficult to consider the time series characteristics of text data, and the E-SVM model has significant dependence on structure and cannot simultaneously consider the spatiotemporal differences of text information. The Meta AI language model performs well in data processing, but it heavily relies on the level of data training, while the MLS-ITM model struggles to meet different information extraction needs, such as keyword localization and relationship recognition, with good integrity and consistency in information extraction on both datasets. When the sample size is 6000, the ROUGE score, completeness, consistency, and grammatical correctness score of the TextRank optimized hybrid model exceed 88, 84, and 94 points, respectively, which can better ensure the semantic integrity and consistency of the text. Compared with literature [8] fusion convolutional neural network, the TextRank optimized hybrid model can effectively ensure the integrity of semantic information. Compared with the automatic extraction of text information by literature [9], the TextRank optimized hybrid model has better information expansion ability. Compared with the information extraction method proposed by literature [11], the TextRank optimized hybrid model can not only extract key information, but also better grasp the correlation between sentence information before and after.

The proposed model has good information extraction performance and efficiency, but there are still issues that need improvement, such as the introduction of self attention mechanism, which will increase the time and cost to a certain extent. Therefore, further optimization of neural network parameters is needed in the future. At the same time, the proposed method will be combined with other intelligent algorithms to improve cross modal information processing and feature fusion, and enhance the good adaptability of data characteristics. It is expected to provide reference for the field of electronic information processing and the improvement of information extraction services.

VI. Conclusion

With the explosive growth of electronic information, how to automatically extract key information from a large amount of text data has become increasingly crucial for information processing and analysis. In this context, the intelligent extraction method of key electronic information based on neural networks has become a hot and challenging research topic. For the key electronic information extraction, the TextRank algorithm is used to construct a key information extraction model and optimize it. A key electronic information extraction model based on optimized TextRank is obtained. Combined with attention mechanism and LSTM network, it facilitates context connection and improves the accuracy of extracting key electronic information. From the research results, with the accumulation of extraction time, the extraction accuracy of the optimized TextRank algorithm increase. The accuracy on the CSL data set exceeded on the NLPCC2017 data set, with a stable extraction accuracy of over 92%. The research method showed high accuracy and stability on both datasets. The keyword extraction and document summary generation effects of the research method were tested. The ROUGE Scores on the THUCNews data set and Sogou news data set were 89.6 and 88.2, respectively. The integrity was 86.2 and 84.1, the consistency was 96.3 and 94.7, and the grammatical correctness was 92.1 and 92.2, respectively. This indicates that the research method has a good effect on extracting key electronic information.

Acknowledgment

The research is supported by: The 2023 Chongqing University Research Project "Construction and Key Technology Application Research of Orange Honey Quality Traceability System Based on Blockchain Service Network (BSN)" (KJZD-K202303502);

The 2022 Chongqing Vocational Education Teaching Reform Research Project "Three dimensional Integration, Five party Collaboration: Exploration and Practice of the Long Term" Green Manufacturing "Talent Integration Training Model" (GZ223113).

References

[1] Zhang M, Bo X U, Xiaoyun L I, Dong FU, Liu J, Baojian WU, Qiu K. Artificial Neural Network-Based QoT Estimation for Lightpath Provisioning in Optical Networks. IEICE Transactions on Communications, 2019, E102.B(11):2104- 2112.

[2] Gao L, Li X, Liu D, Wang L， Yu Z.A Bidirectional Deep Neural Network for Accurate Silicon Color Design. Advanced Materials, 2019, 31(51):1905467.1-1905467.7.

[3] Wang K, Liu M. A feature, ptimized Faster regional convolutional neural network for complex background objects detection. IET Image Processing, 2021,15(2):378-392.

[4] Chi L, Hu L. ISKE: An unsupervised automatic keyphrase extraction approach using the iterated sentences based on graph method. Knowledge-Based Systems, 2021, 223(6):107014.1-107014.12.

[5] Li T, Hu L, Li H, Sun C, Li S, Chi L. Towards unsupervised keyphrase extraction via an autoregressive approach. Knowledge- based systems, 2023,274(Aug.15): 1.1-1.10.

[6] Singh Y, Kaur L. Effective key-frame extraction approach using TSTBTC–BBA. IET Image Processing, 2020, 14(4):638- 647.

[7] Sachan M, Dubey A, Hovy E H, Mitchell TM, Xing EP. Discourse in Multimedia: A Case Study in Extracting Geometry Knowledge from Textbooks. Computational Linguistics, 2019, 45(8):1-35.

[8] Zheng J, Shen S, Jiang T, Zhu W. Deep neural networks design and analysis for automatic phase pickers from three-component microseismic recordings. Geophysical Journal International, 2020, 220(1):323-334.

[9] Sonntag D, Profitlich H J. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. Artificial intelligence in medicine, 2019, 93(JAN.):13-28.

[10] Nasar Z, Jaffry S W, Malik M K. Named Entity Recognition and Relation Extraction: State-of-the-Art. ACM computing surveys, 2022,54(1):20.1-20.39.

[11] Zhang X, Ma L, Peng K. A novel key performance indicator oriented process monitoring method based on multiple information extraction and support vector data description. The Canadian Journal of Chemical Engineering, 2022,100(5):1013- 1025.

[12] Hayat S, Kun S, Shahzad S, Suwansrikham P, Mateen M, Yu Y. Entropy information-based heterogeneous deep selective fused features using deep convolutional neural network for sketch recognition. IET Computer Vision, 2021,15(3):165-180.

[13] Hamouda M, Ettabaa K S, Bouhlel M S. Smart Feature Extraction and Classification of Hyperspectral Images based on Convolutional Neural Networks. IET Image Processing, 2020, 14(10):1999-2005.

[14] Chunhao D, Peng C, Kang O. Enhanced high-order information extraction for multiphase batch process fault monitoring. The Canadian Journal of Chemical Engineering, 2020, 98(10):2187-2204.

[15] A L Y, B W G, D L J C. Keyword guessing attacks on a public key encryption with keyword search scheme without random oracle and its improvement. Information Sciences, 2019, 479:270-276.

[16] Zhu E, Sheng Q, Yang H, Li J. A unified framework of medical information annotation and extraction for Chinese clinical text. Artificial intelligence in medicine, 2023,142(Aug.):1.1-1.12.

[17] Wong R L, Sagar M, Hoffman J, Huang C, Gore JL. Clinical accuracy of information extracted from prostate needle biopsy pathology reports using natural language processing. Journal of Clinical Oncology, 2021, 39(15_suppl):1557- 1557.

[18] Xingchen Z, Yan G, Xian-Min M, Cao Y, Chen X, Chen Z, Du W, Fu L, Luo Z. Extracting photometric redshift from galaxy flux and image data using neural networks in the CSST survey. Monthly Notices of the Royal Astronomical Society, 2022,512(3):4593- 4603.

[19] Luo N, Yu H, You Z, Li Y, Zhou T, Jiao Y, Han N, Liu C, Jiang Z, Qiao S. Fuzzy logic and neural network-based risk assessment model for import and export enterprises: A review. Journal of Data Science and Intelligent Systems, 2023, 1(1): 2-11.

[20] Xiong Y, Chen Y, Chen C, Wei X， Wang P. An Odor Recognition Algorithm of Electronic Noses Based on Convolutional Spiking Neural Network for Spoiled Food Identification. Journal of The Electrochemical Society, 2021, 168(7):1-9.

Optimized Blockchain-Based Deep Learning Model for Cloud Intrusion Detection

Sultan Alasmari
Department of Information System-College of Computer and Information Sciences,
Majmaah University, Majmaah 11952, Saudi Arabia
College of Technology and Business, Riyadh Elm University, King Fahad Road, Riyadh 12734, Saudi Arabia

***Abstract*—Cyberattacks are becoming increasingly complex and subtle. In many different types of networks, intrusion detection systems, or IDSs, are frequently employed to help in the prompt detection of intrusions. Blockchain technology has gained a lot of attention recently as a means of sharing data without a reliable third party. Specifically, it is impossible to change data stored in one block without changing all the following blocks. Create a deep learning (DL) method based on blockchain technology and hybrid optimization to improve the IDS's prediction accuracy. The UNSW-NB15 dataset is gathered via the Kaggle platform and utilized for Python system training. Principal component analysis (PCA) is used in the preprocessing to eliminate errors and duplication. Next, employ association rule learning (ARL) and information gain (IG) approaches to retrieve pertinent characteristics. The greatest features are the ones that improve detection performance through hybrid seahorse and bat optimization (HSHBA) selection. Lastly, create an efficient intrusion detection system by designing Blockchain-based Ensemble DL (BEDL) models, with convolutional neural networks (CNNs), restricted Boltzmann machines (RBM), and generative adversarial networks (GAN). The constructed model's experimental results are verified using pre-existing classifiers, yielding an improved accuracy of 99.12% and precision of 99%.**

Keywords*—*Intrusion detection system; blockchain; deep learning; hybrid optimization; cloud computing; feature selection

I. INTRODUCTION

As the Internet has developed, an innovative technology known as the Internet of Things (IoT) has surfaced and been increasingly integrated into our daily lives. The IoT is directly empowering people and society through applications including healthcare, supply chain management, and RFID-based identity management systems [1]. By merging cloud computing and machine learning (ML), the base technology is starting to show promise for data analysis and modeling. Growth is being driven by the progress of IoT-based development across multiple industries. On the other hand, most IoT applications rely on centralized processing and storage architecture [2, 3]. There are number of security or privacy flaws in the centralized storage model. There are limitations in the underlying working paradigm that will make it difficult to expand IoT-based systems shortly. Decentralized storage models are required to solve these problems. Blockchain technology is one of the newest decentralized architectures [4, 5].

A collection of cryptographic entities can store and manage an assortment of time-stamped, unchanging data records in blocks using a technique called blockchain. The blocks are connected by the cryptographic hashes of the preceding block [6]. The timestamp, hash of the preceding block, and transaction comprise each block. Consequently, any modifications to the transactions need to be applied consistently through the consensus process across all of the blocks that comprise the blockchain [7]. This proves that the blockchain is the best data storage framework and ensures its immutability. Blockchain systems have been used for bitcoins, handling operations, financial services, healthcare, digital ID leadership, and many more uses. Two non-financial blockchain systems are Ethereum and Hyperledger [8, 9].

The proliferation of internet usage has facilitated data exchange and storage. Nevertheless, these data's susceptibility to hackers is greatly increased [10]. Numerous studies suggested the use of firewalls, data encryption, and user authentication to prevent unauthorized users from accessing stored data; yet, malevolent actors continue to find ways to circumvent these security measures and obtain illegal data access [11]. IDS has been proposed by further study to detect hostile intrusions in computer networks and internet-connected devices. Although intrusion detection systems have shown to be effective in detecting malicious activity, their limited perspective makes it difficult to spot coordinated or widespread cyberattacks [12, 13]. Because of this vantage point, certain attacks may go unnoticed or may not be discovered quickly enough. IDSs must swap attack features to quickly identify new assaults because certain attacks have managed to evade detection [14]. Thus, by combining the activities of participating IDS nodes, more harmful behaviors can be halted if IDS nodes communicate this threat information.

Protecting oneself from various kinds of attacks is more crucial than ever in the modern world, where data-centric research demands accurate DL algorithms [15]. Current studies have brought attention to ML systems' susceptibility to intrusion detection, whereby undetectable changes in the input data lead to inaccurate predictions in the output. A range of hostile attacks occur in real-time settings, and workable defense strategies are suggested to fend them off [16, 17]. In recent years, convolutional neural networks have been increasingly prevalent in the DL area as they have demonstrated remarkable performance in a wide range of application domains. High efficiency and long-term reliability are achieved using DL, which teaches neural network layers to see data as a hierarchical structure of principles [18]. When the amount of data increases, DL performs better than traditional ML. Blockchain-enabled DL-based intrusion detection algorithms have grown in popularity and have been used for a wide range of applications in recent years [19, 20]. To improve IDS prediction and improve feature selection, the proposed IDS framework employed hybrid optimization and ensemble DL approaches. The key contribution of the developed model is followed as,

- UNSW-NB15 datasets are collected from the Kaggle website and trained in the system.
- Preprocessing steps include data cleaning, normalization, and dimensionality reduction to prepare the data for DL models.
- Then extract relevant features in the feature extraction phase.
- Moreover, select the most important features using the HSHBA model to enhance the detection results.
- Finally, intrusions are detected using the ensemble DL technique, and the output is predicted based on the majority voting of each classifier.
- To maintain the honesty of the blockchain, cryptographic hash functions can be employed.
- The experimental outcome is validated with existing techniques in terms of accuracy, precision, false rate, and so on to prove the efficiency.

Section II reviews the related study, Section III describes the problem definition, Section IV elaborates on the proposed model, and Section V depicts the experimentation analysis followed by the conclusion in Section VI.

II. Related Work

A deep blockchain framework (DBF) is proposed by Osama et al. [21] to provide privacy-based blockchain technology with intelligent agreements and security-based decentralized intrusion detection in Internet of Things networks. A DL algorithm called Bidirectional Long Short-Term Memory (BiLSTM) uses the IDS to handle sequential network information. Distributed intrusion detection systems are given privacy through the use of the Ethereum library in the development of privacy-based ledgers and smart contract techniques. Both users and cloud providers may find the framework useful as a decision-support tool.

A unique distributed intrusion detection system (IDS) that uses fog computing to identify DDoS attacks towards a mining facility in an IoT network supported by blockchain is proposed by Randhir et al. [22]. Random Forest (RF) and an improved gradient tree boosting system (XGBoost) are trained on dispersed fog nodes to assess performance. The BoT-IoT dataset, which comprises the majority of current attacks discovered in blockchain-enabled IoT networks, is used to evaluate the efficacy of the suggested methodology. In general, RF requires less time for testing and training on widespread fog nodes than XGBoost.

To improve IIoT security, Romany et al. [23] presented the efficient Blockchain-Assisted Cluster-based IDS method known as BAC-IDS. Clustering IoT devices is the goal of the BAC-IDS concept, which intends to enable blockchain-based safe data transfer and discover intrusions. The BAC-IDS method uses a clustering approach based on Harris Hawks Optimization (HHO) to select Cluster Heads (CH) and build clusters based on their preferences. The NSL-KDD2015 dataset shows that the BAC-IDS technique has a lower error rate (0.03) based on experimental results.

In a smart city, Internet of Things security is essential. IoT security is a serious concern because of the many objectives and significant drawbacks that impede the quick adoption of these smart gadgets. Using lightweight information technology, Erukala et al. [24] describe a permission-based blockchain network for safeguarding the key pairs of connected devices. Using the combined ML technique, a collaborative detection tool is utilized to identify DDoS attacks on Internet of Things devices. Next, include a blockchain system that securely distributes alarm warnings to every IoT network node with sufficiently secure identification.

To build, implement, and evaluate an intrusion detection system, Chao Liang et al. [25] presented a hybrid placement method based on a multi-agent framework, blockchain, and DL procedures. The modules that make up the system are reaction, analysis, data administration, and data collecting. The outcomes show how effective DL algorithms are in identifying transport layer threats. According to the experiment, DL algorithms can be used to detect intrusions in Internet of Things networks.

A blockchain-based DL and ML system was developed by Shraddha et al. [26] to enhance IDS performance. First, use a group of classifiers, like Random Forest, Convolutional Neural Network, and XGBoost, to identify anomaly assaults. Next, utilizing the blockchain system, identified assaults are transformed into signatures and transferred further throughout the network. In this step, information is stored and secured at a higher level using the cryptosystem that is a part of the blockchain. The suggested IDS system's experimental results show increased accuracy and greater detection performance.

To efficiently identify Advanced Persistent Threats (APT) attack characteristics on the terminals, Lampis [27] suggests an Intrusion Detection and Prevention System (BIDPS) supported by Blockchain technology. First, identify and stop the methods used by attackers, remove trust from the endpoint, and put it on-chain. To assess the BIDPS's efficacy, a testbed was constructed and more than 10 APT attack tactics were used to target the endpoint. Because the Blockchain is immutable, BIDPS can successfully fend off launched attacks, strengthening its detection and prevention operations.

To effectively detect attacks, Nour et al. [28] offer a federated DL-based intrusion detection system (FED-IDS), which transfers learning data from servers to dispersed vehicle edge nodes. To learn the spatial-temporal depictions of vehicle traffic flows required for categorizing various attack types, FED-IDS employs a context-aware transformer system. Miners validate distributed local upgrades on the blockchain to prevent erroneous updates from being added. The outcomes of the experiment demonstrate the viability of protecting intelligent transportation networking against cyberattacks.

III. Problem Statement

Considerable research has been done on the integration of intrusion detection and blockchain to enhance data privacy and identify current and potential threats, respectively. These methods use learning-based ensemble frameworks to protect data privacy while simultaneously making it easier to identify complicated harmful occurrences [29]. The most arduous and time-consuming task is preprocessing data. Thus, even a rookie researcher should be able to choose pertinent features from the dataset and apply them while creating an IDS model. A new technology called the Internet of Things (IoT) is being developed for the creation of numerous vital applications. These apps still use centralized storage design, though, so they face several serious issues, including single points of failure, privacy, and security. Critical IoT network failure points were made visible by a distributed denial of service (DDoS) attack on the cloud [30]. All of the devices have access to the Internet, making the collected heterogeneous IoT urban information more accessible to the public and more susceptible to intellectual property theft by hackers. Network data exchange requires the guarantee of security and secrecy. The majority of networks that hackers infiltrate are ones where data interchange is highly vulnerable.

A. *Research Gap*

In the realm of cloud intrusion detection, existing approaches primarily focus on either traditional ML methods or centralized DL models, both of which come with significant limitations. Traditional ML techniques struggle to handle the large volumes of diverse data generated in cloud environments and often fail to adapt to new or evolving attack patterns. Centralized DL models, while more capable of detecting complex intrusions, suffer from challenges related to scalability, single points of failure, and the potential for compromised data integrity. Additionally, the lack of a robust, tamper-resistant framework for sharing model updates or alerts across cloud nodes can undermine the reliability and trustworthiness of the system, especially in multi-tenant environments where data privacy and security are paramount. The proposed Optimized Blockchain-Based DL Model aims to fill these gaps by introducing a decentralized, blockchain-anchored architecture that leverages the power of DL for cloud intrusion detection while enhancing security, transparency, and scalability. Unlike centralized systems, this model allows distributed cloud nodes to collaboratively detect intrusions, with each node securely sharing model updates through smart contracts on the blockchain. This approach not only decentralizes the decision-making process but also ensures that model integrity is preserved across all participating nodes, eliminating the risk of single points of failure and enhancing the system’s ability to handle real-time threats. Furthermore, by integrating blockchain, the proposed model introduces immutability and auditability in the detection process, addressing the trust and accountability gaps present in current systems. This work, therefore, marks a significant advancement in cloud security, providing a more resilient and transparent system for intrusion detection.

IV. Proposed Methodology

To stop network attacks, an IDS is essential in the field of cybersecurity. The choice engine being used determines how effective it is. A learning system built on blockchain and DL should be used to discover anomalies to boost the system's adaptability. This study offers an IDS based on DL and blockchain, utilizing optimization approaches to identify network threats. Protecting the accuracy and openness of the system depends on the decentralized and unchangeable ledger that blockchain technology offers. Everything about the intrusion detection process is recorded, including transactions and occurrences. Collaborative DL models, specifically CNNs, RBM, and GAN, can be utilized for intrusion detection assignments. The method used to develop these kinds of models are useful for detecting intrusions because they can identify intricate patterns and anomalies using network traffic data. Hash functions based on cryptography are also used in cloud environments to improve security and identify assaults.

A. *Dataset Collection*

The dataset used in this study to evaluate the suggested framework is the UNSW-NB15 [31]. There are 42 attributes in the UNSW-NB15, comprising three categorical inputs and 39 numerical inputs. The data formats (types) for the numerical inputs are binary, integer, and float. Two data subsets are also included in the UNSW-NB15: one for the training phase and another for testing. A thorough explanation of the attack's UNSW-NB15 subset distribution is given in Table I. The proposed technology is given in Fig. 1.

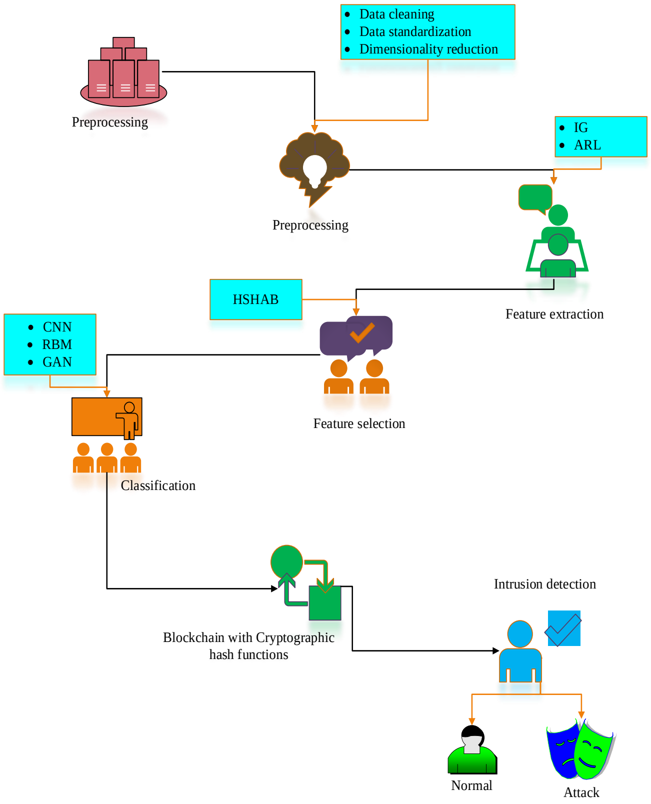


Fig. 1. Process of the developed technique.

TABLE I. DESCRIPTION OF THE UNSW-NB15 DATASET

Types of attack	Total No of data	No of training data	No of validation data	No of testing data	Data Attributes
Generic	40000	30081	9919	18871	1. Flow Features 2. Basic Features 3. Content Features 4. Time and Additional Features
Shellcode	1133	854	279	378	
Exploits	33393	25034	8359	11132	
Worms	130	99	31	44	
Reconnaissance	10491	7875	2616	3496	
Fuzzers	18184	13608	4576	6062	
Analysis	2000	1477	523	677	
Dos	12264	9237	3027	4089	
Backdoor	1746	1330	416	583	
Normal	56000	41911	14089	37000	

Divide the UNSW-NB15 training set into two parts: 25% of the initial training set is employed for validating the trained designs, and the remaining 75% of the original training is utilized to train the models. Furthermore, the validated models are tested using the UNSW-NB15-Test. The 10 assault classes that make up the UNSW-NB15. The data attributes if UNSW-NB15 include:

Flow Features: Attributes like dur (duration of the flow), proto (protocol used), and service (network service on the destination, e.g., HTTP, FTP, DNS).

Basic Features: These describe essential properties of the traffic, such as sbytes (source-to-destination bytes), dbytes (destination-to-source bytes), sttl (source time-to-live), and dttl (destination time-to-live).

Content Features: These focus on the content of packets, including features like sload (source packets per second) and dload (destination packets per second), which help in identifying abnormal patterns in data flows.

Time and Additional Features: These capture specific timing and behavior, such as ct_srv_src (connections to the same service from the same source) and ct_dst_ltm (connections to the same destination in the last 100 connections).

The dataset also includes a class label (label) that distinguishes between normal traffic (0) and attack traffic (1), with various attack types including DoS, backdoor, and shellcode. These attributes provide a detailed perspective of network traffic, enabling in-depth analysis for intrusion detection systems.

B. Preprocessing

One statistical method that makes use of an orthogonal change is PCA [32]. A set of correlated variables is transformed into a set of uncorrelated variables using PCA. For exploratory analysis of data, PCA is employed. Moreover, PCA can be used to investigate the connections between a set of variables. It can therefore be applied to reduce dimensionality.

Let's say a dataset $z^{(1)}, z^{(2)}, z^{(3)}, \dots . z^{(n)}$ has n n-dimension data that must be transformed to dimension inputs i -dimension $(i << n)$ using PCA. Below is a description of PCA:

Data Standardization: Eq. (1) states that the raw data must have a unit variance & a zero mean.

$$k_j^i = \frac{k_j^i - \bar{k}_j}{\sigma_j} \forall j \qquad (1)$$

Using Eq. (2), determine the co-variance vector of the raw data.

$$\sum = \frac{1}{n} \sum_i^n \quad (k_i)(k_i)^t, \sum \in E^{n*n} \qquad (2)$$

Utilizing the formula from Eq. (3), determine the covariance matrix eigenvector and eigenvalue.

$$v^t \sum = \gamma \mu$$

$$V = [\dots v_1 \; v_{2\dots\dots} \; v_n \dots], v_i \in E^n \qquad (3)$$

It is necessary to project raw data onto a n -dimensional subspace: Top n The covariance matrix's eigenvector is selected. Eq. (4) provides the necessary vector calculation.

$$z_i^{new} = [v_1^t z^i \; v_2^t z^i \dots\dots\dots\dots v_k^t z^i \,] \in E^k \quad (4)$$

Thus, if the unprocessed data is with n dimensionality, It'll be condensed into new k data presented in three dimensions.

C. Feature Extraction

To quantify the valuable information extracted from the cleaned dataset, two techniques are employed: Information Gain (IG) and Associate Rule Learning (ARL). When it pertains to feature extraction, IG and ARL work well together to find pertinent traits that make a substantial contribution to the association rules.

1) Information Gain (IG): The value of a feature for forecasting the target variable is measured using IG [33]. When a particular characteristic is known, it computes the decrease in entropy or unpredictability of the target variable. IG is an entropy-based technique for evaluating features that quantify the amount of information a feature provides regarding the target class. IG can identify the features with the highest information based on the target class. Strongly related to the target class, features with high IG are typically chosen to get the greatest classification outcomes. Nevertheless, IG is unable to remove unnecessary features. The amount of information that was available before as well as after the attribute value was known usually determines how much information is gained. For multiple classes, IG can have a maximum value of 1. Eq. (5) provides the formula for entropy analysis of more than two classes.

$$G(Y) = \sum_{i=1}^{k} Q(y_i) Q\,(y_i) \qquad (5)$$

Let, Q is denoted as the number of classes. Moreover, feature $Y$of I_G and the class labels Z is designed in Eq. (6) and Eq. (7).

$$I_G(Y, Z) = G(Y) - G\left(\frac{Y}{Z}\right) \qquad (6)$$

$$G\left(\frac{Y}{Z}\right) = -\sum_j Q(z_i)\sum_i Q\left(\frac{y_i}{z_i}\right)\left(Q\left(\frac{y_i}{z_i}\right)\right) \quad (7)$$

Let, $G(Y)$is denoted as the entropy of Y and $G\left(\frac{Y}{Z}\right)$ is considered as entropy of Y after seeing Z. Since I_G is considered a filter technique, when dealing with big multidimensional data, it scales effectively.

2) Association rule learning: Using ARL [34], one can find intriguing correlations or interactions between variables in huge datasets. These fascinating relationships are typically concealed in unprocessed data, but if they are found and retrieved, they can be effectively used to describe the data. To aid in further identifying intrusions, associations among users and the applications they use must be extracted from the intriguing relationships and added to usage profiles that are both normal and suspicious.

The following is the official description of ARL: Let $K = \{k_1, k_2, \ldots., k_n\}$be a group of n features referred to as system audit data pieces. Let $D = \{d_1, d_2, \ldots. d_m\}$exist a group of entries in this dataset. Each record d_i has a subset of characteristics in K. An assumption of the kind in Eq. (8) is referred to as a rule.

$$R \Rightarrow S, \text{ where } S \subseteq K, \text{and } R \cap S = 0 \quad (8)$$

The subsequent item sets apply to intrusion detection when reading the aforementioned example by this definition: $R = d_1 = users$ and $S = d_2 = used\_users$. The suggestion $R \Rightarrow S,$ is an ARL.

The subsequent item sets apply to intrusion detection when reading the aforementioned example by this definition: the support $Supp(R)$of an item set R, and the assurance $conf(R \Rightarrow S)$of a rule $R \Rightarrow S$as corresponds to Eq. (9).

$$conf(R \Rightarrow S) = \frac{Supp(R \cup S)}{Supp(R)} \quad (9)$$

Support for a set of objects $R\big(Supp(R)\big)$is the percentage of the data set's values that include the item set R. An algorithm for learning association rules consists of two distinct steps: First, determine an appropriate support level threshold and search the data for all likely common sets of items with support values higher than the threshold. Second, rules with confidence values higher than the minimal threshold are constructed using these acquired common item sets.

D. Feature Selection

Hybrid Sea horse and Bat Optimization are used to choose the dataset's best characteristics (HSHBA). It is the result of combining Bat Optimization (BO) [35] with Sea Horse Optimization (SHO) [36]. The movement, predatory behavior, and breeding of sea horses are simulated by the SHO algorithm. The core elements of SHO have three behaviors. The global and locaters techniques are adapted to the motion and hunting behaviors, respectively, to enhance the enhancements of the SHO algorithm. Here, Lévy flying is utilized to mimic the seahorse's movements as it spirals nearer to its greatest advantage. The mathematical relationship is described in Eq. (10).

$$Z_{new}^1(t+1) = Z_i(t) + Levy(\delta)\left(\big(Z_{elite}(t) - Z_i(t)\big) \times a \times b \times c + Z_{elite}(t)\right) \quad (10)$$

Let, $a, b,$ and c are denoted as coordinate vectors in three dimensions (a, b, c) in the spiraling motion, correspondingly. Eq. (11) is used to calculate the Brownian motion with waves.

$$\sigma = \left(\frac{\Gamma(1+\delta) \times sinsin\left(\frac{\pi\delta}{2}\right)}{\Gamma\left(\frac{1+\delta}{2}\right) \times \delta \times 2^{\left(\frac{\delta-1}{2}\right)}}\right) \quad (11)$$

Brownian motion is utilized to imitate the movement of the unit size of the sea horse, which is given by Eq. (12), to the left of the $\overrightarrow{r_1}$ cut-off point to enable to better comprehend the search space in SHO.

$$Z_{new}^1(t+1) = Z_i(t) + rand \times c \times \alpha_t \times (Z_i(t) - \alpha_t \times Z_{elite}) \quad (12)$$

Let, c is considered a constant coefficient, $\alpha_t$is represented as a random walk coefficient for Brownian motion.

Eq. (13) can be used to integrate these two conditions to determine the sea horse's new position at each repetition t.

$$Z_{new}^1(t+1) = \begin{Bmatrix} Z_i(t) + Levy(\delta)\left(\big(Z_{elite}(t) - Z_i(t)\big) \times a \times b \times c + Z_{elite}(t)\right) & \overrightarrow{r_1} > 0 \\ Z_i(t) + rand \times c \times \alpha_t \times (Z_i(t) - \alpha_t \times Z_{elite}) & \overrightarrow{r_1} \leq 0 \end{Bmatrix} \quad (13)$$

This stage involves updating each bat's position $\hat{x}_i$ and velocity $\hat{v}_i$ in a space of dimensions d. $\hat{x}_i$ and $\hat{v}_i$ ought to be updated afterward throughout the iterations. The new solutions $\hat{x}_i(t)$ and velocities $\hat{v}_i(t)$ at time step t. Eq. (14) shows the mathematical algorithm's hybrid efficiency.

$$Z_{new}^2(t+1) = \{\hat{x}_i(t-1) + \hat{v}_i(t)\,\overrightarrow{r_2} > 0.1\ (1-\eta) \times (Z_{new}^1(t) - rand \times Z_{elite}) + \eta \times Z_{new}^1(t)\,\overrightarrow{r_2} \leq 0.1\,\} \quad (14)$$

Let, $Z_{new}^1(t)$ shows the seahorse's updated location at that moment of t, $\overrightarrow{r_2}$ is denoted as a random number [0, 1], It is employed to modify the seahorse's duration of steps during predation, which gets shorter by a linear amount each iteration. Eq. (15) is used to determine the velocity.

$$\hat{v}_i(t) = \hat{v}_i(t-1) + (\hat{x}_i(t-1) - \hat{x}^*)\hat{f}_i \quad (15)$$

Let, β in the range of [0,1] is a vector chosen at random using a uniform range. Here, $\hat{x}^*$ is considered as the current best place (solution) in the world, which is found after evaluating every option among every i bat. Let, $\hat{f}_i$ is considered as frequency and is computed by applying Eq. (16).

$$\hat{f}_i = \hat{f}_{min} + (\hat{f}_{max} - \hat{f}_{min}).\ \beta \quad (16)$$

Let, $\hat{f}_{max}$ and $\hat{f}_{min}$ are denoted as maximum and minimum frequency. Instead, it indicates that the prey is moving more quickly than the seahorse was when it was hunting, allowing the prey to escape and the seahorse to fail to capture it. Using the HSHBA approach, the generated model chooses the best characteristics from the dataset. It improves the IDS system's performance in terms of detection and categorization. To detect intrusion, the chosen features are subsequently modified for the classification stage. The HSHBA algorithm is described in Algorithm.1.

```
Algorithm:1 Feature select process using HSHSA
Start
{
        Initialize $Z_i$
        Compute fitness value// all seahorse
Determine $Z_{elite}$ //best seahorse
While $(t < T)$do
            If $(\overrightarrow{r_1} = random > 0)$do   // movement
            {
            Set constant parameter values
            Execute Brownian motion using eqn. (11)
            Sea horse position updated using eqn.10 and 12
            }
                Else if do
                {
                Update position using eqn. (13).
                }
                End if
Update bat position $\hat{x}_i$ and velocity $\hat{v}_i$  // combine bat
optimization
{
Update the new position using eqn. (14)  //new hybrid model
}
        If $(\overrightarrow{r_2} > 0.1)$
        {
        Select best features
        }
            Else if ( $\overrightarrow{r_2} \leq 0.1$)
            {
            Continue Searching
            }
            End if
Select best features
Enhance prediction accuracy
}
end
```

E. Classification

Ensemble DL models are employed at this stage to enhance the forecast outcomes. For intrusion detection tasks, the created EDL model incorporates convolutional neural networks (CNNs), Generative Adversarial Networks (GAN), and Restricted Boltzmann Machine (RBM). A detailed explanation is provided below.

1) CNN [37]: A multi-layer perceptron specifically created for identifying two-dimensional shapes is called a convolutional network. Convolution kernel characteristics and convolution layer connectivity weights are among the network parameters that are learned during the CNN training procedure. To determine the intrusions, the prediction procedure primarily uses the input data and network parameters. Eq. (17) and Eq. (18) are used to train the chosen features into the CNN algorithm's convolution layer.

$$x^{\leftrightarrow c}_{j} = f\left(v^{\leftrightarrow c}_{j}\right) \tag{17}$$

$$v^{\leftrightarrow c}_{j} = \sum\nolimits_{i \in N_j} x^{\leftrightarrow}{}_{i}^{c-1} * h^{\leftrightarrow c}_{ij} + b^{\leftrightarrow c}_{j} \tag{18}$$

Let, $v^{\leftrightarrow c}_{j}$is denoted as the net activation of the j^{th}network of the convolution layer c, it is obtained by removing the feature map produced by the preceding layer, and convolution averaging $x^{\leftrightarrow}{}_{i}^{c-1}$, $x^{\leftrightarrow}{}_{i}^{c}$is denoted as the output of the j^{th}channel of the convolution layer c. $f(.)$ is considered an activation function and applies tanh and sigmoid operations, among others. N_j is a subset of the feature maps that are utilized as input to compute $v^{\leftrightarrow c}_{j}$, $h^{\leftrightarrow c}_{ij}$ is denoted as convolution kernel matrix, $b^{\leftrightarrow c}_{j}$ is denoted as convolution feature map with bias. Regarding a feature map output $x^{\leftrightarrow c}_{j}$, by every supplied feature map $x^{\leftrightarrow}{}_{j}^{c-1}$might differ. $*$ is denoted as a convolution symbol. Then update $v^{\leftrightarrow c}_{j}$ into the pooling layer via Eq. (19).

$$v^{\leftrightarrow c}_{j} = \sum\nolimits_{i \in N_j} \beta^{\leftrightarrow c}_{j} down\left(x^{\leftrightarrow}{}_{j}^{c-1}\right) + b^{\leftrightarrow c}_{j} \tag{19}$$

Let, $v^{\leftrightarrow c}_{j}$ is denoted as net activation of the j^{th} channel of the pooling layer c, It is produced by balancing and pooling the characteristics map output $x^{\leftrightarrow}{}_{i}^{c-1}$ of the previous layer, β is denoted as pooling layer weighting factor, $b^{\leftrightarrow c}_{j}$ is considered as pooling layer offset, $down(.)$ is denoted as pooling function. Eq. (20) and Eq. (21) are used to weigh the inputs and get the outcome via the activation function, which yields the output of the fully connected layer.

$$x^{\leftrightarrow c} = f(v^{\leftrightarrow c}) \tag{20}$$

$$v^{\leftrightarrow c} = w^{\leftrightarrow c} x^{\leftrightarrow c-1} + b^{\leftrightarrow c} \tag{21}$$

Let, $v^{\leftrightarrow c}$ is considered a fully connected layer net activation function c, it is acquired by removing and filtering the output map. $x^{\leftrightarrow c-1}$is denoted as the previous layer. $w^{\leftrightarrow c}$is considered as a fully connected network weight coefficient, and $b^{\leftrightarrow c}$ is considered a fully connected layer offset c.

2) RBM [38]: The RBM model has a visible layer $\tilde{v}$ with n units and a hidden layer $\tilde{h}$ with m units. In addition, a matrix of actual values $\widetilde{w}_{n \times m}$ replicas the proportions of visible to hidden neurons, where $\widetilde{w}_{ij}$ is denoted as the visible unit connection $\tilde{v}_i$ and the hidden unit $\tilde{h}_j$. The data is primarily received by the visible layer for processing, while its pattern and probability distribution are learned by the hidden layer. Furthermore, probably every layer's unit $\tilde{v}$ and $\tilde{h}$ are binary It came from the distribution of Bernoulli, i.e., $\tilde{v} \in \{0,1\}^n, \tilde{h} \in \{0,1\}^m$. Eq. (22) calculates an RBM's energy function:

$$E\left(\tilde{v}, \tilde{h}\right) = -\sum\nolimits_{i=1}^{n} \tilde{x}_i \tilde{v}_i - \sum\nolimits_{j=1}^{m} \tilde{y}_i \tilde{h}_j - \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{m} \tilde{v}_i \tilde{h}_j \widetilde{w}_{ij} \tag{22}$$

Let, x and y are denoted as hidden and visible unit biases. Furthermore, Eq. (23) simulates the combined likelihood of a specific configuration $\left(\tilde{v}, \tilde{h}\right)$:

$$P\left(\tilde{v}, \tilde{h}\right) = \frac{e^{-E\left(\tilde{v}, \tilde{h}\right)}}{f_p} \tag{23}$$

Let, f_p is considered a partition function, which, while taking into account visible and hidden units, restores the chance over all conceivable configurations. Essentially, an RBM must become familiar with a set of parameters. using an algorithm for training. For every training set, it maximizes the sum of data possibilities ξ, which is described in Eq. (24)

$$argmax_{\Theta} \prod_{\tilde{v}\in\xi} P(\tilde{v}) \tag{24}$$

Implementing the negative of the logarithm functions, which is denoted by the negative log-likelihood (NLL), to describe this problem is an intriguing method. The NLL indicates the distribution estimation of the new information over the original data. Consequently, one can use the partial derivatives to calculate the derivatives of $\tilde{W}$, $\tilde{x}$ and $\tilde{y}$at iteration t. The parameter updating rules are described by Eq. (25–27).

$$\tilde{W}^{t+1} = \tilde{W}^{t} + \eta\left(\tilde{v}P(\tilde{v},\tilde{h}) - \tilde{v}P(v,h)\right) \tag{25}$$

$$\tilde{x}^{t+1} = \tilde{x}^{t} + (\tilde{v} - v) \tag{26}$$

$$\tilde{y}^{t+1} = \tilde{x}^{t} + \left(P(\tilde{v},\tilde{h}) - P(v,h)\right) \tag{27}$$

Let, η is denoted as the learning rate, v is considered as the visible layer's reconstruction $\tilde{v}$, and h is considered as the hidden vector's estimation $\tilde{h}$ given v.

3) GAN [39]: A GAN consists of two parts, a generator g_e and a discriminator d_r, that compete with one another. g_e uses a noise vector as its input $\vec{n}$ and intends to provide high-quality fake data that closely approximates the original data. Moreover, d_r seeks to identify authentic data from artificially created data. The min-max goal function o_f is used to represent Eq. (28).

$$\underset{g_e}{min}\underset{d_r}{max} o_f\left(g_{e,}d_r\right) = E_{\vec{x}_r\sim P}\left[log\ log\ \left(d_r(\vec{x}_r)\right)\right] + E_{\vec{n}\sim M}\left[log\ log\ \left(1 - d_r\left(g_e(\vec{n})\right)\right)\right] \tag{28}$$

Let, $\vec{x}_r \sim P$is considered as the actual distribution of data and $\vec{n} \sim M$ is denoted as Gaussian noise distribution. $d_r(\vec{x})$is denoted as outputs. The generator g_e gather $\vec{n}$ input to classify the intrusion.

a) Ensemble DL techniques: The maximum voting of each classifier determines which of the obtained prediction results is chosen. Max voting [40] entails gathering predictions for every class label and projecting which class label will receive the greatest number of votes using Eq. (29). Soft voting is an additional kind of maximum voting. In soft voting, Eq. (30) illustrates, that predicted chances are gathered for each class identity, and the class identity with the highest probability is predicted.

$$z' = [C_1'(\hat{x}), C_2'(\hat{x}), C_3'(\hat{x})] \tag{29}$$

Let, z' is denoted as the majority vote of each classifier, determines the class label C_1',C_2', and C_3'.

$$z' = argmax_i \sum_{j=1}^{n} W_j' P_{ij}' \tag{30}$$

Let, W_j' is considered as the weight that can be allocated to the j^{th}classifier.

F. Blockchain with Cryptographic Hash Function [41]

A blockchain's primary role is to offer a cryptographically safe method for collecting a permanent and globally verifiable collection of documents, known as blocks, that are systematically arranged by separate time stamps. Blockchains are commonly utilized as a distributed, open database of data transactions since they are frequently shared and synchronized via a peer-to-peer network. Every member of the blockchain network has access to the record data, which they can use to accept, reject, or verify using a consensus procedure. Records are added to the blockchain in the same sequence that they were verified after they are approved.

The foundation of blockchains is the cryptographically secure hash function, a fundamental building block of cryptography. These hashing algorithms $\hat{H}: \{0,1\}^* \to \{0,1\}^n$map an input of any length to an output with a fixed length of n bits, and it needs to meet the security constraints listed below:

Preimage resistance: Considering a hash value $\hat{h}$, It ought to be necessary to $\Phi(2^n)$work involved with computing an $\hat{x}$ such that $\hat{H}(\hat{x}) = \hat{h}$.

Second preimage resistance: the input $\hat{x}$ and hash value $\hat{h} = \hat{H}(\hat{x})$ are needed in $O(2n\ )$ for computing $\hat{x}' \neq \hat{x}$ such that $H(x\ 0\ ) = h.3)$.

Collision resistance: Need $\Phi(2^n)$determination to calculate any two $\hat{x}' \neq \hat{x}$ such that $\hat{H}(\hat{x}') = \hat{h}$

The opponent does not influence the real hash value in collisions. (Second) preimage resistance is particularly important in the context of blockchains because attackers might change current blocks while maintaining the chain if they could identify second preimages with a specific mixfix. The aforementioned security criteria state that an attack of this kind needs to be at least 2^n for a hash function with n bits. The IDS determines each file's hash value when scanning records on a system and matches it to the store. If a similarity is discovered, malware is probably present.

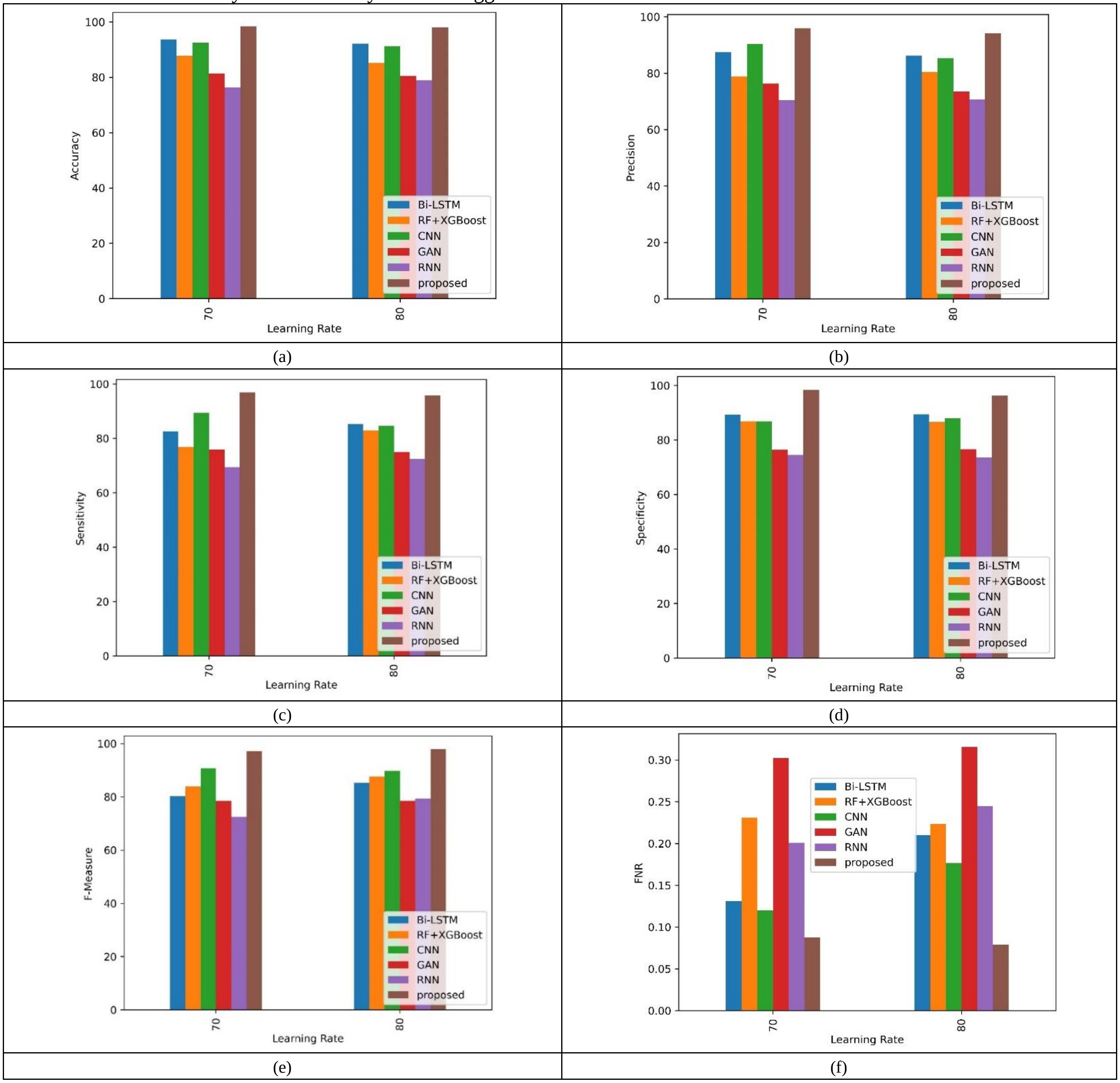
V. Results and Discussion

This study employed a variety of metrics to assess the efficacy of the proposed model. The BEDL model testing and training processes were conducted using Python. In the experiment, two learning rates 70 and 80 were chosen for the analysis of the study. The EDF method produced good results in many classification procedures. The presented method uses the BEDF model and hybrid optimization to increase the resilience of cloud computing. The architecture of proposed BEDL involves a decentralized network of cloud nodes that work collaboratively to detect intrusions. Each node in the blockchain network runs the DL IDS, where updates to the feature selection results are recorded immutably on the blockchain to ensure model integrity. The architecture comprises smart contracts that govern data sharing, model updates, and anomaly detection reporting across nodes. Smart Contracts are designed to trigger automatic actions such as initiating a model retraining process when new intrusion patterns are detected for accurate detections. Additionally, the smart contracts enforce data privacy by facilitating secure, encrypted communications between cloud nodes while ensuring that any detection-related alerts are stored immutably and transparently across the blockchain, guaranteeing auditability and trust. The use of consensus algorithms ensures that only validated model updates are propagated across the network, improving both security and collaboration in intrusion detection. Furthermore, the dataset used for investigation is

UNSW-NB15 which is available in https://www.kaggle.com/datasets/dhoogla/unswnb15. It is a benchmark dataset designed for evaluating IDS. It was created by simulating real-world network traffic at the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS), incorporating modern attack vectors. The dataset contains 49 features and 9 different attack categories, such as DoS, worms, backdoors, and exploits, alongside normal traffic. It provides both labeled and unlabeled data for training and testing machine learning models. The UNSW-NB15 dataset is widely used due to its diversity and realistic network behavior. The performance measures utilized to analyze the efficacy of the suggested technique are F1-score, False Positive Rate (FPR), False Negative Rate (FNR), Matthews Correlation Coefficient (MCC), Negative Predictive Value (NPV), accuracy, precision, sensitivity, and specificity.

A. *Performance Analysis*

Two learning rates 70 and 80 are employed for the training and testing of the developed technique. The experimental findings are tested against a variety of accessible DL classifiers, including Bi-LSTM [21], RF+XGBoost [23], RNN [40], CNN [37], and GAN [39].



(a)

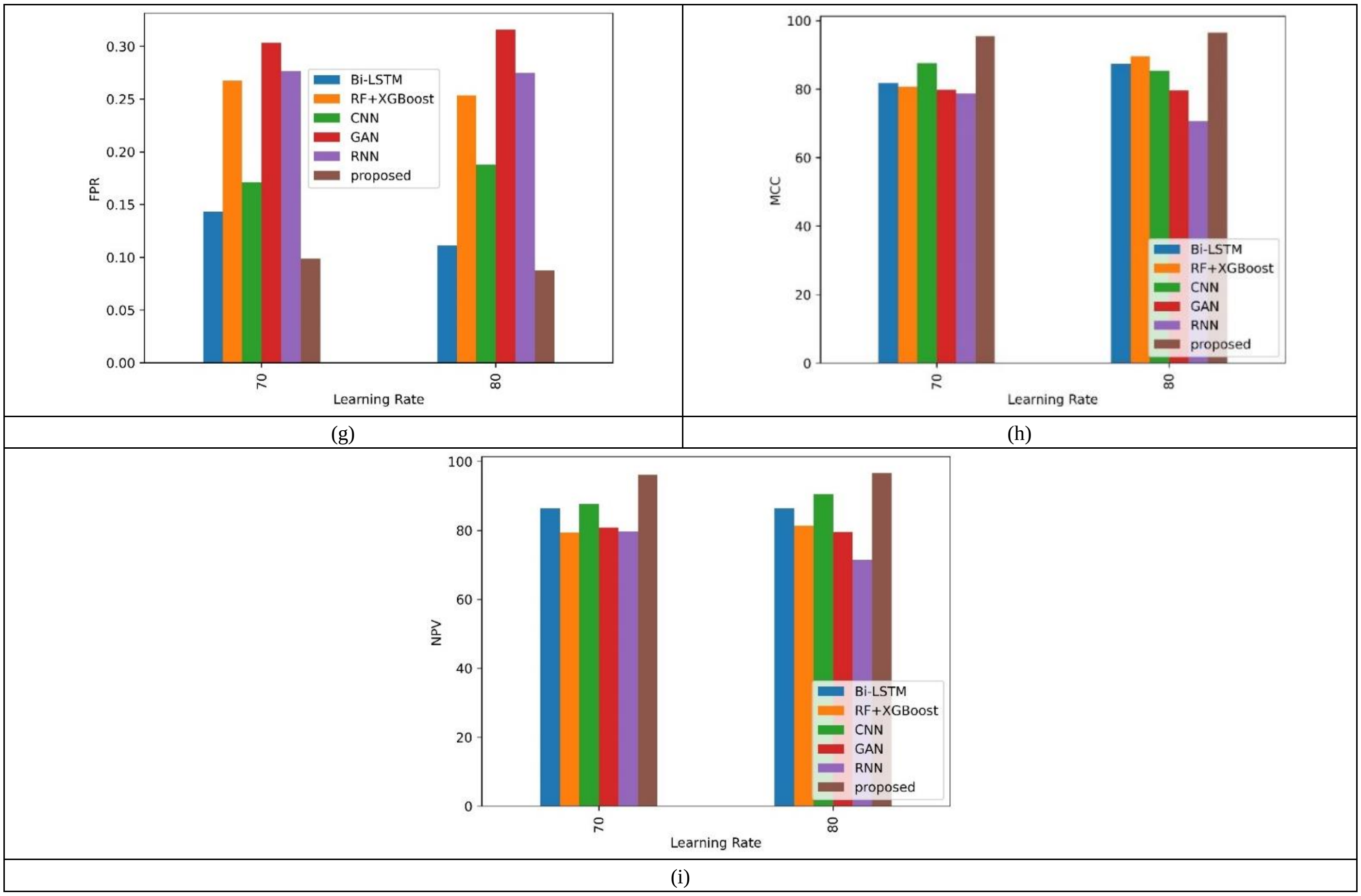
(b)

(c)

(d)

(e)

(f)



(g)

(h)

(i)

Fig. 2. Performance of proposed BEDL over Baseline Models for (a) Accuracy, (b) Precision, (c) Sensitivity. (d) Specificity, (e) F1-Score, (f) FNR, (g) FPR, (h) MCC, and (i) NPV.

1) Accuracy: The accuracy scores that several models obtained on the test are shown in Fig. 2(a), along with a suggested model, for two alternative scenarios: one with a learning rate of 70 and the other with a learning rate of 80. The Bi-LSTM model demonstrated a maximum accuracy of 93.655% at a learning rate of 70, followed by the CNN model's 92.4763% accuracy. The accuracy of the suggested model is 98.4763%. With an accuracy of 81.3652%, the GAN model was the least accurate of the models on the list; the RNN model was somewhat more accurate at 76.3764%. With an accuracy of 87.7766%, the ensemble methods RF+XGBoost fared better in the meantime. When comparing the accuracy ratings of all models in the first scenario to the second, which had a higher learning rate of 80, they generally declined. With 92.1121% accuracy, the Bi-LSTM model was still ahead of the CNN model, which came in second at 91.2344%. The suggested model saw a decline to 97.9987% but still displayed great accuracy. The accuracy of the ensemble methods RF+XGBoost similarly decreased, reaching 85.2235%. With the GAN model at 80.4762% and the RNN model at 78.8776%, the GAN and RNN models displayed similar patterns as in the prior situation.

2) Precision: The precision scores under two distinct learning rates (70 and 80) are shown in Fig. 2(b) for a variety of models, including Bi-LSTM, RF+XGBoost, CNN, GAN, RNN, and a proposed model. Precision is a metric that expresses the percentage of genuine positive predictions among all positive predictions, assessing how accurately a model predicts the future. The Bi-LSTM system achieves 87.54% precision at a learning rate of 70, while the RF+XGBoost model follows with 78.89% precision. 90.38% is the precision achieved by the CNN model, 76.37% by the GAN model, and 70.39% by the RNN model. Remarkably, with a precision of 95.97%, the suggested model beats all others, demonstrating its superiority in correctly predicting positive events at this learning rate. There is a tiny difference in the precision values below a learning rate of 80. The precision of the RF+XGBoost model rises to 80.48%, while that of the Bi-LSTM model falls to 86.23%. The precision of the GAN model falls to 73.48%, the RNN model stays almost the same at 70.77%, and the CNN model's precision reduces to 85.33%. In a similar vein, the precision of the suggested model drops but stays high at 94.22%.

3) Sensitivity: Sensitivity values for a range of models, including CNN, GAN, RNN, RF+XGBoost, Bi-LSTM, and a suggested model, are shown in Fig. 2(c) for various learning rates. True positive rate, another name for sensitivity, is the percentage of real positive cases that the model properly recognized. The suggested model beats the others with a sensitivity of 96.88% at a learning rate of 70, demonstrating its higher capacity to accurately detect positive cases. The CNN and Bi-LSTM models, with corresponding sensitivity values of 89.48% and 82.66%, trail closely behind. With 76.88% and 69.47% sensitivity, respectively, RF+XGBoost and RNN outperform the GAN model, which comes in last with 75.99% sensitivity. The suggested model keeps its high sensitivity at 95.77% as the learning rate rises to 80, demonstrating its efficacy. Notably, the RF+XGBoost model outperforms the CNN and Bi-LSTM models in this configuration, with a sensitivity of 82.96%. However when compared to the other models, the GAN model's sensitivity is still the lowest at 74.99%, suggesting its relative weakness in correctly identifying positive cases.

4) Specificity: The specificity values of several models, as well as a suggested model under two distinct learning rates (70 and 80), are shown in Fig. 2(d). In binary classification, specificity is a metric that shows the percentage of real negative cases that the model correctly classifies as such. The Bi-LSTM model gets the maximum specificity of 89.37% under a learning rate of 70, followed by the CNN model at 87.99%. Additionally, the suggested model functions effectively, with a 98.35% specificity. It's important to keep in mind, though, that the GAN model performs comparatively worse than the others in terms of specificity, only reaching 76.48%. There are some variations in the models' performance when the learning rate is raised to 80. The Bi-LSTM model increases somewhat to 89.38% while maintaining its high specificity. There is also a minor improvement to 87.99% for the CNN model. Though it still performs noticeably better than most models, the suggested model's performance drops to 96.21%. Notably, as compared to its performance at the lower learning rate, the RNN model exhibits a drop in specificity.

5) F1-Score: The F-Measure performance scores under two distinct learning rates, 70 and 80, are displayed in Fig. 2(e) for a variety of models, including Bi-LSTM, RF+XGBoost, CNN, GAN, RNN, and a proposed model. The Bi-LSTM model obtains an F-Measure of 80.36% at a learning rate of 70, whereas RF+XGBoost does marginally better at 83.99%. CNN has the best performance, coming in at 90.78%, and the suggested model comes in at 97.23%. The F-Measure scores of GAN and RNN are lower, at 78.48% and 72.48%, respectively. A learning rate increase of 80 improves performance for the majority of models. Bi-LSTM sees a slight improvement to 85.23%, RF+XGBoost to 87.68%, CNN to 89.74%, and GAN and RNN to 79.39% and 78.56%, respectively. CNN experiences a slight decline. The suggested model demonstrates a significant rise to 97.99%.

6) FNR and FPR: The False Negative Rate (FNR) performance of the various models at the two learning rates 70 and 80 is displayed in Fig. 2(f). A distinct model, such as Bi-LSTM, RF+XGBoost, CNN, GAN, RNN, and a suggested model, is shown by each column. With a FNR of 0.087662, the suggested model beats all others at a learning rate of 70. CNN and Bi-LSTM both show comparatively low FNRs of 0.12004 and 0.13123, respectively. Nevertheless, models with higher FNR values, ranging from 0.20093 to 0.30234, include RF+XGBoost, GAN, and RNN. A discernible change in the models' performance occurs when the learning rate is raised to 80. The suggested model continues to have the lowest FNR (0.078876), indicating its resilience. Comparing the FNR values of Bi-LSTM, RF+XGBoost, and CNN to the 70-learning rate scenario, however, reveals a modest gain. Notably, with values over 0.24, GAN and RNN continue to show greater FNRs than the other models. On the other hand, all models perform somewhat better when the learning rate is raised to 80. Notably, the suggested model keeps the lowest FPR, demonstrating its superiority over the other models even more. Additionally, Bi-LSTM exhibits a significant drop in FPR in Fig. 2(g), demonstrating its sensitivity to variations in learning rate. The FPRs of RF+XGBoost, CNN, and GAN are comparable, while the performance of RNN is largely constant.

7) MCC: The scores for several models and a suggested model are shown in Fig. 2(h) for two separate scenarios: one with a learning rate of 70 and the other with a learning rate of 80. The models perform differently in the first situation when the learning rate is 70. The suggested model achieves 95.56553, while the Bi-LSTM model reaches 81.8009, RF+XGBoost at 80.6775, CNN at 87.6544, GAN at 79.7668, and RNN at 78.74662. Interestingly, the suggested model performs noticeably better than the others, demonstrating its usefulness in this situation. Significant differences in the model's performance can be seen in the second scenario, which uses a higher learning rate of 80. The Bi-LSTM model outperforms RF+XGBoost at 89.657, with an MCC of 87.3766. Nevertheless, the performance of the GAN and RNN models further declines to 79.65564 and 70.6553, respectively, while the CNN model's performance reduces to 85.4773. With an MCC of 96.4878, the suggested model maintains its strong performance despite these modifications, demonstrating its resilience and superiority over the other models in this situation.

8) NPV: The performance measures, namely the net present value (NPV), of various models in two distinct situations are displayed in Fig. 2(i) one with a learning rate of 70 and the other with an 80. A suggested model, CNN, GAN, RNN, RF+XGBoost, Bi-LSTM, and RNN are among the models that are compared. The suggested model performs better than other models in the first scenario with a 70-learning rate, obtaining a value of 6.1232. With an NPV of 87.6598, CNN outperforms the others, while Bi-LSTM, with an NPV of 86.3624, is not far behind. The NPV values of RF+XGBoost, GAN, and RNN are 79.3765, 80.87763, and 79.68773. With an NPV of 96.5886, the suggested model maintains its advantage in the second scenario with an 80-learning rate. CNN outperforms the first scenario by a substantial margin, obtaining the highest NPV among the models after the suggested one, 90.4874. Next, with an NPV of 81.3885, is RF+XGBoost. But with an NPV of 71.3874, RNN's performance significantly deteriorates, and it becomes the least-performing model out of all of them. To enhance the robustness and security of the cloud environment, a cryptographic hash function is generated using blockchain. The gained experimental outcomes are compared with existing classifiers and attained better experimental outcomes. The designed technique gained accuracy of 98.47%, and 97.99% for 70 and 80 learning rates, and also gained less FPR of 0.098, and 0.087 for 70 and 80 learning rates. The developed technique improves the performance and IDS and enhances the efficiency and robustness by using blockchain. In the future, improving blockchain networks' scalability is essential to enabling the widespread deployment of IDS. Subsequent investigations may concentrate on creating innovative consensus processes or layer-2 scaling approaches to manage the growing number of events and transactions produced by IDS sensors.

B. Real-time Implementation Model

A pilot deployment is conducted across a distributed cloud infrastructure. The system's performance is assessed by simulating various intrusion scenarios, including DDoS attacks [41], unauthorized access attempts, and insider threats. Key performance indicators such as detection accuracy, false positive/negative rates, latency in intrusion detection, and blockchain transaction throughput are monitored. Additionally, the scalability of the system is evaluated by increasing the number of cloud nodes and analyzing the consensus efficiency, smart contract execution times, and resource utilization. The evaluation also considers the impact of network delays, data privacy enforcement, and system robustness in handling high-traffic environments, ensuring the solution's practicality and effectiveness for real-world cloud security.

When a DDoS attack occurs targeting the platform hosted on cloud system, the proposed BEDL model detects abnormal traffic spikes. Once an anomaly is detected, a smart contract is triggered, validating the threat and broadcasting it across the blockchain. This ensures all nodes in the network are aware of the attack, preventing it from spreading further. The smart contract also records the event immutably for future audits and triggers automated actions such as load balancing and firewall rule updates to mitigate the threat in real-time, enhancing both the security and resilience of the cloud infrastructure.

VI. Conclusion

This paper designs blockchain-based DL models to enhance the security of cloud computing. Three types of DL techniques are combined such as CNN, GAN, and RBM to enhance the prediction results of the developed model. UNSW-NB15 dataset is collected and they are cleaned, standardized, and reduced dimensionality using preprocessing. They select the best features to improve the attack prediction rate using the HSHBA model. Additionally, detect the intrusion present in the cloud using a blockchain-based EDL model. The final results are predicted based on the majority and soft voting of the designed technique. Future research could explore integrating federated learning [42] with blockchain to further enhance data privacy in cloud intrusion detection, allowing decentralized model training without sharing sensitive data. Additionally, adopting quantum-safe cryptography in the blockchain layer could future-proof the system against quantum computing threats. The scalability of the proposed architecture can be improved through layer-2 blockchain solutions to reduce latency. Research can also focus on adaptive DL models that evolve with emerging threats in real time. Lastly, expanding the system's application to edge computing environments could enhance security in IoT-based cloud ecosystems.

Acknowledgment

The author extends the appreciation to the Deanship of Postgraduate Studies and Scientific Research at Majmaah University for funding this research work through the project number (R-2024-1245).

References

[1] V. Saravanan, M. Madiajagan, S.M. Rafee, P. Sanju, T.B. Rehman, and B. Pattanaik, “IoT-based blockchain intrusion detection using optimized recurrent neural network,” Multimedia Tools and Applications, pp. 1-22, 2023.

[2] A.A. Khan, M.M. Khan, K.M. Khan, J. Arshad, and F. Ahmad, “A blockchain-based decentralized machine learning framework for collaborative intrusion detection within UAVs,” Computer Networks, vol. 196, p. 108217, 2021.

[3] R. Kumar, P. Kumar, R. Tripathi, G.P. Gupta, N. Kumar, and M.M. Hassan, “A privacy-preserving-based secure framework using blockchain-enabled deep-learning in cooperative intelligent transport system,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 16492-16503, 2021.

[4] S. Badri, “HO-CER: Hybrid-optimization-based convolutional ensemble random forest for data security in healthcare applications using blockchain technology,” Electronic Research Archive, vol. 31, no. 9, pp. 5466-5484, 2023.

[5] S. Thiruvenkatasamy, R. Sivaraj, and M. Vijayakumar, “Blockchain Assisted Fireworks Optimization with Machine Learning based Intrusion Detection System (IDS),” Tehnički vjesnik, vol. 31, no. 2, pp. 596-603, 2024.

[6] I. Katib and M. Ragab, “Blockchain-assisted hybrid harris hawks optimization based deep DDoS attack detection in the IoT environment,” Mathematics, vol. 11, no. 8, p. 1887, 2023.

[7] D.K. Jain, W. Ding, and K. Kotecha, “Training fuzzy deep neural network with honey badger algorithm for intrusion detection in cloud environment,” International Journal of Machine Learning and Cybernetics, vol. 14, no. 6, pp. 2221-2237, 2023.

[8] L. Almuqren, K. Mahmood, S.S. Aljameel, A.S. Salama, G.P. Mohammed, and A.A. Alneil, “Blockchain Assisted Secure Smart Home Network using Gradient Based Optimizer with Hybrid Deep Learning Model,” IEEE Access, 2023.

[9] O. Shende, R.K. Pateriya, P. Verma, and A. Jain, “CEBM: collaborative ensemble blockchain model for intrusion detection in IoT environment,” 2021.

[10] E. Ntizikira, L. Wang, J. Chen, and K. Saleem, “Honey-block: Edge assisted ensemble learning model for intrusion detection and prevention using defense mechanism in IoT,” Computer Communications, vol. 214, pp. 1-17, 2024.

[11] S.K. Kanna, M.Y.B. Murthy, M.B. Gawali, S.M. Rubai, N.S. Reddy, G. Brammya, and N.S. Preetha, “A deep learning-based disease diagnosis with intrusion detection for a secured healthcare system,” Knowledge and Information Systems, pp. 1-39, 2024.

[12] S.K. Gupta, M. Tripathi, and J. Grover, “Hybrid optimization and deep learning based intrusion detection system,” Computers and Electrical Engineering, vol. 100, p. 107876, 2022.

[13] A. Aljabri, F. Jemili, and O. Korbaa, “Convolutional neural network for intrusion detection using blockchain technology,” International Journal of Computers and Applications, vol. 46, no. 2, pp. 67-77, 2024.

[14] W. Dhifallah, T. Moulahi, M. Tarhouni, and S. Zidi, “Intellig_block: Enhancing IoT security with blockchain-based adversarial machine learning protection,” International Journal of Advanced Technology and Engineering Exploration, vol. 10, no. 106, p. 1167, 2023.

[15] S. Siddamsetti and M. Srivenkatesh, "Implementation of Blockchain with Machine Learning Intrusion Detection System for Defending IoT Botnet and Cloud Networks," Ingénierie des Systèmes d'Information, vol. 27, no. 6, 2022.

[16] A.A. Sharadqh, H.A.M. Hatamleh, S.S. Saloum, and T.A. Alawneh, "Hybrid chain: blockchain enabled framework for bi-level intrusion detection and graph-based mitigation for security provisioning in edge assisted IoT environment," IEEE Access, vol. 11, pp. 27433-27449, 2023.

[17] E. Ashraf, N.F. Areed, H. Salem, E.H. Abdelhay, and A. Farouk, "FIDChain: Federated intrusion detection system for blockchain-enabled IoT healthcare applications," In Healthcare, vol. 10, no. 6, p. 1110, June 2022.

[18] R. Kumar, P. Kumar, M. Aloqaily, and A. Aljuhani, "Deep-learning-based blockchain for secure zero touch networks," IEEE Communications Magazine, vol. 61, no. 2, pp. 96-102, 2022.

[19] A. Albakri, B. Alabdullah, and F. Alhayan, "Blockchain-assisted machine learning with hybrid metaheuristics-empowered cyber attack detection and classification model," Sustainability, vol. 15, no. 18, p. 13887, 2023.

[20] E.M. Onyema, S. Dalal, C.A.T. Romero, B. Seth, P. Young, and M.A. Wajid, "Design of intrusion detection system based on cyborg intelligence for security of cloud network traffic of smart cities," Journal of Cloud Computing, vol. 11, no. 1, p. 26, 2022.

[21] O. Alkadi, N. Moustafa, B. Turnbull, and K.K.R. Choo, "A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks," IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9463-9472, 2020.

[22] R. Kumar, P. Kumar, R. Tripathi, G.P. Gupta, S. Garg, and M.M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," Journal of Parallel and Distributed Computing, vol. 164, pp. 55-68, 2022.

[23] R.F. Mansour, "Blockchain assisted clustering with intrusion detection system for industrial internet of things environment," Expert Systems with Applications, vol. 207, p. 117995, 2022.

[24] E.S. Babu, B.K.N. SrinivasaRao, S.R. Nayak, A. Verma, F. Alqahtani, A. Tolba, and A. Mukherjee, "Blockchain-based Intrusion Detection System of IoT urban data with device authentication against DDoS attacks," Computers and Electrical Engineering, vol. 103, p. 108287, 2022.

[25] C. Liang, B. Shanmugam, S. Azam, A. Karim, A. Islam, M. Zamani, S. Kavianpour, and N.B. Idris, "Intrusion detection system for the internet of things based on blockchain and multi-agent systems," Electronics, vol. 9, no. 7, p. 1120, 2020.

[26] S.R. Khonde and V. Ulagamuthalvi, "Blockchain: Secured Solution for Signature Transfer in Distributed Intrusion Detection System," Computer Systems Science & Engineering, vol. 40, no. 1, 2022.

[27] L. Alevizos, M.H. Eiza, V.T. Ta, Q. Shi, and J. Read, "Blockchain-enabled intrusion detection and prevention system of APTs within zero trust architecture," IEEE Access, vol. 10, pp. 89270-89288, 2022.

[28] M. Abdel-Basset, N. Moustafa, H. Hawash, I. Razzak, K.M. Sallam, and O.M. Elkomy, "Federated intrusion detection in blockchain-based smart transportation systems," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 3, pp. 2523-2537, 2021.

[29] G.G. Gebremariam, J. Panda, and S. Indu, "Blockchain-based secure localization against malicious nodes in IoT-based wireless sensor networks using federated learning," Wireless communications and mobile computing, 2023.

[30] R. Kumar, P. Kumar, R. Tripathi, G.P. Gupta, A.N. Islam, and M. Shorfuzzaman, "Permissioned blockchain and deep learning for secure and efficient data sharing in industrial healthcare systems," IEEE Transactions on Industrial Informatics, vol. 18, no. 11, pp. 8065-8073, 2022.

[31] https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15

[32] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, R. Kaluri, D.S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," IEEE Access, vol. 8, pp. 54776-54788, 2020.

[33] D. Stiawan, M.Y.B. Idris, A.M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," IEEE Access, vol. 8, pp. 132911-132921, 2020.

[34] S. Sivanantham, V. Mohanraj, Y. Suresh, and J. Senthilkumar, "Association Rule Mining Frequent-Pattern-Based Intrusion Detection in Network," Computer Systems Science & Engineering, vol. 44, no. 2, 2023.

[35] X.S. Yang and A. Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization," Engineering computations, vol. 29, no. 5, pp. 464-483, 2012.

[36] F.A. Özbay, "A modified seahorse optimization algorithm based on chaotic maps for solving global optimization and engineering problems," Engineering Science and Technology, an International Journal, vol. 41, p. 101408, 2023.

[37] J. Cheng, Y. Liu, X. Tang, V.S. Sheng, M. Li, and J. Li, "DDoS Attack Detection via Multi-Scale Convolutional Neural Network," Computers, Materials & Continua, vol. 62, no. 3, 2020.

[38] G.H.D. Rosa, M. Roder, D.F. Santos, and K.A. Costa, "Enhancing anomaly detection through restricted Boltzmann machine features projection," International Journal of Information Technology, vol. 13, pp. 49-57, 2021.

[39] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.Y. Wang, "Generative adversarial networks: introduction and outlook," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588-598, 2017.

[40] M.A. Khan, "HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system," Processes, vol. 9, no. 5, p. 834, 2021.

[41] A. I.Gide, and A.A. Mu'azu, "A Real-Time Intrusion Detection System for DoS/DDoS Attack Classification in IoT Networks Using KNN-Neural Network Hybrid Technique", Babylonian Journal of Internet of Things, pp. 60-69, 2024.

[42] R. T. Hameed, and O. A. Mohamad, "Federated Learning in IoT: A Survey on Distributed Decision Making", Babylonian Journal of Internet of Things, Vol. 2023, pp. 1-7, 2023.

Deep Learning for Stock Price Prediction and Portfolio Optimization

Ashy Sebastian, Dr. Veerta Tantia
Department of Commerce, Christ University, Bengaluru, India

***Abstract*—Using deep learning for stock market predictions and portfolio optimizations is a burgeoning field of research. This study focuses on the stock market dynamics in developing countries, which are often considered less stable than their developed counterparts. The study is structured in two stages. In the first stage, the authors introduce a stacked LSTM model for predicting NIFTY stocks and then rank the stocks based on their predicted returns. In the second stage, the high-return stocks are selected to form 30 different portfolios with six different objectives, each comprising the top 7, 8, 9, and 10 NIFTY stocks. These portfolios are then compared based on risk and returns. Experimental results show that portfolios with five stocks offer the best returns and that adding more than nine stocks to the portfolio leads to excessive diversification and complexity. Therefore, the findings suggest that the proposed two-stage portfolio optimization method has the potential to construct a promising investment strategy, offering a balance between historical and future information on assets.**

Keywords*—Deep learning; long-short term memory; stock price prediction; portfolio optimization; emerging markets; Indian stock market

I. Introduction

Portfolio optimization is an essential component of a trading system. The goal of optimization is to determine the optimal asset allocation within a portfolio to maximize returns for a given level of risk. This concept, widely known as modern portfolio theory (MPT), was pioneered by study [1].

The main advantage of creating an optimal portfolio is that it encourages diversification, which helps stabilize the equity curve and results in a higher return per unit of risk than trading a single asset. Nevertheless, despite the undeniable power of such diversification, the selection and implementation of the right asset allocations in a portfolio can be challenging due to significant fluctuations in financial market dynamics over time. For instance, assets that exhibited strong negative correlations in the past could be positively correlated in the future, and individual assets in the same asset class often show high positive correlations [2]. This adds extra risk to the portfolio, degrades its subsequent performance and undermines investor confidence

Most traditional portfolio research overlooks the selection of high-quality assets and instead focuses on enhancing strategy performance. However, pre-selecting high-quality assets is crucial for optimal portfolio formation [3]. Zhou et al. [4] noted that the success of portfolio management relies on the initial selection of high-quality stocks. Investors attempt to predict the future returns of stocks and determine the optimal weight based on the highest predicted returns to construct a portfolio [5]**.** During the portfolio optimization process, the expected return on an asset is a vital consideration, highlighting the importance of preliminary asset selection in effective portfolio management [6]. Doing this substantially reduces the scope of possible assets to choose from. Typically, high profits are associated with high risk. However, in portfolio optimization, risk can be minimized by selecting optimal securities and assets. Thus, combining forecasting theory with portfolio selection could improve portfolio returns [7].

With advancements in machine learning and deep learning, though predicting asset returns has become feasible, these prediction results are not yet effectively utilized in practice for portfolio creation and optimization. As a result, many portfolios fail to fully capitalize on the available predictive insights, limiting their potential for improved performance and risk management. The challenge lies in effectively utilizing these predicted returns to construct an optimal investment portfolio. Considering this, the research seeks to address the issue by exploring how to integrate advanced forecasting information into the portfolio selection process.

Literature evidence that the strong functioning of stock markets has a significant effect on the overall growth of an economy, especially in a developing one. Nevertheless, forecasting in emerging markets is more difficult than in developed ones [8] due to their greater volatility, which can be affected by external reasons such as oil prices and the performance of developed markets. Emerging markets, such as China and India, comprise a significant portion of the global economy. Trading in these markets requires a different approach than in developed markets, as they possess unique characteristics such as higher volatility and potentially greater average returns. These markets often lack full information efficiency, owing to institutional barriers that impede information flow and the inexperience of market participants in quickly incorporating new information into security prices. In [9], a genetic algorithm was used to select stock portfolios in the emerging Asian markets of Vietnam, Thailand, Philippines, Singapore, Malaysia, and Indonesia. Although the Indian stock market has made significant advancements, there is a dearth of research on prediction-based portfolios. Furthermore, NSE (National Stock Exchange) and BSE (Bombay Stock Exchange), India's two major stock exchanges, have around 1600 and 5000 listed companies, respectively, traded on them daily. Hence, it is challenging for individual investors to decide on the number and type of stocks. Therefore, the development of appropriate asset selection and portfolio optimization models can assist investors in emerging markets to maximize profits and minimize risk. Sen et al. [10] conducted a study on the Indian stock wherein the top five stocks from nine different sectors of NSE were taken for creating minimum-variance and optimum-risk portfolios. An LSTM (Long-short term memory) model was then designed to predict the future prices of the stocks in each portfolio. Five months after the portfolio construction, the actual return and the return predicted by the LSTM model are computed and compared. Nevertheless, the study was limited to return prediction and did not consider asset pre-selection. Hence, to fill this gap, this study attempts to select high-quality assets from the Indian stock market through deep learning-based forecasting and build a competitive portfolio for improved returns.

The purpose of this paper is to construct an optimized portfolio based on the asset returns forecasted by deep learning. The authors argue that a complete portfolio consists of two stages, where the first stage is the pre-selection of high-quality assets, and the second stage is the determination of optimal weights for the portfolio. In the empirical part, this study selects the constituent stocks of NIFTY50 to test the proposed methodology. In the primary stage, this paper proposes a stacked LSTM for stock price prediction and then calculates returns based on these predicted values. In the second stage, the stocks are ranked based on these returns, and the top N stocks are selected for portfolio formation and optimization. These selections are then compared against a benchmark index to establish superior performance.

This study addresses the following questions.

- Does asset pre-selection before optimal portfolio formation enhance the portfolio performance compared to the traditional approach without pre-selection?
- Does the performance of different-sized portfolios show significant variations?
- What is the optimal number of stocks to be held in a portfolio for maximum return?

The major contributions of this paper are:

This research makes significant theoretical contributions by developing a stacked LSTM model for predictions in the Indian stock market. By demonstrating the model's efficacy in handling the intricacies of the Indian stock market, characterized by high volatility and noise, this study fills a crucial gap in the literature.

Most relevant studies concentrate on the stage of optimal portfolio formation using the mean-variance framework but overlook the pre-selection of stocks, which occurs before the formation of the optimal portfolio. Markowitz advised against putting all "eggs in one basket," emphasizing risk reduction through diversification. However, the mean-variance approach does not tackle the issue of deciding the number of assets to invest in. The application of LSTM for identifying high-potential stocks before constructing portfolios introduces a novel approach to portfolio management.

By accurately predicting the future performance of various stocks, the model helps investors identify those with the highest potential returns and the most favorable risk profiles. This pre selection process ensures that only the most promising stocks are included in the portfolio, improving overall performance. This further helps to determine the optimum number of stocks to be held while creating a portfolio.

Additionally, this study highlights the significant potential for advancing research on deep learning in emerging markets, integrating new academic insights to better understand financial decision-making and time-series forecasting. Consequently, this work enhances the existing literature on the application of deep learning models for emerging markets.

To the best of the authors' knowledge, this is the first study of its kind that combines forecasting theory with portfolio optimization for the Indian stock market index NIFTY50, and it is believed that this paper is making a substantial contribution to the related literature.

The remainder of this article is organized in the following order. Previous literature is discussed in Section II and Methodology is in Section III. The experimental results and discussion are in Section IV. In Section V, the authors mark their concluding thoughts, and Section VI discusses limitations and future scope.

II. Literature Review

This literature review of this study is divided into two sections. The first section discusses the concept of deep learning and the second section deals with portfolio optimization. This is followed by the research gap.

A. Deep Learning

Deep learning, a subset of machine learning, employs artificial neural networks (ANN) to address complex problems. ANNs are composed of an input layer, hidden layers, and an output layer, with each node in a layer connected to every node in the subsequent layer. By adding more hidden layers, the network becomes deeper, leading to the formation of deep neural networks (DNNs). The term "deep" in deep learning refers to this increased complexity, achieved by adding more hidden layers between the input and output layers.

They have been used for financial forecasting, including economic recession predictions [11], portfolio optimization [2],[12],[13], sentiment analysis [14],[15] and much more. Several studies have investigated the use of deep learning models like LSTM, CNN, RNN and their variations for predicting price movements and trends in financial markets [16]. These studies have shown promising outcomes, demonstrating that deep learning techniques can effectively capture and extract important features and patterns from trading data. Dixon et al. [17] employed deep neural networks to predict daily market movement directions and validated their model through backtesting with a basic trading strategy. A CNN was used by [18] to forecast the hourly direction of BIST 100 stocks, utilizing a "specifically ordered feature set." This research extracted various technical indicators, price data, and temporal features while using chi-square feature selection to minimize noise and dimensionality. Another study in [19] focused on predicting the next-minute direction of the SPDR S&P 500 by proposing slim versions of LSTM models. Their results indicated that Slim LSTM when combined with technical indicators, outperformed the standard LSTM model. Moreover, [20] examined the predictability of intraday movements over different time intervals, analyzing their model's performance during high-volatility periods. The findings revealed the presence of "time-delayed correlations" in S&P 500 stocks in both stable and volatile market conditions and provided evidence supporting the effectiveness of deep learning methods in forecasting trends across numerous interrelated time series. In study [21], several ML and DL-based techniques, such as MNB classifier, SVM, Naïve Bayes, linear regression, and LSTM, were effectively created and executed. These methods utilized the general public's sentiment, viewpoints, news, and past stock prices to predict BSE and Infosys stock prices. Shah et al. [22] developed a model for predicting the NIFTY closing values by combining a CNN, LSTM, and dense layers within a 20-day time frame. Ghosh et al. [23] utilized both Random Forests and LSTM networks to assess their performance in predicting out-of-sample directional movements of S&P 500 constituent stocks for intraday trading. Their analysis spanned from January 1993 to December 2018. More recently, [24] successfully predicted daily stock movements of the BIST 30 index using an ensemble learning algorithm combined with a set of ten different variable groups.

1) LSTM: Despite the existence of numerous deep learning techniques, considerable efforts have been dedicated to demonstrating that LSTM can outperform other methods in the prediction of time-series data. Li et al. [25] conducted a systematic review investigating the application of deep learning models in predicting stock market trends using technical analysis. The review found that the LSTM model was the most commonly used and preferred algorithm for stock market prediction due to its ability to store memory and address the gradient vanishing problem. In their research, [26] utilized ten stock technical indicators along with ten years of historical data from the Tehran stock exchange across multiple machine learning and deep learning models. The study showed that the LSTM model outperformed others in terms of prediction accuracy, displaying the lowest error rate and the highest capacity to fit the data effectively. Abbasimehr et al. [27] proposed a technique of a multi-layer LSTM network with a grid search method. The suggested approach looks for the LSTM network's ideal hyperparameters. The authors used actual demand data from a furniture company to test the efficacy of their suggested strategy and compared it to other cutting-edge time series forecasting methods. The researchers concluded that the proposed model outperformed the alternatives significantly and can be applied to real-world scenarios, like stock price prediction, weather forecasting, and energy demand forecasting. Furthermore, LSTM neural networks have emerged as leading models for a diverse range of machine learning tasks, varying greatly in scale and characteristics. The core concept underlying LSTM architecture is the presence of a memory cell capable of retaining its state over time, complemented by non-linear gating units that control the flow of information into and out of the cell [28]. Consequently, the existing literature inspires us to adopt LSTM for this study because of its ability to analyze relationships among time-series data through its memory function.

B. Portfolio Optimization

Portfolio formation involves two primary concerns: choosing assets with the potential for higher returns and determining the optimal asset composition to achieve the goal of maximizing returns while minimizing risk. A quantitative approach is often employed in making these investment decisions. During the portfolio optimization process, the expected return on an asset is a vital consideration, highlighting the importance of preliminary asset selection in effective portfolio management [6]. Selecting the appropriate asset allocations for a portfolio is challenging because financial market conditions can fluctuate significantly over time.

1) Return prediction and asset pre-selection: Consequently, integrating stock return predictions with portfolio optimization models is essential for effective financial investment [29]. Scholars often use predicted returns instead of historical averages to enhance portfolio optimization models [30], [31].

Huang [32] introduced a model for stock selection combining SVR with genetic algorithms. In this model, SVR was used to forecast the future returns of individual stocks, while the genetic algorithm optimized the model's parameters and input features. The highest-ranked stocks were then equally weighted to construct a portfolio. The findings demonstrated that this proposed model outperformed the benchmarks in investment performance. Hao et al. [31] used an Auto Regressive-Multi Resolution Neural Network (AR-MRNN) and SVM for return prediction, and then prediction-based portfolio selection models were developed using these methods. Comparing the prediction accuracy, the SVM predictor outperforms the AR-MRNN predictor. Additionally, the SVM-based portfolio selection model surpassed the AR-MRNN-based and mean-variance models in performance. The analysis also showed that higher prediction accuracy leads to better returns. Performance comparison of an RNN, GRU, and LSTM for predicting stock prices was conducted by [33]. Their experimental results indicated that the LSTM neural network outperformed the other models. Additionally, they developed portfolios based on predictive thresholds using the LSTM neural network's forecasts. This approach was more data-driven compared to traditional models in portfolio design. Experimental results revealed that these portfolios achieved promising returns.

Wang et al. [3] proposed a mixed method of LSTM and mean-variance model for creating an optimal portfolio. The LSTM was used for return prediction, and then the stocks with the highest predicted returns were selected for portfolio formation. The effectiveness of this methodology is validated by comparing it with five baseline strategies. The proposed model significantly outperforms these strategies in terms of annual cumulative return, Sharpe ratio per three-year period, and average monthly return relative to risk over each three-year period, demonstrating superior potential returns and risk management. Ma et al. [34] integrated return prediction into portfolio formation by utilizing two machine learning models—Random Forest and SVR—along with three deep learning models: LSTM, Deep Multilayer Perceptron (DMLP), and CNN. Specifically, it first applies these prediction models for stock preselection prior to portfolio formation. The predictive results are then used to enhance the mean-variance and omega portfolio optimization models.

2) Problems of index investing: The introduction of index mutual funds in the 1970s, followed by the rapid growth of exchange-traded funds (ETFs) in the 2000s, made it cheaper for ordinary investors to own well-diversified portfolios. This development had two significant consequences. First, many investors who previously held individual stocks switched to passive indexing to reduce transaction and asset management expenses. Second, the affordability of index funds allowed numerous households who had not previously invested in stocks to enter the equity market [35]. But in spite of this, index investing often yields lower returns compared to actively managed portfolios or strategic asset pre-selection methods. This is primarily because index funds aim to replicate the performance of a market index rather than outperform it. Consequently, they are limited by the underlying index's returns, which may not capture high-growth opportunities or effectively manage risks through selective asset allocation. As a result, investors seeking higher returns and better risk-adjusted performance may find more success with approaches that involve active management and careful pre-selection of high-potential assets. However, actively managed funds have a very high expense ratio, which makes them non-feasible and less attractive. Accordingly, there is growing research on combining forecasting theory with portfolio optimization.

C. Research Gap

The effectiveness of portfolio construction largely hinges on the anticipated performance of stock markets. Traditional portfolio theory often relies heavily on expected returns and neglects future information [4]. Advances in machine and deep learning have introduced substantial opportunities to integrate predictive analytics into portfolio selection. Despite this potential, the concept of prediction-based portfolios has been underexplored in academic research, with notable contributions like the one by [30] standing out. Consequently, integrating deep learning predictions to assist in selecting the best investment strategies represents a valuable and promising avenue for future research [36]**.** Many researchers [31], [32], [37], [38], [39] have applied these models in the stock pre-selection process prior to portfolio formation and achieved promising and satisfying results. However, to the best of the authors' knowledge, there are no studies existing with regard to the Indian stock market. We extend the work by [9] where GA was used to select stock portfolios in six different Asian markets, excluding India. This is the first study of its kind, and consequently, it is believed that this paper is making a substantial contribution to the related literature.

III. Methodology

The following sections provide details on the methodology of this study. The study is undertaken in two stages.

A. Prediction model for NIFTY 50

In the first stage, this study aims to build an LSTM-based prediction model for the NIFTY 50 stocks.

1) Data description: Financial time-series forecasting is always explained by historical values or lagged observations [40]. Hence, this dataset consists of historical values of NIFTY stocks spanning 12 years taken from the official NSE website[1]. The study opts for the NSE over the BSE because of its larger size and greater market participation. The Nifty 50 is the primary index of the National Stock Exchange, encompasses 50 diversified stocks across 13 sectors of the economy, and represents the country's leading blue-chip companies. The selected sample includes liquid stocks from various sectors and sizes, thus minimizing sample bias and avoiding concentration on a specific group of stocks.

The values included six features- Open, High, Low, Close (OHLC), adjusted close, and volume. Kumar et al. [41] conducted a survey on stock market forecasting using computational techniques and reported that only 8% of the studies used a combination of historical values and technical indicators for prediction purposes. Hence, this study uses a synthesis of historical data and STIs as the predictor or input variables. The output or target variables are the close values of subsequent days. The total dataset consists of daily trading data of 2,956 trading days (April 2012 – March 2024). This data covers two stock market crashes, the crypto crash in 2018 and the 2020 COVID crisis, and the Russia-Ukraine war in 2022 so that the extreme volatilities of the assets can be considered for optimal portfolio construction. This research chooses a sliding window approach of a 30-day time frame [42], [43] for developing the prediction model. This implies that data from the preceding 30 days will be utilized to forecast close values for the 31st day. Pandas library is used to import data. Matplotlib and Seaborn are used for data visualizations.

2) Data pre-processing: The accuracy of predictions significantly depends on the quality of the data. Therefore, it is vital to preprocess the raw data before incorporating it into the model-building process. The collected sample of 2956 trading days was removed from duplicates and NAN values. The cleaned dataset consisted of 2906 trading days. The outliers identified using a boxplot are treated using the winsorization technique [44], [45], [46]. Winsorization helps to eliminate outliers by capping extreme values, thereby making the distribution of the transformed data more symmetrical and closer to a normal distribution [45], [47]. This data is then normalized using a min-max scalar. Data normalization involves transforming real numerical attributes to a scale between 0 and 1, resulting in a training model that is less

[1] https://www.nseindia.com/reports-indices-historical-index-data

affected by the variable scales [26]. This also ensures that all values fall within a range of [0,1], thus leading to faster convergence. Normalization is very useful for improving the accuracy of neural network models [43]. The equation for normalization is as follows.

$$X\_scaled = \frac{\mathrm{X-XMin}}{XMax-XMin} \tag{1}$$

X is the feature's initial value, *Xmin* is the lowest X value, *Xmax* is the highest X value, and *X_scaled* is the new scaled X value between 0 and 1.

3) *Proposed model*

a) STI: Statistical Technical Indicators (STIs) are mathematical calculations based on factors like price, volume, or other relevant metrics related to stocks, securities, or contracts. Unlike fundamental analysis, they do not take into account business fundamentals such as earnings, revenue, or profit margins. The primary goal of technical analysis is to predict future price movements, and deep learning algorithms enhance the accuracy of these predictions. By combining these two approaches, the reliability of price forecasts can be significantly strengthened. While technical indicators are essential for identifying stock price patterns, trends, and momentum, it's important to note that many studies limit their use to trend indicators, often overlooking key momentum, volatility, and strength indicators that are equally crucial for comprehensive financial analysis [22]. Hence, this study uses a total of ten momentum and volatility STIs as identified by [46] and is calculated through Ta-Lib. They are listed in Table I.

b) LSTM: Hochreiter et al. [51] introduced LSTM in 1997 to address the issue of vanishing gradients in conventional RNNs, which hindered their ability to capture long-term relationships in sequential data effectively. This problem occurs because gradients tend to become smaller and smaller as they propagate back through time, making it difficult for the network to update the weights in earlier layers. The main advantage of LSTMs over traditional RNNs is their capability to choose whether to remember or forget information from earlier time steps, enabling them to handle long-term dependencies more effectively. This is achieved through a set of specialized memory cells and gating mechanisms that balance information flow through the network. "An LSTM layer is composed of one or more LSTM units, and an LSTM unit consists of cells and gates to perform classification and prediction based on time series data" [49]. The cell contains three gating mechanisms: the input gate *i*, the output gate *o*, and the forget gate *f*. The quantity of new information added to the cell state is dictated by the input gate, the amount of old information that is discarded from the cell state is regulated by the forget gate, and the quantity of information that is transferred from the cell state to the next time step is controlled by the output gate [26]. The cell state is the memory of the LSTM and can be thought of as a conveyor belt that runs through the entire LSTM chain, enabling the transmission of information from one time step to the subsequent one.

This study uses a double-layered LSTM since deeper LSTM architecture is known to yield superior prediction outcomes compared to a single LSTM network [50]. Furthermore, since the input variables are 18 in total, a PCA is conducted for dimensionality reduction.

c) Working of the stacked LSTM model: The input data for the LSTM model is organized into a three-dimensional array, where each dimension captures a different aspect of the data. The time dimension corresponds to the sliding time window, which is used to capture temporal dependencies in the data. The sample dimension represents the size of the dataset used for training and testing the model. Finally, the feature dimension indicates the number of input features provided to the LSTM model, allowing it to process multiple attributes simultaneously for more accurate predictions. This study chose 30 days as a time window, and the input features are the OHLC, volume, and STI values reduced as PC's through PCA. The input layer is linked to the LSTM layer with 32 neurons, which is the hidden layer. This is connected to another LSTM with 16 neurons. Two LSTM layers are stacked sequentially, where the output of the first LSTM layer serves as the input to the second LSTM layer. This stacking enables the model to better capture and process sequential patterns and long-term dependencies in the data. The second LSTM layer is then connected to a dense layer with a single neuron, which serves as the output layer for making predictions, such as forecasting stock prices or classifying trends. This setup allows the model to refine complex temporal relationships and generate more accurate results.

Dropouts and Early stopping are used as regularization techniques. The idea behind dropout [51] is to randomly "drop out" or disable some of the neurons in a layer during each training iteration. This is done by setting the output of some of the neurons to zero with a certain probability (usually around 0.5). The exact neurons that are dropped out are randomly selected during each training iteration. The reason for the dropout is to prevent the neural network from depending too much on any particular set of neurons. By randomly dropping out neurons during training, the network is forced to learn more robust and generalized useful features across a wider range of inputs. At test time, the full network is used without any dropout. However, the output of each neuron is multiplied by the dropout probability to ensure that the expected output of each neuron is the same as during training. A dropout rate of 0.2 is used in this case. EarlyStopping is a Keras callback that allows you to stop training when a monitored quantity (like validation loss or accuracy) has stopped improving. It helps avoid overfitting by terminating the training process once the model performance on the validation set no longer improves. This ensures that the model does not waste time training for too many epochs, which can lead to overfitting. Here, it is configured to monitor the validation loss and stop training if accuracy does not improve for 8 consecutive epochs.

The model is initialized with random weights and biases. Each LSTM layer receives an input consisting of the previous 30 time steps and attempts to predict the next time step in the sequence, which corresponds to the closing value for the 31st day; this output from the LSTM is passed on to the dense layer, which gives the final output values. ReLU and Linear activation functions are utilized in the hidden and output layers, respectively. Mean Squared Error (MSE) is used as the loss function. Each LSTM network computes its individual loss, and the total loss is calculated by summing the losses from both LSTM networks and the dense layer. The Adam optimization algorithm is used during training to minimize this total loss, and the number of epochs is set to 100, following the approach used by [2]. Optimizers are techniques utilized to adjust the model's features, including parameters like learning rate and weights, to minimize losses. An epoch represents one complete pass of the entire training dataset [26] by the LSTM model. The training process continues until the validation loss stops improving for a specified number of epochs, as determined by early stopping, or until the maximum number of iterations is reached.

TABLE I. LIST OF STIS

STI	Description	Indicator Type
MACD	Moving average convergence divergence	Momentum
RSI	Relative strength index	Momentum
STOCH	Stochastic Oscillator	Momentum
CCI	Commodity Channel Index	Momentum
ADX	Average Directional Index	Momentum
ROC	Rate of change	Momentum
WILLR	William percent R	Momentum
ATR	Average True Range	Volatility
NATR	Normalized Average True Range	Volatility
TRANGE	True Range	Volatility

Source:[46]

The model’s parameter settings are established through trial-and-error experiments to optimize performance. The implementation is carried out in a Python 3.7 environment using Keras, a high-level API built on Google's TensorFlow framework. Table II provides an overview of the parameter settings used in the experiments.

B. Portfolio Optimization

1) Assumptions: In this study, several key assumptions are necessary for the analysis of portfolio performance. Although these assumptions may seem idealistic compared to real-world conditions, they serve to simplify the complexities associated with investment, such as costs and trading prices. By making these assumptions, the study can focus more effectively on comparing the relative performance of portfolios.

- No transaction cost and tax.
- The stocks are sold and bought at the closing price.
- Investors are not risk-averse.
- A 3-year average return on treasury bills is considered as risk-free return.

Once an LSTM-based prediction model has been developed for predicting close values, the next stage involves the preselection of assets for portfolio creation and optimization. Since the test period consists of approximately three years, datasets from April 2021 – March 2024 (Test period) are considered for portfolio creation as well. Average returns for a period of three years are calculated for all 50 stocks based on the predicted close values. Then, they are arranged in descending order of their predicted returns to select the top return-generating stocks.

2) Top N stocks: Several studies have shown that holding too many stocks makes it difficult to manage and keep them under control, particularly for individual investors. Building a portfolio with fewer than ten stocks is taken into consideration in several portfolio optimization research [52]. A portfolio with an average of seven stocks performs better than other portfolios with a variable number of stocks, according to [37]. Wang et al. [3] noted that a portfolio with ten stocks is ideal, as it outperforms portfolios with any other number of stocks. According to [40], the optimal number of stocks for an individual investor’s portfolio construction is seven. As a result, this study selected a group of the top 10,9,8 and seven stocks for portfolio creation. Six different objectives-based portfolios are created. This means that for each objective, five different portfolios with the number of stocks N =10,9,8 and 7 and, all NIFTY stocks with N=50 are formed to evaluate the performance of different-sized portfolios.

TABLE II. PARAMETER SETTINGS

Parameters	Values
Optimizer	Adam
Epochs	100
Batch size	64
Step size	30
Drop out	0.2
Activation function	ReLU (Hidden layer) Linear (Output layer)

Source:[46]

3) Portfolio objectives

- Objective 1: Minimum volatility portfolio
- Objective 2: Maximum returns portfolio
- Objective 3: Maximum Sharpe ratio with No Constraints
- Objective 4: Maximum Sharpe ratio with Constraints [L2 regularizer, gamma=2, $\sum W = 1$]
- Objective 5: Uncorrelated assets portfolio
- Objective 6: Equally weighted portfolio

4) Comparison criteria: The portfolios created are compared on the following metrics of return and risk [39], [56], [57].

a) Sharpe ratio: The Sharpe ratio is a widely used industry standard for assessing investment risk adjustment return in finance. It is computed by deducting the risk-free investment return from the stock or investment portfolio's actual return and dividing the result by the stock or portfolio's standard deviation.

$$\text{sharpe ratio} = \frac{R_\rho - R_f}{\sigma_P} \qquad (2)$$

where, R_P is the Return of the portfolio, R_f is the Risk-free rate, σ_P is the Standard deviation of the portfolio. A Sharpe ratio >1 is always preferred, implying that the investment generates one unit of excess return for every unit of risk taken.

b) Sortino ratio: Sharpe ratio considers both upside and downside risks since standard deviation calculates deviation from mean returns. This deviation could either be positive or negative. Conversely, the Sortino ratio [55]—a variant of the Sharpe ratio—only accounts for negative or downward volatility. Upside volatility is generally considered a benefit of investing and is not dangerous [55]. Consequently, the total standard deviation in the Sharpe ratio is replaced by this downside risk or volatility in the Sortino ratio. A higher Sortino ratio is always desired by the investor since it indicates the return per unit of downside risk.

$$sortino\ ratio = \frac{R_\rho - R_f}{DR} \quad (3)$$

where, R_P is the Return of the portfolio, R_f is the Risk-free rate and DR is the Downside risk.

c) Cumulative returns: Cumulative returns measure the total growth of an investment from the start to the end of a given period. It represents how much a portfolio has increased in value, assuming daily compounding of returns. The study uses the *cumprod()* function in Python. This cumulative product function calculates the running product of the growth factors or daily growth over the entire period. This effectively compounds the daily returns, simulating how an investment grows day by day.

d) Annual returns (CAGR): Annual returns convert total cumulative growth into an annual growth rate. It accounts for the effect of compounding, showing the consistent annual rate that equates to the total growth observed. The study uses 252 Trading Days as the standard number of trading days annually to annualize the returns.

$$\text{Annual return} = (1+\text{ Cumulative return})^{252/\text{ Trading days}} - 1 \quad (4)$$

Here, the number of trading days is calculated as 252 days* 3 years

e) Volatility: Volatility is a measure of the risk factor of the portfolio and is calculated as the standard deviation of the portfolio's returns by using the covariance matrix of the asset returns and the optimized portfolio weights.

f) Beta: Beta is the measure of the systematic risk of a portfolio or stock. It indicates the relative volatility of a portfolio as against the benchmark or index. It helps investors understand how much risk they are taking in comparison to the market. Investors looking for higher returns with higher risk might prefer high-beta stocks, while those seeking stability might opt for low-beta stocks.

The study used the PyportfolioOpt package in Python to create these portfolios. This package generated weights based on the given objective functions. For every N number of stocks, six different portfolios corresponding to each objective are formed. After the weights have been assigned, portfolio returns are calculated based on these allotted weights and the market returns, which are NIFTY returns in this case. Next, *cum.prod()* function is used to calculate cumulative returns and CAGR is calculated from that value. Volatility is measured as the standard deviation of the returns. The study considered a rate of 6.85% as the average risk-free rate of the past three years and 252 as the average trading days for calculating Sharpe and Sortino ratios.

IV. Experimental Results and Discussion

A. Generation of PC's

In the first experiment, the study attempts to develop a stacked LSTM for the prediction of NIFTY 50 stocks. Historical values and STIs of the past 30 days are used as input variables to predict the close values of the 31st day. Since there are 18 input variables and 12 years of data, as mentioned in the methodology section, the study uses PCA to generate PCs to reduce dimensionality. The PCA run for all the 50 stocks generated 5 or 6 PCs. To evaluate the variance associated with each PC, the explained variance ratio was calculated. This ratio is obtained by dividing the variance of each component by the total variance. The variance ratio of each PC of the top 5 NIFTY stocks, in terms of market capitalization, is explained in Table III. It is clear that more than 95% of the information was preserved even after PCA.

The number of generated PCs determines the architecture of the stacked LSTM. Since the number of PC's are 5 and 6, the input layer will also have the same number of nodes. The architecture of the developed LSTMs is illustrated in Fig. 1.

B. Learning Curves

After obtaining the PCs, the datasets are separated as training data (70%; 2,035 days), validation data (10%;291 days), and testing data (20%; 580 days) to train, validate and test the proposed LSTM.

The study employed training and validation curves to chart the model's performance, indicated by loss (MSE), on both the training and validation datasets across epochs. During training, the model aims to minimize the loss function, which measures the disparity between the predicted values and the actual values in the training dataset. Fig. 2 illustrates the learning curve of the LSTM model applied to the top five stocks, The X-axis represents the number of training epochs, and the Y-axis represents the loss value, indicating the model's error in prediction.

TABLE III. EXPLAINED VARIANCE RATIO

PCs	HDFC	RIL	ICICI	INFY	L&T
0	0.5912883	0.5722852	0.5063156	0.5477141	0.5843656
1	0.1384574	0.1520931	0.1390794	0.1500667	0.1258094
2	0.08379	0.0760954	0.1303963	0.0789791	0.1029366
3	0.0500956	0.0707812	0.0786461	0.0639012	0.0575905
4	0.0427779	0.0428246	0.0574367	0.0537327	0.0469288
5	0.0384285	0.0343057	0.0338025	0.0424565	0.0334498
6	0.0261741	0.0154152	0.0179384	0.0298618	
Total	0.9710117	0.9638004	0.9636151	0.9667119	0.9510807

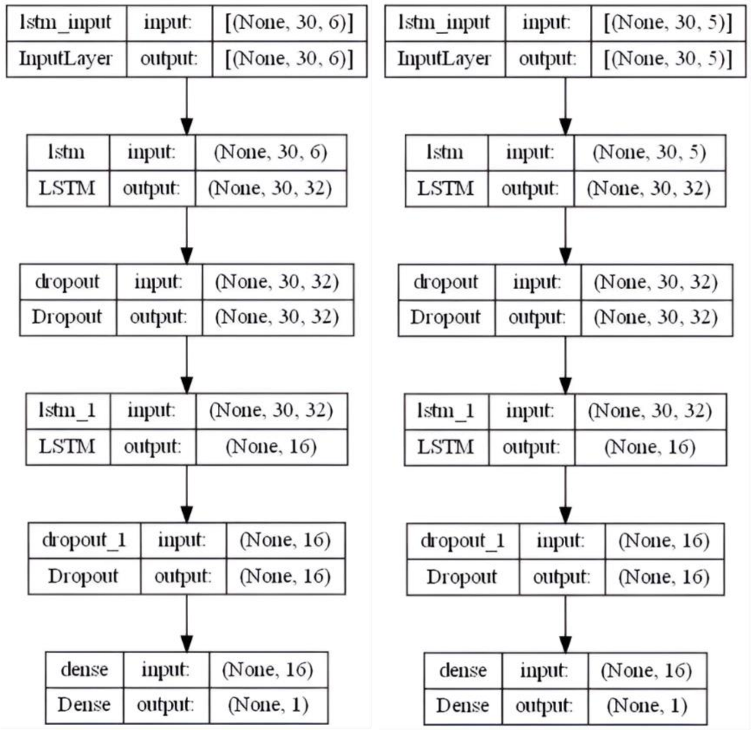
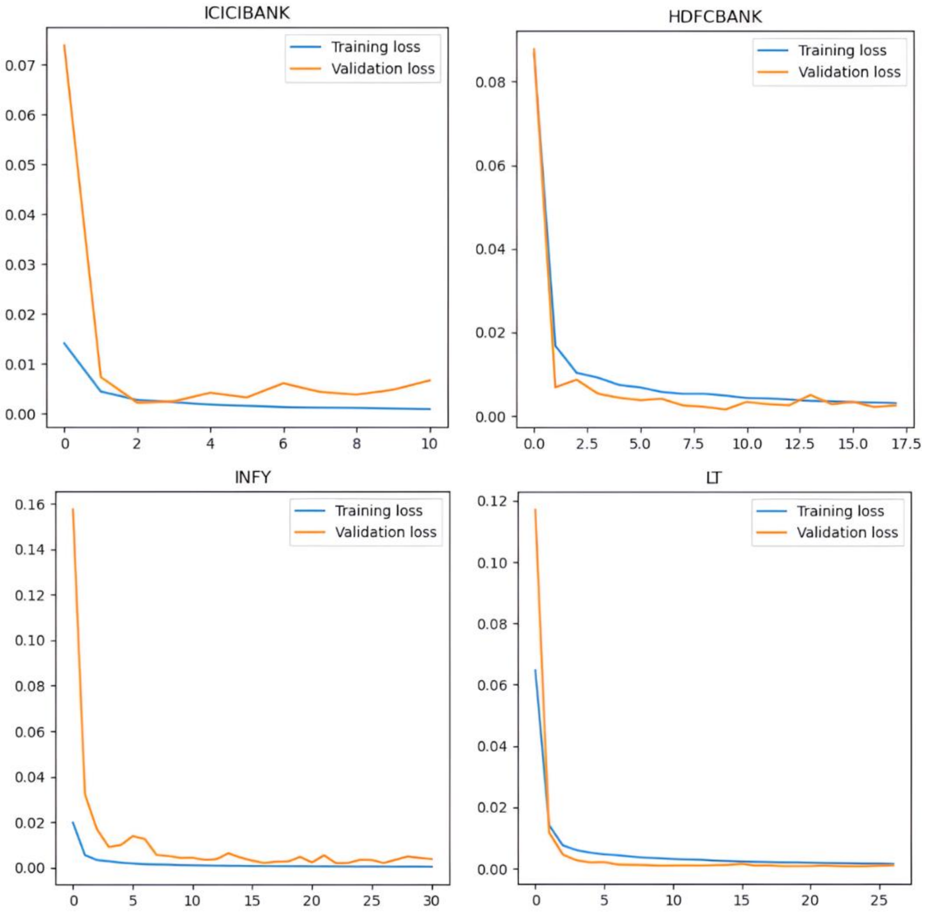

lstm_input | input: | [(None, 30, 6)]
InputLayer | output: | [(None, 30, 6)]
lstm | input: | (None, 30, 6)
LSTM | output: | (None, 30, 32)
dropout | input: | (None, 30, 32)
Dropout | output: | (None, 30, 32)
lstm_1 | input: | (None, 30, 32)
LSTM | output: | (None, 16)
dropout_1 | input: | (None, 16)
Dropout | output: | (None, 16)
dense | input: | (None, 16)
Dense | output: | (None, 1)
lstm_input | input: | [(None, 30, 5)]
InputLayer | output: | [(None, 30, 5)]
lstm | input: | (None, 30, 5)
LSTM | output: | (None, 30, 32)
dropout | input: | (None, 30, 32)
Dropout | output: | (None, 30, 32)
lstm_1 | input: | (None, 30, 32)
LSTM | output: | (None, 16)
dropout_1 | input: | (None, 16)
Dropout | output: | (None, 16)
dense | input: | (None, 16)
Dense | output: | (None, 1)

Fig. 1. Architecture of the LSTM.



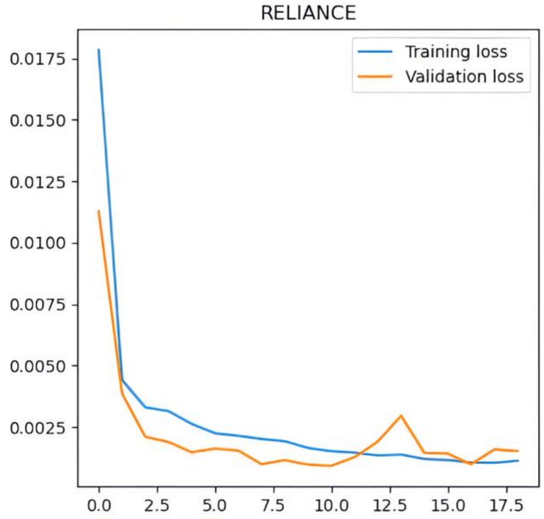


Fig. 2. Learning curves.

The study made the following observations

- Good Generalization: The learning curves for both training and validation loss decrease rapidly at the start and remain closely aligned throughout the training. The parallel behavior of the training and validation losses suggests that the model is generalizing well and not overfitting or underfitting the training data.
- Convergence: Both curves stabilize and converge to a low value towards the end of the epochs, indicating that the model has reached a point where additional training does not significantly change the loss, demonstrating a well-trained model.

C. Metric Evaluation

Appropriate assessment metrics are needed to validate the deep learning models. This study chooses the following indicators: Mean of absolute error (MAE) and root mean square error (RMSE). They have been extensively used in the literature [56], [57], [58]. Lower MAE and RMSE values would increase the prediction accuracy. Accuracy is measured as [58]. The equations are as follows.

$$\text{PM}\Sigma\text{E}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i|^{2}} \quad (5)$$

Where n is the number of data points, yi is the actual value of the i th data point, ŷi is the predicted value of the i th data point, Σ represents summation, | | represents absolute value.

$$\text{A}\chi\chi\upsilon\rho\alpha\chi\psi = 1 - \text{MAE} \quad (6)$$

$$\text{MAE}= \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \quad (7)$$

Results show that the model obtained an average accuracy of 90.58%, RMSE of 0.1057, and MAE of 0.0942. ADANIENT, from the Metals and Mining Industry, and EICHERMOT, belonging to the Automobile industry, recorded the lowest and highest accuracies of 53.12% and 97.33%, respectively. Industry-wise analysis shows that the Oil, Gas & Consumable Fuels industry has the highest and Metals and Mining has the lowest accuracy.

D. Asset Preselection

In the second experiment, the focus is on creating and optimizing stock portfolios. The process involves filtering the top stocks based on their average returns over a three-year test period. The returns are calculated on the basis of their adjusted closing values. Table IV shows that the top 10 stocks include COAL INDIA, which achieved the highest average return at 38.72%, followed by SUN PHARMA with 25.22%, and ONGC with 23.83%.

Interestingly,30% of these stocks are in the energy sector (COALINDIA, BPCL, ONGC) and another 30% in the automobile sector (HEROMOTOCO, EICHERMOT, BAJAJ-AUTO), with 20% in banking and financial services (INDUSINDBK, AXIS BANK) and 10% each in pharmaceuticals (SUN PHARMA) and FMCG(ITC), despite financial services (37.72%) and information technology (14.11%) constituting the majority share in NIFTY50.

E. Portfolio Formation

This study created portfolios with the number of stocks N= 10, 9, 8 and 7. With each N, six objectives-based portfolios are created. For optimization purposes, the PyportfolioOpt package is used to achieve optimum weight allocation in each scenario. These portfolios are compared on the basis of the six key factors: Sharpe ratio, Sortino ratio, Annual returns, Cumulative returns, Annual volatility, and Beta. Furthermore, optimal portfolios are also compared against the market benchmark NIFTY 50 and portfolio with N=50 to evaluate the performance of portfolios with and without asset preselection.

1) Performance of different-Sized portfolios: For Objective 1, the top-performing portfolios are analyzed with a focus on their respective metrics. The top 7 portfolios stand out with a Sharpe ratio of 1.75 and a Sortino ratio of 3.15, indicating high risk-adjusted returns and effective downside risk management. These portfolios achieve an annual return of 29.67% and a cumulative return of 114.27%.

TABLE IV. AVERAGE RETURNS OF TOP 10 STOCKS

Stock	Annual returns
COALINDIA	38.7258
SUN PHARMA	25.2218
ONGC	23.8309
INDUSINDBK	22.3735
AXIS BANK	22.164
ITC	21.5993
HEROMOTOCO	19.5488
EICHERMOT	17.0569
BAJAJ-AUTO	16.6815
BPCL	15.8402

TABLE V. OBJECTIVE-WISE ANALYSIS

OBJECTIVE 1

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns %	Cumulative Returns
Top 7	1.75	3.15	0.75	14	29.67	114.27
Top 8	1.72	3.13	0.76	13.9	29.02	111.13
Top 9	1.93	3.41	0.74	13.6	31.41	122.78
Top 10	1.87	3.33	0.74	13.5	30.53	118.46
All 50	1.46	3.04	0.63	10.26	20.62	73.32

OBJECTIVE 2

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns	Cumulative Returns
Top 7	1.63	3.35	0.82	20.51	46.43	205.99
Top 8	1.65	3.40	0.80	17.79	41.48	176.69
Top 9	1.51	3.14	0.91	19.58	41.32	175.76
Top 10	1.44	3.06	0.90	19.41	38.83	162.01
All 50	1.15	2.52	0.99	14.57	25.29	93.69

OBJECTIVE 3

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns	Cumulative Returns
Top 7	2.35	3.93	0.72	17.2	44.51	194.41
Top 8	2.35	3.93	0.72	17.2	44.51	194.41
Top 9	2.51	4.17	0.72	15.7	44.03	191.57
Top 10	2.51	4.17	0.72	15.7	44.03	191.57
All 50	2.73	4.47	0.73	14.3	43.85	190.49

OBJECTIVE 4

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns	Cumulative Returns
Top 7	2.14	3.41	0.81	18	42.06	179.99
Top 8	2.11	3.39	0.81	17.7	40.85	173.06
Top 9	2.24	3.61	0.79	16.4	40.9	173.31
Top 10	2.2	3.55	0.80	16.3	39.86	167.48
All 50	2.08	3.32	0.94	15.5	36.62	149.69

OBJECTIVE 5

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns	Cumulative Returns
Top 7	1.66	3.6	0.67	15.83	37.47	154.32
Top 8	1.6	3.44	0.71	15.29	35.1	141.65
Top 9	1.89	3.91	0.72	15.78	42.47	182.37
Top 10	1.73	3.55	0.80	16.09	34.6	165.55
All 50	1.19	2.64	0.92	13.88	25.07	92.72

OBJECTIVE 6

	Sharpe ratio	Sortino ratio	Beta	Annual Volatility	Annual Returns	Cumulative Returns
Top 7	1.78	2.83	0.89	15.92	31.06	121.08
Top 8	1.7	2.76	0.90	15.7	28.98	110.96
Top 9	1.8	2.94	0.88	15.27	30.11	116.39
Top 10	1.74	2.79	0.88	15.18	28.75	109.82
All 50	1.44	2.23	0.95	13.8	20.77	73.91

Expanding the portfolio to the top 9 increases the Sharpe ratio to 1.93 and the Sortino ratio to 3.41, with annual returns rising to 31.41% and cumulative returns to 122.78%, suggesting that including more assets up to this point optimizes performance. However, the top 10 portfolios show a slight decrease in the Sharpe ratio to 1.87 and Sortino ratio to 3.33, with annual returns at 30.53% and cumulative returns at 118.46%, indicating a slight reduction in performance efficiency. The beta values remain relatively low across these portfolios, suggesting they maintain a lower level of market risk while delivering strong returns.

Under Objective 2, the top 7 portfolios have a Sharpe ratio of 0.63 and an impressive Sortino ratio of 3.35, indicating excellent management of downside risk. These portfolios achieve the highest annual returns of 46.43% and cumulative returns of 205.99%. The top 8 portfolios, with a higher Sharpe ratio of 1.65 and Sortino ratio of 3.40, offer annual returns of 41.48% and cumulative returns of 176.69%, providing better overall risk-adjusted performance. The top 9 portfolios show a slight decrease in the Sharpe ratio to 1.51 and Sortino ratio to 3.14, with annual returns at 41.32% and cumulative returns at 175.76%, suggesting a minor decline in performance. The top 10 portfolios further decrease in performance with a Sharpe ratio of 1.44 and Sortino ratio of 3.06, achieving annual returns of 38.83% and cumulative returns of 162.01%. The beta values indicate that these top portfolios are slightly more volatile, which aligns with their higher returns.

For Objective 3, the portfolios excel in risk-adjusted returns and overall performance. The top 7 and top 8 portfolios share the highest Sharpe and Sortino ratios of 2.35 and 3.93, respectively, achieving annual returns of 44.51% and cumulative returns of 194.41%. Interestingly, the top 9 and top 10 portfolios, with slightly higher Sharpe ratios of 2.51 and Sortino ratios of 4.17, deliver similar annual returns of 44.03% and cumulative returns of 191.57%, suggesting that an increase in the number of assets does not significantly impact performance. The beta values are slightly higher, reflecting moderate volatility but efficient risk management.

In Objective 4, the top 7 portfolios exhibit a high Sharpe ratio of 2.14 and a Sortino ratio of 3.41, with annual returns of 42.06% and cumulative returns of 179.99%. The top 8 portfolios maintain a Sharpe ratio of 2.11 and a Sortino ratio of 3.39, with annual returns of 40.85% and cumulative returns of 173.06%. Expanding to the top 9 portfolios slightly increases the Sharpe and Sortino ratios to 2.24 and 3.61, respectively, while maintaining annual returns of 40.9% and cumulative returns of 173.31%. The top 10 portfolios show a slight decrease in performance with a Sharpe ratio of 2.2 and Sortino ratio of 3.55, achieving annual returns of 39.86% and cumulative returns of 167.48%. The beta values indicate that these portfolios are more volatile but compensate with higher returns, reflecting efficient risk management.

Objective 5 portfolios show significant variation in performance metrics. The top 9 portfolios stand out with a higher Sharpe ratio of 1.89 and a Sortino ratio of 3.91, achieving the highest annual returns of 42.47% and cumulative returns of 182.37%. The top 7 portfolios have a Sharpe ratio of 1.66 and a high Sortino ratio of 3.6, with annual returns of 37.47% and cumulative returns of 154.32%. The top 8 portfolios maintain a Sharpe ratio of 1.6 and Sortino ratio of 3.44, achieving annual returns of 35.1% and cumulative returns of 141.65%. The top 10 portfolios show a slight decline in performance with a Sharpe ratio of 1.73 and Sortino ratio of 3.55, achieving annual returns of 34.6% and cumulative returns of 165.55%. The beta values suggest moderate volatility, consistent with the achieved returns.

For Objective 6, the performance metrics highlight a steady trend. The top 7 portfolios achieve a Sharpe ratio of 1.78 and a Sortino ratio of 2.83, with annual returns of 31.06% and cumulative returns of 121.08%. The top 8 portfolios maintain a Sharpe ratio of 1.7 and Sortino ratio of 2.76, achieving annual returns of 28.98% and cumulative returns of 110.96%. The top 9 portfolios show a slight increase in the Sharpe ratio to 1.8 and Sortino ratio to 2.94, with annual returns of 30.11% and cumulative returns of 116.39%. The top 10 portfolios exhibit a Sharpe ratio of 1.74 and Sortino ratio of 2.79, achieving annual returns of 28.75% and cumulative returns of 109.82%. The beta values indicate that these portfolios are relatively more volatile, reflecting their higher returns. Overall, the top portfolios under Objective 6 maintain good performance with balanced risk management despite diminishing returns with larger portfolio sizes. Fig. 3 shows a performance comparison of the best portfolios from each objective.

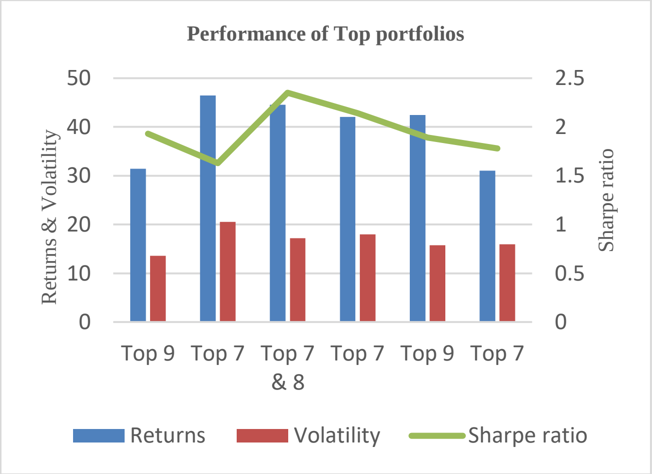


Fig. 3. Performance comparison of Top portfolios.

2) Best performing portfolio:'Investors in the financial markets rarely pursue a risk-minimization approach; instead, they are more than willing to take on more considerable risks if the accompanying profits are even higher' [59]. Consequently, the study considers returns as the primary criteria for selecting the best portfolio. Though the top 7 stocks under objective 2 (maximum returns) generated the highest return of 46.43%, the volatility is 20.51%, which is higher than any other portfolio. Therefore, based on the detailed analysis across various objectives, the portfolio that consistently delivered the best performance is the top 9 and 10 portfolios under Objective 3, which is the maximum Sharpe ratio with no constraints. Though nine stocks were initially added, the weights were allocated to only 5 stocks [Bajaj Auto, Coal India, ITC, ONGC, and Sun pharma] and the rest were given zero weights. Weights were allocated in such a way that it maximizes the objective function of the Sharpe ratio. Hence, effectively only five stocks constituted the optimal portfolio.

Fig. 4 demonstrates the weight allocation for portfolio construction. It is also interesting to note that these 5 selected stocks are not the top 5 return-generating stocks but are randomly selected by the software.

Portfolio performance metrics:

- Sharpe Ratio: 2.51
- Sortino Ratio: 4.17
- Beta: 0.72
- Annual Volatility: 15.7%
- Annual Returns: 44.03%
- Cumulative Returns: 191.57%

The high ratios indicate that this portfolio achieved the best risk-adjusted returns and managed downside risk effectively.

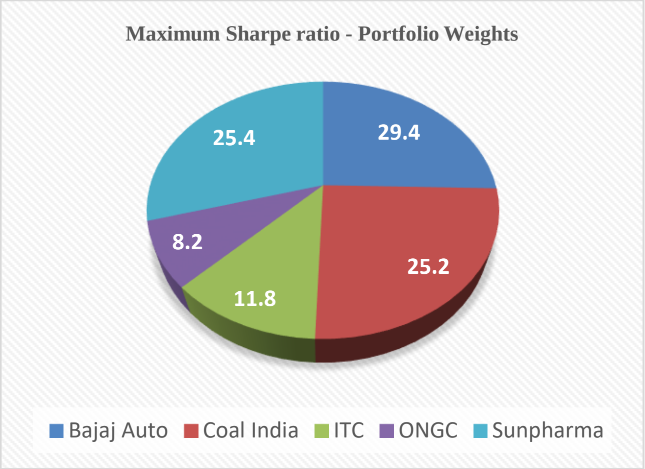


Fig. 4. Weight allocation of optimal portfolio.

The beta value indicates a moderate level of market risk, suggesting that the portfolio is not overly volatile while still capturing substantial returns. Annual Returns of 44.03% and Cumulative Returns of 191.57% reflect the highest annual and cumulative returns, demonstrating the portfolio's superior performance over the period. Hence, the study concludes that a five-stock portfolio provides the best performance across all analyzed metrics. They achieve the highest risk-adjusted returns, manage downside risk effectively, and deliver the highest annual and cumulative returns.

3) Pre-selection V/s All 50: Table V shows a comparison of portfolios constructed after pre-selection and without pre-selection. The pre-selected stocks consist of top 7,8,9, and 10 stocks from NIFTY 50, whereas the alternate portfolios consist of all NIFTY 50 stocks. It is evident that the former achieved better returns than the latter in terms of annual returns. Fig. 5 shows the excess returns earned by the top-performing portfolios as compared to NIFTY all 50 stocks from each objective. The optimal portfolios obtained an excess return of 10.79%, 21.14%, 0.66%,5.44 %, 17.4 %, and 10.29 %. For Objective 3, the portfolio aimed to maximize the Sharpe ratio without constraints but only achieved a 0.66% outperformance in annual returns. This underperformance suggests that the unconstrained approach might have led to a skewed portfolio, possibly concentrating too heavily on high-risk stocks. While the goal was to achieve the best risk-adjusted returns, the lack of risk controls likely reduced the portfolio's ability to generate higher excess returns. In contrast, Objective 4, with constraints, achieved a much higher excess return of 17.4%, highlighting the importance of controlled risk management when optimizing for the Sharpe ratio. Nevertheless, the top portfolios earned an average excess return of 10.95% as against NIFTY all 50 portfolios. Thus, the experiment proves that pre-selection of your assets can help better your investment fortunes than too much diversification. Over-diversification reduces your risk but also brings down one's returns.

4) Superiority over benchmark models: The study evaluated the performance of the proposed portfolio optimization method against the NIFTY 50 index, a well-established market benchmark. This benchmark includes a diverse range of leading companies in the Indian stock market. By using the NIFTY 50 as a reference, the study can compare the proposed portfolio allocations to the performance of a globally recognized benchmark index.

a) All 50 v/s Index: This analysis involves the creation of six different portfolios, each based on a distinct objective, utilizing all NIFTY 50 stocks. These portfolios were then compared to the NIFTY Index in terms of returns and volatility. The results demonstrate that the newly constructed portfolios significantly outperformed the NIFTY Index (Fig. 6). Specifically, the portfolios showed higher returns, with "All 50 (3)" achieving the highest return of 43.85% and "All 50 (4)" following with 36.62%. Even the lowest-performing portfolio, "All 50 (1)," delivered a return of 20.62%, surpassing the NIFTY Index's return of 14.85%. In terms of volatility, while some portfolios exhibited higher volatility than the NIFTY Index (13.75%), such as "All 50 (4)" with 15.5% and "All 50 (2)" with 14.57%, others managed to maintain lower or comparable volatility levels, like "All 50 (1)" with 10.26%. This indicates that the newly created portfolios not only provided superior returns but also effectively managed risk, outperforming the NIFTY Index overall.

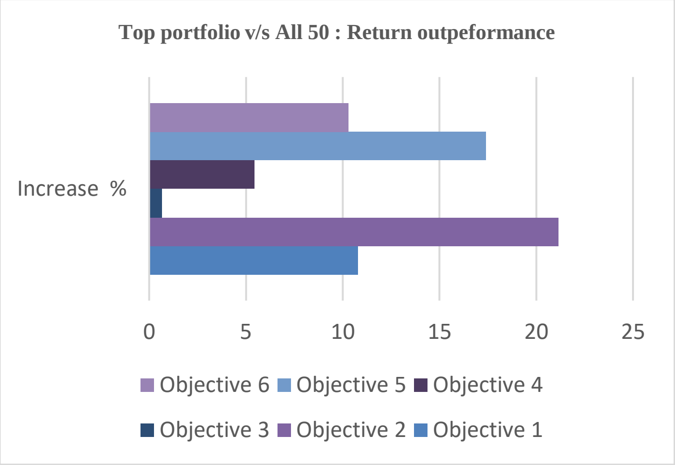


Fig. 5. Return outperformance of top portfolios.

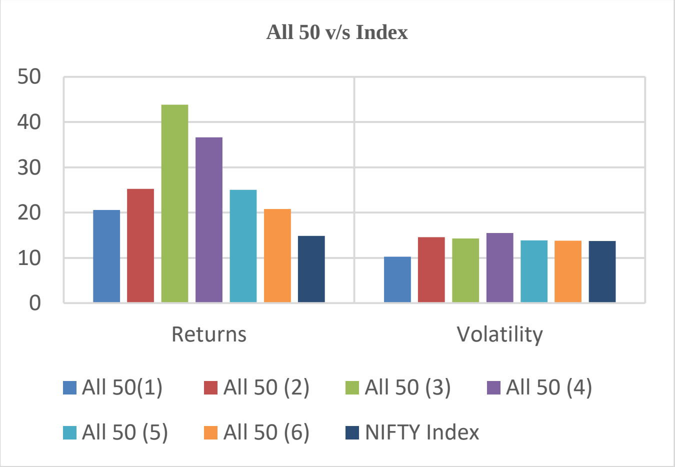


Fig. 6. Comparison of returns and volatility.

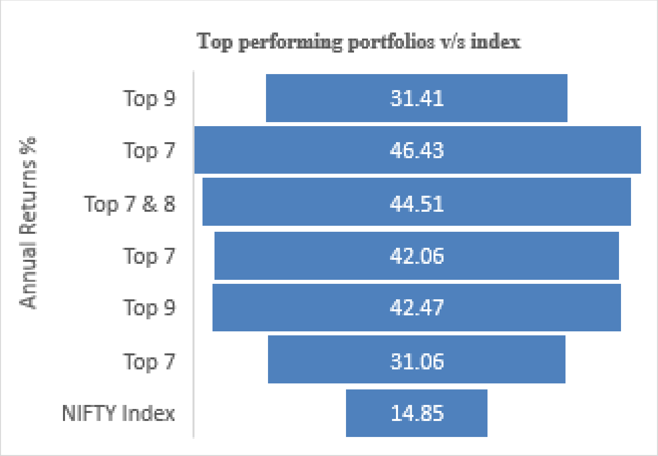


Fig. 7. Performance comparison of top portfolios and NIFTY index.

b) Top performing portfolios v/s index: This analysis compared the annual returns of top-performing portfolios from each objective with the NIFTY index returns over the past three years. Fig. 7 shows that the top 7 portfolios in objective 2 could generate returns as high as 46%, an excess return of 31.58% more than the NIFTY index. Even the portfolio with the lowest returns of 29.02% (Top 8 in objective 1) earns 14.7 % more than the benchmark index, demonstrating consistent and notable outperformance across different strategies. All the top-performing portfolios earn an average excess return of 27.51%. The analysis conclusively demonstrates that the top-performing portfolios provide superior returns and significantly exceed the NIFTY index's benchmark performance. This suggests that the strategies employed in these portfolios are highly effective, yielding substantial excess returns even in the least performing portfolio among the top contenders. Investors may find these strategies attractive for achieving higher returns compared to the standard benchmark, emphasizing the value of strategic portfolio selection and management.

F. Discussion

This work aims to extend the existing literature on deep learning-based prediction and asset pre-selection for portfolio optimization. An LSTM-based prediction model was developed using data from 12 years of historical and technical indicators. This model was applied to forecast NIFTY stocks for a test period of 580 days. Then, the predicted returns were used to filter the top ten assets for portfolio creation. The major findings of the study are:

- Results show that the model obtained an average accuracy of 90.58%, RMSE of 10.57%, and MAE of 9.42%. The authors compare these results to [62],[63], which reported an accuracy of 72% and 90%, respectively.

- This study attempted to select high-quality assets from the Indian stock market through deep learning-based forecasting and build a competitive portfolio for improved returns. Results indicated that portfolios coupled with pre-selected assets generated better results than the portfolios with the entire NIFTY 50 stocks. Preselection helps filter out underperforming or overly volatile assets, leading to a more robust and resilient portfolio that aligns with specific investment objectives. This is consistent with the studies of [63],[64], and [65].

- It is evident from the study that a portfolio consisting of 5 stocks provides the optimal balance between diversification, risk management, and return maximization, which is consistent with the results of [62] but contradicting [3], [37] and [40]. While diversification is crucial to reduce unsystematic risk, excessive diversification beyond nine stocks leads to diminishing returns and unnecessary complexity. For instance, the top 10 and all 50 portfolios have significantly lower returns and Sharpe ratios compared to the top 7, 8 and 9 portfolios.

V. Conclusion

With advancements in machine learning and deep learning, though predicting asset returns has become feasible, these prediction results are not yet effectively utilized in practice for portfolio creation and optimization. As a result, many portfolios fail to fully capitalize on the available predictive insights, limiting their potential for improved performance and risk management. The challenge lies in effectively utilizing these predicted returns to construct an optimal investment portfolio. Considering this, this research seeks to tackle the issue by exploring how to integrate advanced forecasting information into the portfolio selection process.

The study has dual stages. In the first stage, the study developed a stacked LSTM capable of forecasting close values of all NIFTY 50 stocks by following a sliding window approach of 30 days. The model obtained an average accuracy of 90%. The second stage is asset pre-selection, where the top ten stocks, based on their predicted returns, were filtered for portfolio creation. Five portfolios each per objectives were created resulting in a total of 30 different portfolios. The results concluded that portfolios constituting five stocks result in best returns as high as 44%. Investors should avoid expanding their portfolios beyond nine stocks, as excessive diversification can lead to diminishing returns and unnecessary complexity. The proposed portfolios beat the benchmark NIFTY index as well as portfolios with no asset pre-selection, comprising all 50 stocks.

The findings of this study indicate that the proposed two-stage portfolio optimization method has the potential to construct a promising investment strategy due to its trade-off between historical and future information on assets. The results demonstrate the reliability and effectiveness of the asset selection approach in identifying high-performing assets, providing competitive risk-adjusted returns for portfolio optimization, beneficial for both portfolio managers and individual investors. Using real-time market predictions, the algorithm enables investors to choose assets with higher returns and apply the model, which accounts for recent data dynamics in expected return and risk. This makes the approach more practical. Consequently, the proposed method offers a systematic decision-making framework that assists in determining which assets to hold and their investment proportions to achieve the maximum risk-adjusted return and optimal risk-return balance. The study hence concludes that combining forecasting theory with portfolio selection could improve portfolio returns.

VI. Limitations and Future Scope

The assumptions used to test the portfolios do not accurately reflect their performance in real-world conditions. Real-world investments incur costs such as taxes, transaction fees, indivisibility of assets, and unexpected transaction prices. However, these assumptions do not impact the relative performance when compared to benchmarked portfolios. The portfolio construction in this study considered only stocks. Future studies could include assets from different classes to evaluate their performance.

Despite the improved performance, the proposed model used only historical values and STIs as input values for LSTM forecasting. Future studies could explore the integration of other sources of data, such as news articles and social media sentiment analysis, to improve the model's predictive power. Including exogenous factors, such as interest rates, inflation rates, and exchange rates, could also provide more comprehensive forecasting results. Moreover, the proposed model was validated only on the NIFTY stocks and on a single time-frame, limiting the generalizability of the results to other stock markets and time frames. Future work could explore the model's performance across different markets and under varying market conditions, such as highly volatile markets or those experiencing sudden shocks. This would help assess the model's robustness and applicability in diverse financial environments, providing insights into its potential for broader use in real-world scenarios.

References

[1] H. M. Markowitz, “Portfolio selection,” Journal of finance, vol. 7, no. 1, pp. 71–91, 1952.

[2] Z. Zhang, S. Zohren, and S. Roberts, “Deep Learning for Portfolio Optimization,” Journal of Financial Data Science, vol. 2, no. 4, pp. 8–20, 2020, doi: 10.3905/jfds.2020.1.042.

[3] W. Wang, W. Li, N. Zhang, and K. Liu, “Portfolio formation with preselection using deep learning from long-term financial data,” Expert Syst Appl, vol. 143, p. 113042, 2020, doi: 10.1016/j.eswa.2019.113042.

[4] Z. Zhou, Z. Song, T. Ren, and L. Yu, “Two-Stage Portfolio Optimization Integrating Optimal Sharp Ratio Measure and Ensemble Learning,” IEEE Access, vol. 11, no. December 2022, pp. 1654–1670, 2023, doi: 10.1109/ACCESS.2022.3232281.

[5] Y. H. Chou, S. Y. Kuo, and Y. T. Lo, “Portfolio Optimization Based on Funds Standardization and Genetic Algorithm,” IEEE Access, vol. 5, pp. 21885–21900, 2017, doi: 10.1109/ACCESS.2017.2756842.

[6] J. B. Guerard, H. Markowitz, and G. Xu, “Earnings forecasting in a global stock selection model and efficient portfolio construction and management,” Int J Forecast, vol. 31, no. 2, pp. 550–560, 2015, doi: https://doi.org/10.1016/j.ijforecast.2014.10.003.

[7] Y. Zhang, X. Li, and S. Guo, “Portfolio selection problems with Markowitz’s mean–variance framework: a review of literature,” Fuzzy Optimization and Decision Making, vol. 17, no. 2, pp. 125–158, 2018, doi: 10.1007/s10700-017-9266-z.

[8] M. Johnson, “Forecasting in Emerging Markets: Challenges and Solutions,” Journal of Emerging Market Finance, vol. 18, no. 1, pp. 1–14, 2019.

[9] L. T. Quang, “Application of Artificial Intelligence-Genetic Algorithms to Select Stock Portfolios in the Asian Markets,” International Journal of Advanced Computer Science and Applications, vol. 13, no. 12, pp. 469–476, 2022, doi: 10.14569/IJACSA.2022.0131257.

[10] J. Sen, A. Dutta, and S. Mehtab, “Stock Portfolio Optimization Using a Deep Learning LSTM Model,” 2021 IEEE Mysore Sub Section International Conference, MysuruCon 2021, pp. 263–271, 2021, doi: 10.1109/MysuruCon52639.2021.9641662.

[11] Z. Wang, K. Li, S. Q. Xia, and H. Liu, “Economic Recession Prediction Using Deep Neural Network,” Journal of Financial Data Science, vol. 4, no. 3, pp. 108–127, 2022, doi: 10.3905/jfds.2022.1.097.

[12] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Appl Stoch Models Bus Ind, vol. 33, no. 1, pp. 3–12, 2017, doi: 10.1002/asmb.2209.

[13] A. M. Rather, “LSTM-based Deep Learning Model for Stock Prediction and Predictive Optimization Model,” EURO Journal on Decision Processes, vol. 9, no. May, p. 100001, 2021, doi: 10.1016/j.ejdp.2021.100001.

[14] F. Ploessl, T. Just, and L. Wehrheim, “Cyclicity of real estate-related trends: topic modelling and sentiment analysis on German real estate news,” Journal of European Real Estate Research, vol. 14, no. 3, pp. 381–400, 2021, doi: 10.1108/JERER-12-2020-0059.

[15] A. Chamekh, M. Mahfoudh, and G. Forestier, “Sentiment Analysis Based on Deep Learning : A Comparative Study,” Lecture Notes in Computer

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13369 LNAI, pp. 498–507, 2022, doi: 10.1007/978-3-031-10986-7_40.

[16] Q. Liu, Z. Tao, Y. Tse, and C. Wang, "Stock market prediction with deep learning: The case of China," Financ Res Lett, vol. 46, no. June, p. 102209, 2022, doi: 10.1016/j.frl.2021.102209.

[17] M. Dixon, D. Klabjan, and J. H. Bang, "Classification-based financial markets prediction using deep neural networks," Algorithmic Finance, vol. 6, no. 3–4, pp. 67–77, 2017, doi: 10.3233/AF-170176.

[18] H. Gunduz, Y. Yaslan, and Z. Cataltepe, "Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations," Knowl Based Syst, vol. 137, pp. 138–148, 2017, doi: 10.1016/j.knosys.2017.09.023.

[19] G. Taroon, A. Tomar, C. Manjunath, M. Balamurugan, B. Ghosh, and A. V. N. Krishna, "Employing Deep Learning in Intraday Stock Trading," Proceedings - 2020 5th International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2020, pp. 209–214, 2020, doi: 10.1109/ICRCICN50933.2020.9296174.

[20] B. Moews and G. Ibikunle, "Predictive intraday correlations in stable and volatile market environments: Evidence from deep learning," Physica A: Statistical Mechanics and its Applications, vol. 547, no. xxxx, p. 124392, 2020, doi: 10.1016/j.physa.2020.124392.

[21] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," PeerJ Comput Sci, vol. 7, pp. 1–21, 2021, doi: 10.7717/peerj-cs.476.

[22] A. Shah, M. Gor, M. Sagar, and M. Shah, "A stock market trading framework based on deep learning architectures," Multimed Tools Appl, vol. 81, no. 10, pp. 14153–14171, 2022, doi: 10.1007/s11042-022-12328-x.

[23] P. Ghosh, A. Neufeld, and J. K. Sahoo, "Forecasting directional movements of stock prices for intraday trading using LSTM and random forests," Financ Res Lett, vol. 46, no. June, p. 102280, 2022, doi: 10.1016/j.frl.2021.102280.

[24] M. S. Sivri and A. Ustundag, "An adaptive and enhanced framework for daily stock market prediction using feature selection and ensemble learning algorithms," Journal of Business Analytics, vol. 00, no. 00, pp. 1–21, 2023, doi: 10.1080/2573234X.2023.2263522.

[25] A. W. Li and G. S. Bastos, "Stock market forecasting using deep learning and technical analysis: A systematic review," IEEE Access, vol. 8, pp. 185232–185242, 2020, doi: 10.1109/ACCESS.2020.3030226.

[26] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. Shahab, "Deep learning for stock market prediction," Entropy, vol. 22, no. 8, 2020, doi: 10.3390/E22080840.

[27] H. Abbasimehr, M. Shabani, and M. Yousefi, "An optimized model using LSTM network for demand forecasting," Comput Ind Eng, vol. 143, no. March, p. 106435, 2020, doi: 10.1016/j.cie.2020.106435.

[28] I. Valova, N. Gueorguieva, T. Aayushi, P. Nikitha, and H. Mohamed, "Hybrid Deep Learning Architectures for Stock Market Prediction," Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science, pp. 1–8, 2023, doi: 10.11159/cist23.121.

[29] P. N. Kolm, R. Tütüncü, and F. J. Fabozzi, "60 Years of portfolio optimization: Practical challenges and current trends," Eur J Oper Res, vol. 234, no. 2, pp. 356–371, Apr. 2014, doi: 10.1016/J.EJOR.2013.10.060.

[30] F. D. Freitas, A. F. De Souza, and A. R. de Almeida, "Prediction-based portfolio optimization model using neural networks," Neurocomputing, vol. 72, no. 10–12, pp. 2155–2170, 2009, doi: 10.1016/j.neucom.2008.08.019.

[31] C. Hao, J. Wang, W. Xu, and Y. Xiao, "Prediction-Based Portfolio Selection Model Using Support Vector Machines," in 2013 Sixth International Conference on Business Intelligence and Financial Engineering, 2013, pp. 567–571. doi: 10.1109/BIFE.2013.118.

[32] C. F. Huang, "A hybrid stock selection model using genetic algorithms and support vector regression," Appl Soft Comput, vol. 12, no. 2, pp. 807–818, Feb. 2012, doi: 10.1016/J.ASOC.2011.10.009.

[33] S. Il Lee and S. J. Yoo, "Threshold-based portfolio: the role of the threshold and its applications," Journal of Supercomputing, vol. 76, no. 10, pp. 8040–8057, 2020, doi: 10.1007/s11227-018-2577-1.

[34] Y. Ma, R. Han, and W. Wang, "Portfolio optimization with return prediction using deep learning and machine learning," Expert Syst Appl, vol. 165, no. September 2020, p. 113973, 2021, doi: 10.1016/j.eswa.2020.113973.

[35] G. Li, "Information sharing and stock market participation: Evidence from extended families," Review of Economics and Statistics, vol. 96, no. 1, pp. 151–160, 2014.

[36] Y. Zhang, X. Li, and S. Guo, "Portfolio selection problems with Markowitz's mean–variance framework: a review of literature," Fuzzy Optimization and Decision Making, vol. 17, no. 2, pp. 125–158, 2018, doi: 10.1007/s10700-017-9266-z.

[37] F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, and W. M. Duarte, "Decision-making for financial trading: A fusion approach of machine learning and portfolio selection," Expert Syst Appl, vol. 115, pp. 635–655, 2019, doi: 10.1016/j.eswa.2018.08.003.

[38] Y. Ma, R. Han, and W. Wang, "Prediction-Based Portfolio Optimization Models Using Deep Neural Networks," IEEE Access, vol. 8, pp. 115393–115405, 2020, doi: 10.1109/ACCESS.2020.3003819.

[39] V. D. Ta, C. M. Liu, and D. A. Tadesse, "Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading," Applied Sciences (Switzerland), vol. 10, no. 2, 2020, doi: 10.3390/app10020437.

[40] W. Chen, H. Zhang, M. K. Mehlawat, and L. Jia, "Mean–variance portfolio optimization using machine learning-based stock price prediction," Appl Soft Comput, vol. 100, p. 106943, 2021, doi: 10.1016/j.asoc.2020.106943.

[41] G. Kumar, S. Jain, and U. P. Singh, "Stock market forecasting using computational intelligence: A survey," Archives of computational methods in engineering, vol. 28, no. 3, pp. 1069–1101, 2021.

[42] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," Neural Comput Appl, vol. 32, no. 13, pp. 9713–9729, 2020, doi: 10.1007/s00521-019-04504-2.

[43] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction : A case study of China stock market," in IEEE International Conference on Big Data (Big Data), 2015, pp. 2823–2824. [Online]. Available: http://arxiv.org/abs/1506.00019

[44] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," Financial Innovation, vol. 5, no. 1, 2019, doi: 10.1186/s40854-019-0138-0.

[45] L. Cao and F. E. H. Tay, "Financial forecasting using support vector machines," Neural Computing and Application, no. 10, pp. 184–192, 2001, doi: 10.1016/S0925-2312(03)00372-2.

[46] A. Sebastian and V. Tantia, "Transforming Finance With Deep Learning Predictions," in Navigating the Future of Finance in the Age of AI, 2024, ch. 12, pp. 227–252. doi: 10.4018/979-8-3693-4382-1.ch012.

[47] A. Sebastian and V. Tantia, "From data to decisions: Harnessing AI and big data for advanced business analytics," in Social Reflections of Human-Computer Interaction in Education, Management, and Economics, 2024, ch. 6, pp. 97–124. doi: 10.4018/979-8-3693-3033-3.ch006.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997.

[49] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," J Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00333-6.

[50] S. Chen and L. Ge, "Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction," Quant Finance, vol. 19, no. 9, pp. 1507–1515, 2019, doi: 10.1080/14697688.2019.1622287.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.

[52] S. Almahdi and S. Y. Yang, "An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown," Expert Syst Appl, vol. 87, pp. 267–279, 2017, doi: 10.1016/j.eswa.2017.06.023.

[53] V. Mohan, J. G. Singh, and W. Ongsakul, "Sortino Ratio Based Portfolio Optimization Considering EVs and Renewable Energy in Microgrid

Power Market," IEEE Trans Sustain Energy, vol. 8, no. 1, pp. 219–229, 2017, doi: 10.1109/TSTE.2016.2593713.

[54] Z. Zhou, Z. Song, T. Ren, and L. Yu, "Two-Stage Portfolio Optimization Integrating Optimal Sharp Ratio Measure and Ensemble Learning," IEEE Access, vol. 11, no. December 2022, pp. 1654–1670, 2023, doi: 10.1109/ACCESS.2022.3232281.

[55] F. A. Sortino, The Sortino Framework for Constructing Portfolios: Focusing on Desired Target ReturnTM to Optimize Upside Potential Relative to Downside Risk. Elsevier, 2009.

[56] O. B. Ansari and F. M. Binninger, "A deep learning approach for estimation of price determinants," International Journal of Information Management Data Insights, vol. 2, no. 2, p. 100101, 2022, doi: 10.1016/j.jjimei.2022.100101.

[57] I. H. Shakri, "Time series prediction using machine learning: a case of Bitcoin returns," Studies in Economics and Finance, vol. 39, no. 3, pp. 458–470, 2022, doi: 10.1108/SEF-06-2021-0217.

[58] G. Ding and L. Qin, "Study on the prediction of stock price based on the associated network model of LSTM," International Journal of Machine Learning and Cybernetics, vol. 11, no. 6, pp. 1307–1317, 2020, doi: 10.1007/s13042-019-01041-1.

[59] P. Singh and M. Jha, "Portfolio Optimization Using Novel EW-MV Method in Conjunction with Asset Preselection," Comput Econ, 2024, doi: 10.1007/s10614-024-10583-8.

[60] P. Singh and M. Jha, "Portfolio Optimization Using Novel EW-MV Method in Conjunction with Asset Preselection," Comput Econ, no. 0123456789, 2024, doi: 10.1007/s10614-024-10583-8.

[61] A. Chaweewanchon and R. Chaysiri, "Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning," International Journal of Financial Studies, vol. 10, no. 3, 2022, doi: 10.3390/ijfs10030064.

[62] W. Chen, H. Zhang, M. K. Mehlawat, and L. Jia, "Mean–variance portfolio optimization using machine learning-based stock price prediction," Appl Soft Comput, vol. 100, p. 106943, 2021, doi: 10.1016/j.asoc.2020.106943.

An Efficient Hierarchical Mechanism for Handling Network Partitioning Over Mobile Ad Hoc Networks

Ali Tahir, Fathe Jeribi*
College of Engineering and Computer Science, Jazan University, Jazan 45142, Kingdom of Saudi Arabia

***Abstract*—Mobile ad hoc networks exhibit distinctive challenges e.g., limited transmission range and dynamic mobility of the participating nodes. These challenges serve as the reasons for the frequent occurrence of network partitioning in mobile ad hoc networks. Network partitioning happens when a linked network topology is partitioned into two or more independent partitions. Because of this phenomenon, the participating node in one partition maintains no linkage with a node in another partition. Network partitioning results in the inaccessibility of mapping knowledge, logical labeling space, and logical structure of the participating nodes. As a result, the performance of a distributed hash tables (DHTs)-oriented routing mechanism is severely affected. In DHT-oriented routing methodologies, the logical network identifier of a new participating node is calculated by considering the logical network identifiers of all the physical neighboring nodes. The logical network identifiers are utilized for routing of packets from a source participating node to a destination participating node in the network. In the event of network partitioning, the incorrect computation of logical network identifiers happens concerning the physical proximity of the participating nodes. This research work suggests an effective routing mechanism to deal with the aforementioned network partitioning-related issues. Simulation results prove the superiority of the suggested scheme over the existing mechanisms.**

Keywords*—Mobile Ad Hoc networks; network partitioning; distributed hash tables; logical cluster member node; logical cluster leader; logical network identifier

I. Introduction

In mobile ad hoc networks (MANETs), there are two fundamental issues [1-4]. One of the issues is the restricted transmitting range. Another issue is the dynamic mobility of the participating nodes. Due to the identified issues, merging and partitioning of network architecture occur in mobile ad hoc networks. In the partitioning of a network, the linked structure is partitioned into two or more independent isolated segments [5-6]. Because of this phenomenon, the participating node in a segment has no contact with a participating node in another segment. On the other hand, the merging of network topology involves the integration of more than two independent partitioned segments [7-13]. This phenomenon occurs when the nodes in one independent partition start receiving hello messages from nodes of another isolated partition. It infers that the nodes in two independent partitioned networks lies in the identical transmitting range.

In DHT-based routing mechanisms, arrangement of the participating nodes is done considering the exploited logical identifier structure like chord, ring, or tree shapes [14-17]. In these topologies (chord, ring, or tree), there are restrictions on the number of routing paths. Some of the exploited topologies, the logical identifier space exhibit exclusively single routing path among the participating nodes. Consequently, these exploited topologies experience minimal resilience in choosing another routing path. In multi-dimensional DHT-based routing mechanisms, high resilience is observed during the selection of alternate routing paths. In mobile ad hoc networks (MANETs), there is a higher possibility that the logical identifier structure is partitioned. The partitioning of the logical identifier is dependent on the construction exploited by the logical identifier space. Hence, an effective logical identifier structure offers resilience related to the adaptation of routing paths when forwarding the data packets. It infers that there are alternate routing paths available in case of participating node failure or mobility. This type of construction is independent of redundant routing paths towards the destination. The reason is that exploited construction offers alternative routing paths (typically more than one) towards the destination participating node in the logical identifier space. In case a node is not available due to the partitioning of the network, then an alternate routing path can be exploited to reach that destination node in the network. It means that multiple routing paths assist in maintaining availability to a participating node in the network. In addition, logical identifier structure is separated in case of merging of two or more physical networks. It is a serious issue to recognize the happening of network merging at the logical identifier structure after physical networks merging. In addition, the smooth merging of two or more logical networks after the detection of network merging is a big issue. Therefore, there is a need to develop a DHT-based routing mechanism that should address the network merging detection and the merging of logical networks afterwards.

As discussed in study [18-22], the process of disintegrating the linked network topology into more than two isolated networks is referred as partitioning of the network. In the existing DHT dependent routing mechanisms, typically the attention is on providing an efficient resolution to the mismatch issue [23-26]. All these DHT dependent routing schemes neglect a critical challenge that is partitioning of logical networks. The partitioning of the network is responsible for the disruption of connectivity among the sender and receiver participating nodes. Because of the network partitioning, the participating nodes are unable to contact each other in the isolated partitions. The reason for this inaccessibility is the disruption of communication among the participating nodes. There are several reasons for the occurrence of network partitioning in mobile ad hoc networks. In mobile ad hoc networks, the dynamic movement of the participating nodes, the self-organized character and the restricted transmission scope are mainly the reasons for the persistent isolation of linked networks. In DHT dependent routing mechanisms, network partition is the cause of diverse critical challenges. The critical challenges due to partitioning of the network consist of DHT construction decomposition, depletion of the logical identifier space and participating nodes' mapping information (MPI) unavailability. In DHT dependent routing mechanisms, these critical challenges have deteriorated influences. Because logical identifiers instead of universal identifiers (IP or MAC address) are exploited for interaction between the participating nodes. The accessibility of mapping information stored at the anchor node plays a critical role. Because it provides the logical identifier information of the receiver participating node in the logical network. In DHT dependent routing schemes over MANETs, effectiveness can be achieved by efficiently recognizing the critical links or nodes. These critical links or nodes are responsible for activating the isolation or partitioning of physical topology. Therefore, recognition of critical nodes or links should be done promptly. By doing so, information depletion is alleviated significantly. Additionally, interruption of connection is greatly minimized.

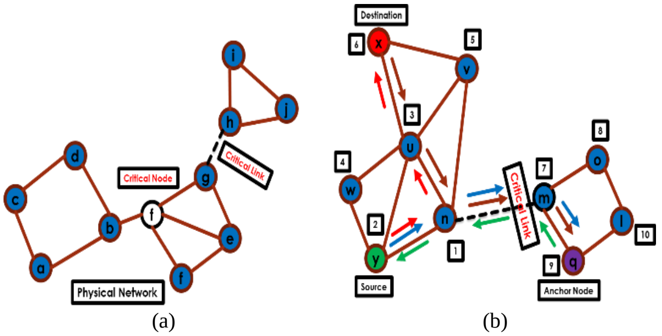
The critical link is a critical association in the logical network. If the critical link is unavailable, then partitioning of the network occurs. On the other hand, a critical participating node is a critical node in the logical network. If the critical participating node becomes unavailable, then partitioning of the network occurs. In the following Fig. 1(b), both the cases are well described. Recognition of critical links or nodes is efficiently done considering the prevalent neighboring participating nodes (x-hop) in a distributed manner.

To recognize the critical association and corresponding critical participating nodes in an efficiently distributed manner, mutual neighborhood connectivity information (x-hop) is exploited. In Fig. 1(b), n↔m association is perceived as a critical one (x-hop) in case the adjoining participating nodes of 'n' and 'm' are not in contact after the failure or breakage of the critical association. The association n↔m in Fig. 1(b) is considered one hop critical (x=1) in case neighborhood connectivity (one hop) is separated after failure or breakage of the critical association. The association n↔m is considered two hops critical (x=2) in case of unavailability of a mutual neighboring node among the neighborhood connectivity (two hops) of critical nodes 'n' and 'm'. This association is considered globally critical. The reason is the unavailability of a mutual neighboring node among the critical nodes 'n' and 'm'. For this purpose, consider x=3, 4 and so on. In the Fig. 1 (b), the node 'y' is considered as the critical participating node (globally). The reason is the separation of neighborhood connectivity of the participating node 'y' into two independent partitioned networks.

In the literature review [22-32], it is observed that research community relating to DHT-based routing does not highlight or elaborate the vital issues involving the partitioning of network. These critical issues should be effectively addressed by the research community. In this research work, a novel solution to address the crucial issues involving network partitioning is presented. It is called the 3DcPR (three-dimensional clustered partition detection with dynamic replication). This novel partition recognition strategy utilized the local neighborhood connectivity knowledge of every participating node in the 3D environment i.e., 3D logical identifier construction for efficient recognition of critical association in the physical topology. In addition, this strategy presents an efficient and dynamic solution for replication management. An efficient and dynamic replication methodology is considered beneficial to decrease the information dissipation and interruption of connection. Additionally, a novel retrieval methodology of misplaced logical identifier space is suggested. Due to the partitioning of network, the logical identifier space depletion phenomenon is observed. As far as we are aware, the suggested research work falls in the category of innovative ones in dealing with the partitioning of networks involving the DHT-based hierarchical routing protocol in mobile ad hoc networks (MANETs). This research shares the following remarkable innovations and advantages in comparison to the existing methodologies:

- We have delineated and instigated a novel partition detection and replication management mechanism termed as 3DcPR for working at the networking layer over MANETs.
- We have proposed a scheme for the identification of critical link and corresponding critical node. In this scheme, the local neighborhood knowledge of the participant nodes is utilized.
- We have suggested a replication management mechanism to smoothly establish and maintain the connectivity considering occurrence of the network partitioning over MANETs. This novel replication scheme significantly decreases the lookup latency.
- We have proposed a technique by integrating logical clustering along with DHTs to cope with the network partitioning scenarios in an effective manner.
- We have proposed an efficient technique that offers significant resilience against node failures or movements by utilizing the gateway nodes in the network.

The following tables (Table I and Table II) present the descriptions of key terms alongside their abbreviations being utilized in the current research work.



(a) (b)

Fig. 1. (a) Critical nodes and corresponding critical association, (b) Stocking and retrieval of mapping details before network partitioning.

TABLE I. ACRONYMS AND ABBREVIATIONS

Abbreviation	Description
BNT	Base Network Topology
LNT	Logical Network Topology
BNI	Base Network Identifier
LNI	Logical Network Identifier
LCL	Logical Cluster Leader
LCMN	Logical Cluster Member Node
LGPN	Logical Gateway Participating Node
MD	Mapping Details
LAPN	Logical Anchor Participating Node
MPTA	Mapping Petition Alarm
MRA	Mapping Reply Alarm
SCPN	Source Cluster Participating Node
DCPN	Destination Cluster Participating Node
DHTs	Distributed Hash Tables
MANETs	Mobile Ad Hoc Networks

TABLE II. DEFINITIONS OF IMPORTANT TERMINOLOGIES

Terminology	Definition
Base Network Topology	The physical neighborhood connection of the participating nodes.
Logical Network Topology	The logical network construction by utilizing the logical network identifiers of the participating nodes.
Logical Network Participating Node	The participating node in the logical network pattern on top of MANETs.
Base Network Identifier	The distinct identifier of logical network participating node in the physical network pattern.
Logical Network Identifier	The tag of the participating node in logical network pattern.
Logical Cluster Leader	It is the controller of the logical cluster in logical network pattern. It is selected by considering the highest participating node degree.
Logical Cluster Participating Node	It is the member node in a logical cluster.
Logical Anchor Participating Node	The participating node in the logical cluster which is responsible for stocking the mapping details of other participating nodes.
Logical Gateway Participating Node	It is participating node in logical cluster that comes under communication zones of more than one logical cluster leaders.

The rest of this paper is organized as follows: Section II presents the problem formulation with a detailed example scenario. Section III elaborates on the proposed novel mechanism for handling network partitioning on top of MANETs. In Section IV, the evaluation of the proposed mechanism is discussed. Ultimately, this research is wrapped up in Section IV alongside the future research directions.

II. PROBLEM FORMULATION

The disintegration of a linked topology into two or more isolated partitions is termed as network partitioning [33]. The nodes in a partition are not able to contact nodes participating in another isolated partition [34-35]. MANETs often experience the phenomenon of network partitioning [36-41]. The reasons behind it are high node mobility, restricted transmitting zone and the self-organized construction. In DHT-oriented routing mechanisms, mainly two critical challenges are faced due to the partitioning of networks in mobile ad hoc networks. These critical challenges are central reasons behind the degraded efficacy of the routing mechanisms based on DHT, one of which, is the inaccessibility of anchor nodes in the isolated and non-isolated topologies. Secondly, the depletion of logical identifier space is the consequence of network partitioning, and the lingering lookup delay associated with critical challenges.

A. Anchor Node Inaccessibility

In routing mechanisms dependent on DHT, the anchor node is responsible for stocking MD (mapping details) of another participating node over mobile ad hoc networks. Mapping details of the receiver participating node is indispensable for transmission between the sender and the receiver participating nodes. If the anchor node of the receiver participating node is unavailable in the connected topology, then MD of the receiver participating node is inaccessible. This phenomenon is the reason for the interruption of transmission among the sender and the receiver participating nodes in the logical topology. It infers that the accessibility of the receiver anchor node in the logical topology guarantees the persistent transmission between the sender and the receiver participating nodes.

Indeed, the partitioning of the network is responsible for the anchor node unavailability in the logical topology. Because of network partitioning of DHT-dependent logical topology into two or more independent networks, the sender and the receiver are members of one partition, but their respective anchor node lies in another isolated partition. In this scenario, the sender and the receiver participating nodes exist in similar partitions and are accessible to each other. Still, the sender and the receiver participating nodes are incapable of connecting. This phenomenon occurs because of the unreachability of mapping details stored at the receiver anchor node.

In Fig. 1 (b), a linked logical topology with a critical association n↔m is elaborated. This association between the participating nodes ‘n’ and ‘m’ is considered critical at a global level in the logical network over mobile ad hoc networks. The reason behind this consideration is that the participating nodes in the neighborhood of ‘n’ and ‘m’ are isolated. In a DHT dependent routing mechanism, every participating node is held responsible for stocking its MD (mapping details) at its anchor node for routing aims.

For example, a participating node ‘x’ stocks its mapping details (MD) at the anchor node ‘q.’ Now, if any other participating node, say the participating node ‘y’, wants to connect with the participating node ‘x’, then the sender participating node ‘y’ should first fetch the stocked mapping details from the anchor node ‘q’ of the participating node ‘x’. Afterwards, the participating node ‘y’ can initiate transmission with the receiver participating node ‘x’ in the logical topology.

In Fig. 2, the isolation of the critical association n↔m occurs in the logical topology over mobile ad hoc networks. Owing to the isolation of critical linkage in the logical topology, the partitioning of network topology takes place as presented in Fig. 2. After isolation of the network topology, the participating sender node ‘y’ and the receiver node ‘x’ remain in the identical partition. Although, the relative anchor node (i.e., node ‘q’) of the receiver node ‘y’ exists in another isolated partition in the logical network. In this scenario, the sender participating node ‘y’ is incapable of fetching the mapping details of the participating receiver node ‘x’ from the anchor node ‘q’.

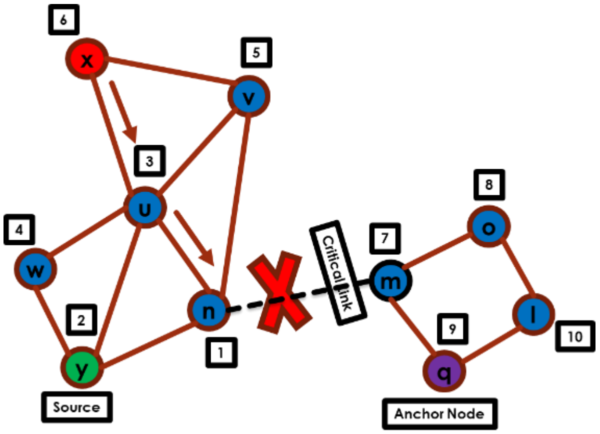


Fig. 2. Impossible retrieval of mapping details after network partitioning due to critical link failure.

Consequently, this phenomenon leads to the suspension of the interaction among the sender and the receiver participating nodes. Also, no participating node can fetch the mapping information of the receiver participating node ‘x’ in the absence of its anchor node in the logical environment. For smooth transmission, with other participating nodes in the logical topology, the receiver participating node ‘x’ should choose an updated anchor node and stock its mapping details. Afterward, the lookup queries targeting the participating node ‘x’ are efficiently rectified. However, the selection of another anchor node and stocking the mapping details on it becomes a reason for prolonged end-to-end delays and information depletion as well. Correspondingly, in case of failure or movement of anchor nodes to another partition, retrieval, or accessibility of stocked mapping details at that failed or moved anchor node should be assured to avoid the disruption of communication among the participating nodes in the logical network topology. It is considered as a critical issue in logical network topologies over mobile ad hoc networks that should be addressed efficiently in DHT-based routing schemes.

In the Fig. 3, the anchor node ‘q’ of the receiver participating node ‘x’ leaves the communication zone of another participating node ‘m’, and a new logical network identifier is allocated to it considering its neighborhood connectivity. Furthermore, the receiver participating node in this scenario chooses another anchor node ‘o’ and stocks its mapping details at the newly selected anchor node ‘o’ in the logical network topology. Simultaneously, this phenomenon becomes a reason for irretrievable enroute lookup requests by other participating nodes in the logical network topology. As a result, no transmission among the participating nodes is observed. In consequence, the DHT-based routing mechanism exhibits prolonged lookup delays and information depletion over mobile ad hoc networks.

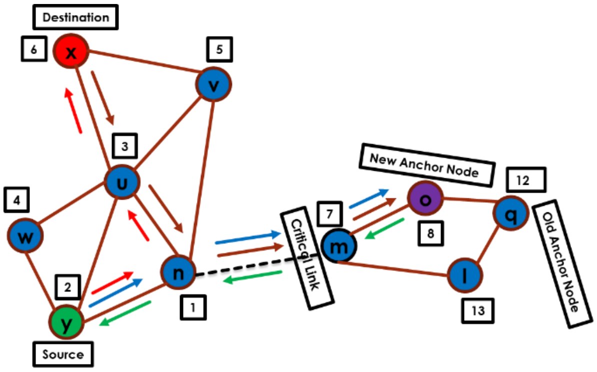


Fig. 3. New anchor node computation and stocking mapping details.

B. Depleted Logical Identifier Space

Another critical issue related to network partitioning that should be efficiently addressed is the depletion of logical identifier space. In a logical environment, the logical identifier space of an isolated partition can be reclaimed in another isolated partition. In consequence, the uniform division of the logical identifier space is achieved. Besides, isolated partitions can interact with each other due to the uniform division of logical identifier space. Therefore, retrieval of the logical identifier space is a well related critical challenge in DHT dependent routing schemes implemented over mobile ad hoc networks.

For this purpose, the pre-established network partitioning criterion performs an important responsibility for recognizing the occurrence of network partitioning in the logical identifier space over mobile ad hoc networks. There should be prompt, efficient pre-recognition of network partitioning until the authentic isolation of a logical network occurs in the logical environment. Pre-recognition is a necessary step towards effectively retrieving the logical identifier space in case of network partitioning activity in the logical environment over mobile ad hoc networks.

In a DHT-based routing mechanism, efficaciously recognition of network partitioning in a distributed manner is essentially required. The research community should rectify the identified challenge effectively to meet the requirements for the development of an efficacious DHT-based routing mechanism over mobile ad hoc networks. Moreover, in the scenario of network partitioning in MANETs, the prolonged lookup delays are observed. The prolonged lookup delay is a crucial associated issue that becomes worse when a critical challenge of network partitioning occurs over mobile ad hoc networks. This pertinent issue can be effectively resolved in a distributed way. For this purpose, the local neighborhood connectivity information of each participating node can be exploited. For the research community in this domain, resolving the said issue in a distributed fashion solicits strenuous exercise.

III. Solution Formulation

To rectify the issues related to network partitioning previously mentioned, a three-dimensional clustered distributed partition detection and replication routing mechanism referred to as 3DcPR is suggested. It exploits the local neighborhood connectivity information of the participating nodes for presenting the solution in a distributed fashion. The distributed resolution of the network partitioning-related challenges commensurate with the obligations of DHT dependent routing schemes in terms of scalability. In our mechanism for partition recognition, the following improvements have been explicitly suggested:

1) To offer the guarantee that the mapping details remain accessible in the disassociated partitions in the logical network. For this purpose, our mechanism dynamically replicates the mapping details in the logical network to handle the post-partitions replication challenges effectively.

2) To deal with the network partitioning, an effective mechanism is formulated to offer the recognition of network partitioning in a distributed fashion. This distributed partition recognition mechanism performs a decisive responsibility in identifying the network partitioning by considering several pre-established criteria. One of the criteria is the identification of the critical links among the participating nodes in the logical topology. Secondly, it focuses on offering an efficient strategy for the retrieval of the depleted logical identifier space due to partitioning of the network topology.

3) To encounter with the lingering lookup query delays, our mechanism provides the required efficacy in executing the replication management. Also, the influences of our replication management scheme considering the prolonged lookup delays are observed. It is found that lookup query delays are greatly minimized by exploiting our dynamic replication management strategy.

In 3DcPR, respective neighborhood connectivity information is exploited. 3DcPR exploits the neighborhood connectivity information up to two hops. It also utilizes the degree of the participating node for recognition of the critical links among the participating nodes in the logical environment. Furthermore, these parameters are considered for dynamic mapping details replication in the logical topology.

By doing so, 3DcPR guarantees the accessibility of mapping details in the isolated partitions after the occurrence of the network isolation phenomenon in the logical topology. As a result, the depleted information is significantly minimized. Also, our methodology is effectively coping with the partitioning of the network with remarkable reduction in the lingering delays of the lookup queries in the logical topology, achieved with no extra control overhead.

A. Recognition of Critical Linkage

In our methodology, HELLO messages among the logical cluster member nodes (LCMNs) are utilized for efficient recognition of network partitioning events in the logical environment over mobile ad hoc networks. In 3DcPR, every logical cluster member node reciprocates pre-established HELLO messages interval with each of its neighboring one-hop cluster members. In the HELLO message, information about the logical network identifier (LNI), logical space segment and a base network identifier of the participating cluster member is contained. Besides, the exchanged HELLO message consists of the one-hop neighborhood connectivity knowledge of the participating cluster member. It infers that each participating cluster member preserves the two-hop physical neighborhood connectivity knowledge in the logical topology for effective recognition of the network partitioning.

3DcPR recognizes a critical association among the two participating cluster members, say ‘n’ and ‘m.’ The association between the cluster members ‘n’ and ‘m’ is considered critical when one-hop neighborhood connectivity of cluster members ‘n’ and ‘m’ remains disconnected in case of failure or removal of the critical association between cluster members ‘n’ and ‘m.’ If a cluster member ‘n’ is one-hop critical of another cluster member ‘m,’ then all the neighborhood connectivity (one-hop) of the cluster member ‘n’ is not reachable from another cluster member ‘m,’ in case the cluster member ‘n’ is moved or failed. Due to this reason, the association among the cluster members ‘n’ and ‘m’ is considered critical. Besides, the association among the cluster members ‘n’ and ‘m’ is perceived as two hops critical in case the neighborhood connectivity (two-hops) of the participating cluster members ‘n’ and ‘m’ is not approachable in the absence of the critical association among the cluster members ‘n’ and ‘m.’

In 3DcPR, the state of the cluster members around the critical association is considered critical. Contrarily, the state of the cluster members, around the critical association, is perceived as non-critical. In 3DcPR, every cluster member informs all the neighbors (one-hop) about its state of being critical or not. For this purpose, each cluster member exploits HELLO messages. In the subsequent portions, the network partition recognition in a distributed manner along with a dynamic replication scheme is elaborated considering the clustering environment. Mostly, the gateways of the neighboring clusters serve as the critical cluster members around the critical association.

In 3DcPR, the k-hop neighborhood connectivity is considered for recognition of partition. Also, k-hop neighborhood connectivity information is considered for retrieval of a logical identifier structure. The k-hop neighborhood connectivity information is utilized for critical nodes recognition in the network. To attain the said objective, HELLO messages are occasionally exploited among adjoining cluster members to share neighborhood connectivity information (k-1 hop). In the HELLO messages, each cluster member shares neighborhood connectivity record (one-hop) with neighboring cluster members considering ‘k’ that equals 2. Moreover, the base network identifier, logical network identifier, and logical space segments are exchanged in the periodic HELLO messages. Hence, every cluster member keeps neighborhood connectivity information up to two-hops in the network.

The imperative step towards successful and efficient recognition of network partitioning is the timely detection of critical association among critical cluster members. For example, consider two participating critical cluster members ‘n’ and ‘m’ around the critical association n↔m. The occurrence of network partitioning phenomenon is provoked by the 3DcPR in a distributed fashion. This partition detection strategy considers precautionary actions involving the adjustment of a logical identifier structure. Besides, a specific duration of time is set. This set timer is termed as partition timer. Network partitioning happens if a critical cluster member ‘n’ around the critical association n↔m does not receive the HELLO messages from another participating cluster member ‘m’ after the pre-established HELLO break. In our mechanism for partition recognition, the pre-established partition timer is three times the HELLO break.

In the same way, the other participating cluster member recognizes the partitioning of the network. If partitioning of networks happens, then the cluster members around the critical association are responsible for retrieval of the misplaced logical identifier structure. For this purpose, the cluster members utilize the misplaced logical identifier structure in the separated independent partitions.

Mostly, the exploited logical identifier construction plays a vital role in the retrieval process of the logical identifier structure. The LIS recovery mechanism depends on the logical structure used. Every participating cluster member around the critical association retrieves the logical identifier structure in the three-dimensional space. For this purpose, every cluster member around the critical association retrieves it by just altering the value of the dimension. Correspondingly, the retrieval process in the chord construction is somewhat different. In the chord construction, logical network identifiers of the participating nodes around the critical association are altered for retrieval of a misplaced logical identifier structure. They modify their logical network identifiers considering the linkage (successor or predecessor) among the participating nodes around the critical association.

In the chord construction, the participating node (successor) retrieves the logical identifier structure by altering the logical network identifier to S. Also, the participating node (predecessor) in the chord construction modifies the logical network identifier to E for the successful retrieval of the misplaced logical identifier structure. The precedent participating node concerning the logical chord formation revives the logical identifier space by altering its logical network identifier to E, and also the descendent participating node in the chord formation recuperates the logical identifier space by modifying its logical network identifier to S. The exploitation of the misplaced logical identifier structure in the separated independent partitions has great benefits. One of the benefits is the uniform distribution of logical identifier structure in the separated independent partitions.

Our strategy suggests an efficient distributed methodology for recognition of partition. In it, the recognition is entirely distributed in nature. It exploits the local neighborhood connectivity knowledge of the participating cluster members for partition recognition in a distributed manner. It does not consider the dissipation of control knowledge at the global level.

B. *Partition Recognition and Replication Strategy*

To test 3DcPR, let us consider a 3DcRP [27] example as depicted in the following scenarios, for explaining the feasibility of our methodology for partition recognition and replication.

Initially, 3DcPR recognizes the critical association and corresponding critical participating cluster members. For this purpose, HELLO messages are exploited by our partition recognition and dynamic replication methodology, i.e., 3DcPR. Every cluster member occasionally exchanges local neighborhood connectivity knowledge (one-hop) with the neighboring cluster members by exploiting the HELLO messages. By doing this, every cluster member in the network contains the neighborhood connectivity knowledge (two-hops). In the HELLO messages, a logical network identifier, base network identifier, and logical space segment are exchanged. Maintaining the neighborhood connectivity knowledge (two-hops) facilitates in recognition of critical association and corresponding critical cluster members in the logical network.

To understand this, consider the example scenario depicted in Fig. 4. In this scenario, one-hop neighboring cluster members of ‘w’, ‘x’ and ‘z’ are {‘u’, ‘y’, ‘x’}, {‘v’, ‘u’, ‘w’, ‘x’, ‘z’}, and {‘x’, ‘y’, ‘v’}, respectively. When the participating cluster members, ‘w’, ‘x’ and ‘z,’ exchange the local neighborhood connectivity knowledge (one-hop), then another participating cluster member ‘v’ in cluster 1 with LNI 1 contains the local neighborhood knowledge (two-hops). The participating cluster member ‘v’ reviews the neighborhood connectivity knowledge (one-hop lists) of all its adjoining cluster members (cluster members ‘x’ and ‘z’) for searching the nexus cluster member among them. The participating cluster member ‘v’ does this exercise for broadcasting its state (either critical or not critical). In this exercise, the participating cluster member ‘v’ does not include itself. After reviewing the neighborhood connectivity knowledge (one-hop lists) of all its adjoining cluster members, the participating cluster member ‘v’ finds that a cluster member (also the gateway cluster member), say ‘y’, is the nexus cluster member among the neighborhood member ‘v’ failure does not have an impact on the connectivity. This connectivity {‘x’, ‘u’, ‘w’, ‘y,’ ‘z’} exists despite its (‘v’) failure or movement. Therefore, the participating cluster member ‘v’ announces its state as non-critical. Likewise, the participating cluster member 'y' hears the local neighborhood connectivity knowledge (one-hop) from each of its adjoining cluster members ‘’w’, ‘x’, ‘z’, and ‘m.’ This one-hop neighborhood connectivity lists are {‘u’, ‘x’}, {‘v’, ‘u’, ‘w’, ‘y’, ‘z’}, {‘v’, ‘x’, ‘y’}, {‘r’, ‘p’, ‘q’, ‘y’} from ‘w’, ‘x’, ‘z’, and ‘m’, respectively.

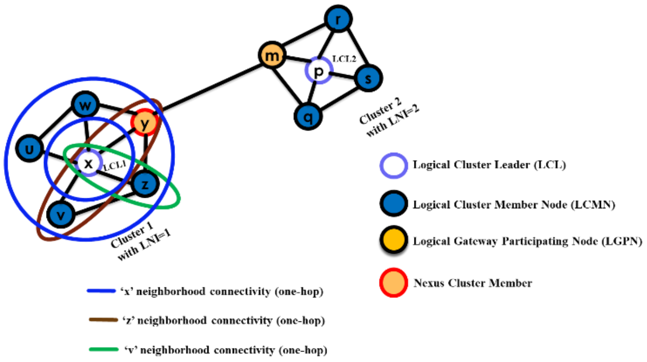

r
m
p
LCL2
s
q
Cluster 2 with LNI=2
w
y
u
x
LCL1
z
v
Cluster 1 with LNI=1
Logical Cluster Leader (LCL)
Logical Cluster Member Node (LCMN)
Logical Gateway Participating Node (LGPN)
Nexus Cluster Member
‘x’ neighborhood connectivity (one-hop)
‘z’ neighborhood connectivity (one-hop)
‘v’ neighborhood connectivity (one-hop)

Fig. 4. Nexus cluster member identification procedure.

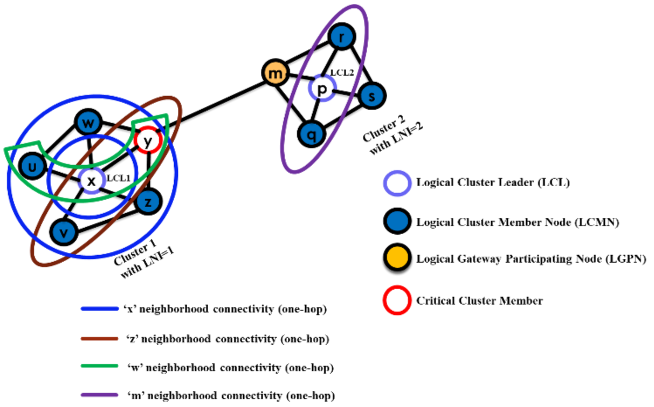

r
m
LCL2
p
s
q
Cluster 2 with LNI=2
w
y
u
x
LCL1
z
v
Cluster 1 with LNI=1
Logical Cluster Leader (LCL)
Logical Cluster Member Node (LCMN)
Logical Gateway Participating Node (LGPN)
Critical Cluster Member
'x' neighborhood connectivity (one-hop)
'z' neighborhood connectivity (one-hop)
'w' neighborhood connectivity (one-hop)
'm' neighborhood connectivity (one-hop)

Fig. 5. Critical cluster member identification.

It is depicted in Fig. 5 that the nexus cluster members among the adjacent cluster members ('w,' 'x,' 'z') of 'y' exists. It constitutes a linked network without the cluster member 'y.' Although, the one-hop neighborhood connectivity of the participating cluster member 'p' of cluster 2 with LNI 2 (also the gateway cluster member) does not become a part of the linked network of cluster members 'w', 'x', and 'z'.

It infers that the state of the two participating cluster members 'y' and 'm' is set as critical. They both also act as the gateway cluster members in clusters 1 and 2, respectively. In consequence, the association among the critical cluster members 'y' and 'm' is considered as critical.

For the example scenario as depicted in Fig. 6, the critical association is y ↔ m. Correspondingly, every participating cluster member refreshes its state of being critical or not in the network periodically. When a newly joined cluster member, say 'w,' in cluster 1 calculates its logical network identifier and it stocks its mapping details at its anchor cluster member, say 'q,' afterward. For this purpose, the newly joined cluster member 'w' sends cluster member 'y' stocks the mapping details of the newly joined cluster member 'w' if that mapping details are not sent by the other critical cluster member in the network.

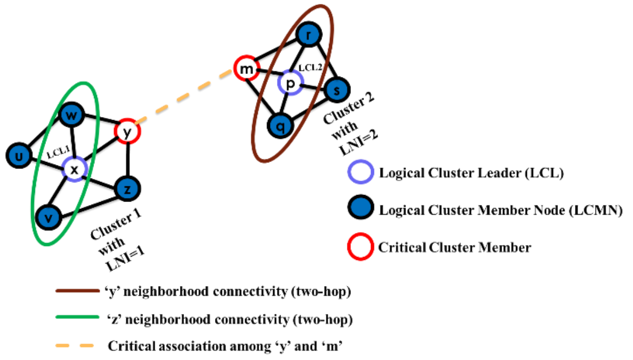

r
m
LCL2
p
s
q
Cluster 2 with LNI=2
w
y
u
LCL1
x
z
v
Cluster 1 with LNI=1
Logical Cluster Leader (LCL)
Logical Cluster Member Node (LCMN)
Critical Cluster Member
'y' neighborhood connectivity (two-hop)
'z' neighborhood connectivity (two-hop)
Critical association among 'y' and 'm'

Fig. 6. Critical association identification.

The critical cluster members 'y' and 'm' stock the mapping details for forwarding it further. Therefore, the critical cluster members 'y' and 'm' around the critical association y↔m, and logical cluster leaders (LCL1 and LCL2) maintain a copy of the mapping details included in the MPTA message. The phenomenon of replicating and retrieving mapping details before the occurrence of network partitioning is depicted in the Fig. 7. This dynamic replication strategy facilitates avoiding communication disruption in the separated partitions after the occurrence of network partitioning. The cluster members in two partitioned networks can acquire the mapping details from the critical cluster members 'y' and 'm' across the critical association and the logical cluster leaders (LCL1 & LCL2) as well.

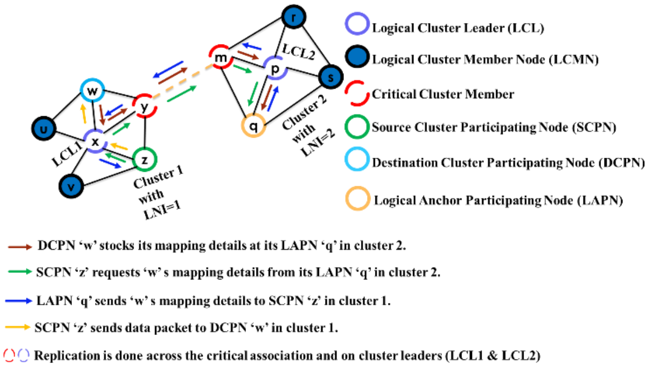

r
m
LCL2
p
s
q
Cluster 2 with LNI=2
w
y
u
LCL1
x
z
v
Cluster 1 with LNI=1
Logical Cluster Leader (LCL)
Logical Cluster Member Node (LCMN)
Critical Cluster Member
Source Cluster Participating Node (SCPN)
Destination Cluster Participating Node (DCPN)
Logical Anchor Participating Node (LAPN)
DCPN 'w' stocks its mapping details at its LAPN 'q' in cluster 2.
SCPN 'z' requests 'w' s mapping details from its LAPN 'q' in cluster 2.
LAPN 'q' sends 'w' s mapping details to SCPN 'z' in cluster 1.
SCPN 'z' sends data packet to DCPN 'w' in cluster 1.
Replication is done across the critical association and on cluster leaders (LCL1 & LCL2)

Fig. 7. Replication around the critical association and on cluster leaders.

By doing so, the participating cluster members can communicate in a partition after the failure or removal of the critical association in the network. In the Fig. 8, an example scenario is depicted to elaborate the phenomenon after the occurrence of network partitioning. In this example scenario, a sender logical cluster member node 'z' in cluster 1 having cluster leader 'x' wants to contact another receiver logical cluster member node 'w.' For this purpose, the sender cluster member 'w' first finds the anchor cluster member for the receiver cluster member 'w.' After finding the anchor cluster member, i.e., 'q' of the receiver cluster member 'w,' the sender cluster member 'z' acquires the stocked mapping details from the anchor cluster member 'q' of 'w.' The stocked mapping details is not accessible for the sender cluster member 'z' if the critical association among the critical cluster members 'y' and 'm' fails or removes.

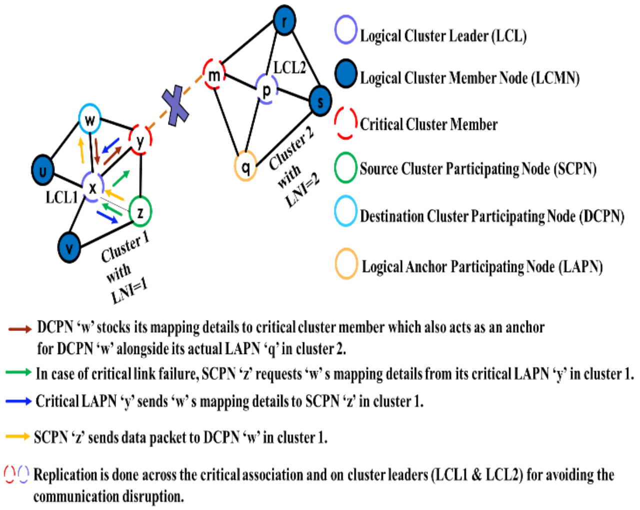

r
m
LCL2
p
s
q
Cluster 2 with LNI=2
w
y
u
LCL1
x
z
v
Cluster 1 with LNI=1
Logical Cluster Leader (LCL)
Logical Cluster Member Node (LCMN)
Critical Cluster Member
Source Cluster Participating Node (SCPN)
Destination Cluster Participating Node (DCPN)
Logical Anchor Participating Node (LAPN)
DCPN 'w' stocks its mapping details to critical cluster member which also acts as an anchor for DCPN 'w' alongside its actual LAPN 'q' in cluster 2.
In case of critical link failure, SCPN 'z' requests 'w' s mapping details from its critical LAPN 'y' in cluster 1.
Critical LAPN 'y' sends 'w' s mapping details to SCPN 'z' in cluster 1.
SCPN 'z' sends data packet to DCPN 'w' in cluster 1.
Replication is done across the critical association and on cluster leaders (LCL1 & LCL2) for avoiding the communication disruption.

Fig. 8. Avoiding communication disruption after network partitioning.

Our methodology provides an effective solution to this unavailability of mapping details in case of critical association failure. It gives assurance to provide accessibility of mapping details after the failure of critical association (occurrence of network partitioning). In our methodology, pre-partitioning precautions like recognition of the critical association/critical cluster members and dynamic replication assist in avoiding the unavailability of the mapping details. In the above example scenario, the pre-partitioning criterion (replicas placement) assists the sender cluster member 'z' to retrieve the mapping details of the receiver cluster member 'w' from the critical cluster member 'y' although the network partitioning occurred due to the failure or removal of critical association among critical cluster members as depicted in Fig. 8. Even if the critical cluster member 'y' fails or moves to another location, the sender cluster member 'z' can obtain the replicated mapping details of the receiver cluster member 'w' from the cluster leader 'x.' Therefore, our methodology effectively addresses the identified issue by replicating the mapping details on both the critical cluster members around the critical association and cluster leaders.

Furthermore, 3DcPR activates the occurrence of network partitioning in case the critical cluster members around the critical association are not able to receive HELLO messages from each other for a specific pre-established time duration (partition-timer). Although, it is a must for the critical cluster members to be both one-hop and two-hops critical. However, the exception is for the critical cluster member that is two-hops critical but has a state of one-hop non-critical as well. In the Fig. 9, it is depicted that critical cluster members 'y' and 'm' around the critical association y↔m are not able to receive the HELLO messages from each other, then this phenomenon provokes the network partition after the expiry of the pre-established partition timer.

In this scenario, the respective critical cluster members retrieve or reiterate the depleted logical identifier space. If the anchor cluster member is considered as the critical cluster member, then it mirrors its MD (mapping details) around the critical association in the network. To avoid the communication disruption after the network partition, the mirroring deployment and retrieval of logical identifier space is an important step. This situation may give rise to uniform and distributed separated partitions though there exists network partitioning. The retrieval process of logical identifier space is different for every logical construction. Because this retrieval process is dependent on the construction of logical identifier space, the exploited logical construction may include a ring, a cord, a tree, or a 3D.

In the case of a high mobility environment, DHT-oriented routing methodologies, over mobile ad hoc networks, face prolonged lookup delays and decreased overall performance. Our methodology utilizes the dynamicity of the mobile ad hoc networks. It means the relative connectivity, i.e., two-hop topological knowledge. Moreover, our methodology utilizes the local network variation (one-hop connectivity knowledge). For relative connectivity and local network variation, periodically sent HELLO messages are utilized by the 3DcPR. By doing so, it offers assured reachability/availability of the network. Besides, it is done with no additional control overhead. As a result, end-to-end delays are effectively minimized for lookup queries in the network.

IV. Results and Discussion

To assess the effectiveness of the proposed mechanism, the network simulator version 2.35 is exploited for creating and running every simulation. NS-2.35 is an open-source event network simulated adopted by the research community. The propagation model utilized by the proposed mechanism is Two Ray Ground. This propagation model is exploited to simulate IEEE 802.11 concerning the standard values for physical and link layers. In Table III, the simulation parameters are demonstrated. The effectiveness of the proposed mechanism i.e., 3DcPR is evaluated with the existing logical networks over MANETs strategies like 3DDR [23] and 3DRP [24]. These previous strategies exploit the three-dimensional shape to efficiently cope with the mismatch issue among the logical and physical network topologies. Besides, the existing strategy i.e., 3DDR provides notable resiliency against participating node movement or failure situations. The participating node calculates its logical network identifier considering the LNIs of each neighboring participating nodes within a network. Consequently, a significant escalation concerning control, computation and routing traffic overheads is observed. Conversely, a logical cluster leader (LCL) is devoted for distributing the logical network identifiers among its logical cluster member nodes residing within its logical cluster concerning the proposed mechanism i.e., 3DcPR. This phenomenon remarkably minimizes the control, computational, and routing traffic overheads within the network.

The playground area size is considered as 1000m * 1000m as per the proposed methodology for efficiently conducting the simulations in NS 2.35. The utmost broadcasting span is set as 50 m. Remember, the broadcasting span is adopted in a manner that passes over each participating nodes a specific distance i.e., two hop. In 3DcPR, HELLO notifications are regularly interchanged and utilized for effectively preserving the neighborhood connectivity among the adjacent participating nodes in the network. The uniform distribution is implemented when establishing a tangible network simulation environment. In addition, as demonstrated in Table III, the network simulations are employed concerning motion patterns by exploiting the high mobility of the participating nodes i.e., 7 m/s to 25m/s. The exact cause of constricting the speed of the participating nodes from 7m/s to 25m/s is the unsustainability of MANETs considering considerably high or low mobility situations. Moreover, a consistent selection of the participating nodes' speed is adopted as 7m/s to 25m/s. For maintaining the connection in the physical network, BonnmotionV2 is exploited in the current scheme for generating the mobility cases in accordance with RWP (random way point) network model. RWP is the first choice of renowned researchers of the wireless domain for perfectly assessing the effectiveness of the ad hoc networks alongside its other prominent benefits like ease and quick functionality. Remember, OLSR is being exploited by the proposed scheme as an underlying strategy. Moreover, 10 unique files are commissioned by every participating node as per the proposed strategy. Also, for modelling routing traffic, the random traffic pattern is employed by 3DcPR. For establishing the data traffic, the CBR (constant-bit-rate) flows are utilized over the UDP protocol. In addition, it is supposed to be the connectivity of network formation for both 3DcPR and 3DDR. In our experiments, ten executions per case are done. The aggregate results are depicted by utilizing graphs. Lastly, the simulation duration is established as 500 seconds for performing different scenarios for accurately evaluating the performance of the proposed and existing mechanisms.

TABLE III. SIMULATION PARAMETERS

Parameter	Value
Playground Size	[1000 m × 1000 m]
Number of Nodes	[25-400]
Transmission Range	50 m
Simulation Time	500 s
Data Rate	[1- 500pps]
Start of Data Transmission	[70, 300]
End of Data Transmission	[250, 499]
Node Speed	[7 m/s – 25 m/s]
Traffic Model	Random Traffic Pattern
Mobility Model	Random Way Point
Radio Propagation Model	Two Ray Ground

In this research article, three performance parameters are considered for efficiently assessing the performance of the suggested mechanism. These performance parameters are evaluated against the high mobility of the participating nodes and rising network size.

1) Routing Traffic Overhead (RTO): It is the entire control overhead packets as utilized by the routing protocol in the mobile ad hoc network.

2) Packet Delivery Ratio (PDR): The ratio between the total mapping request packet (MREQ) initiated and the total MREQ entertained successfully by receiving the mapping reply packets (MREP).

3) Average End-To-End Delay: The average time elapsed between when the source node initiates MREQ and the source node gets MRPY.

A. Average End-To-End Delay

To evaluate the performance of a DHT-based routing mechanism, the average end-to-end delay is considered as an important criterion. Through this criterion, the entire network performance can be assessed as well. It is required to perform a detailed investigation concerning the effect of rising network size alongside the speed of the participating node over aggregated end-to-end delay. The detailed impact analysis of increasing network size and node moving speed over average end-to-end delay is carried out and demonstrated in the following Fig. 9.

In the Fig. 9, the average end-to-end delay of 3DcPR, 3DDR [23], and 3DRP [24] is computed against the changing speed participating nodes ranging from 7m/s to 25m/s considering the increasing network size It is demonstrated in Fig. 9 that the average end-to-end delay of the proposed mechanism i.e., 3DcPR is notably decreased in contrast to the existing schemes 3DDR, and 3DRP. The central reason behind this significant reduction in average end-to-end delay is novel partition discovery and effective replication strategy in a distributed way which discovers the censorious participating nodes first and clones the mapping details for facilitating the uninterrupted connection. The average end-to-end delay is considerably minimized as the extra clones are deployed in the network to avoid transmitting the message request to the actual anchor participating node rather than a nearby participating node carrying the mapping details acting as a clone. As a result, the traffic overhead is significantly decreased considering the proposed mechanism i.e., 3DcPR in comparison to the existing schemes 3DDR, and 3DRP.

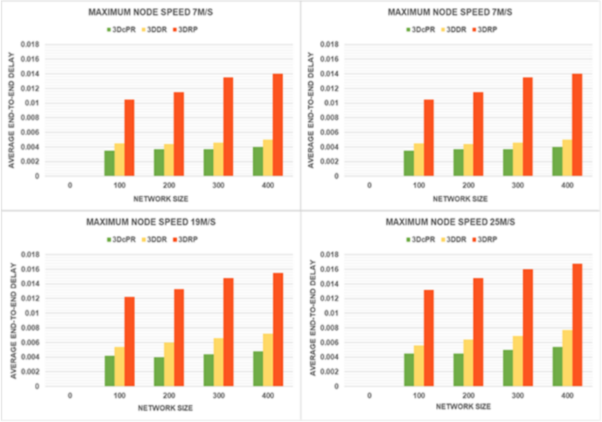


Fig. 9. Average end-to-end delay vs. network size.

Moreover, the notable reduction in routing traffic overhead minimizes the contention to establish the connection with the medium at media access layer considering IEEE 802.11. Furthermore, this phenomenon assists in remarkably decreasing the average end-to-end delay concerning the suggested mechanism.

B. Routing Traffic Overhead

One of the critical parameters to evaluate the performance of the proposed mechanism considering the increasing network size alongside the fluctuating participating node speed is the routing traffic overhead. Specifically, the routing traffic overhead gains immense value when it is computed for each lookup query in the network. As demonstrated in Fig. 10, the routing traffic overhead is computed for the proposed mechanism i.e., 3DcPR in comparison to the existing strategies 3DDR, and 3DRP considering the fluctuating participating node speeds i.e., 7m.s to 25m/s and changing network size. A remarkable decrease in the routing traffic overhead is observed as depicted in Fig. 10 for the proposed mechanism considering the varying network size and fluctuating participating node speeds in comparison to the existing schemes. The reason behind this significant decrease in the routing traffic overhead is the novel partition discovery mechanism alongside the effective replication scheme. This proposed strategy discovers the critical participating nodes in the network and replicates the mapping details which results in minimized lookup request overhead in the network. Remember, as the participating nodes over MANETs have fluctuating speeds which results in frequent network construction changes. Consequently, this phenomenon increases the routing traffic and lookup requests overhead over MANETs.

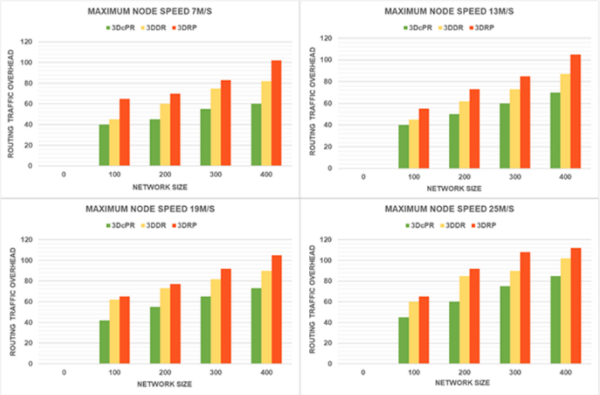


Fig. 10. Routing traffic overhead vs. network size.

The primary reasons behind this rise in the routing traffic overhead are the relocation the anchor participating nodes, stocking the mapping details at newly relocated anchor participating nodes, the retrieval of the logical identifier construction, and re-computation of logical network identifiers happen due to the network partitioning. As demonstrated in Fig. 10, the calculated routing traffic overhead for the proposed scheme i.e., 3DcPR is much lesser than the existing strategies due to the discovery of critical participating nodes and corresponding critical association between them in the network and the novel replica management strategy. Besides, the proposed scheme i.e., 3DcPR utilizes the clustering environment and three-dimensional network construction for arranging the participating nodes in logical network over MANETs.

C. Packet Delivery Ratio (PDR)

As discussed earlier, the potential of a routing mechanism is the effective delivery of data packets towards the destination participating node in the network which is termed as the packet delivery ratio. Remember, the increasing number of participating nodes minimizes the packet delivery ratio concerning the routing mechanism. As known, with the increase in number of participating nodes raises the collision of packets at media access control layer concerning IEEE 802.11. It infers that the reduction in successful message request and message reply alarms happens in the network because of escalating collision of packets in the network. Additionally, the number of hops among the source and destination participating nodes rises in the network. Consequently, this phenomenon amplifies the total amount of transmission in the network. Hence, packet delivery delays are observed in the network and there are more chances of collision of packets concerning media access layer. As a result, the effective transmitting of message request and message reply alarms is badly affected in the network.

In Fig. 11, the packet delivery ratio for the proposed mechanism i.e., 3DcPR and the existing schemes is computed. As demonstrated in Fig. 11, the effect of increasing number of the participating nodes alongside the fluctuating speeds is observed lower for the proposed mechanism i.e., 3DcPR in contrast to the existing strategies 3DDR, and 3DRP. Hence, the proposed mechanism 3DcPR achieves better results concerning packet delivery ratio in comparison to the existing schemes which proves the efficaciousness of the proposed methodology. Besides, this particular potential of the proposed mechanism makes it a strong candidate for its execution and efficient transmitting of packets in a large-scale mobile ad hoc networks.

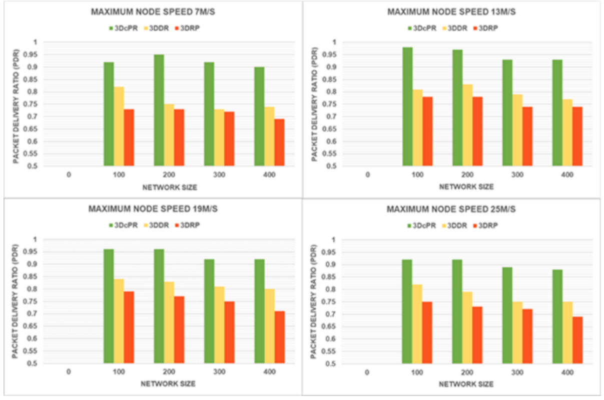


Fig. 11. Packet delivery ratio vs. network size.

In addition, the packet delivery ratio concerning 3DcPR is preserved and enhanced by utilizing the proposed partition discovery and replica management strategy. The proposed methodology clones the mapping details on each censorious connectivity down the route towards the logical anchor participant. This phenomenon is the reason behind the enhanced packet delivery ration for the proposed methodology as the mapping details are accessible even in the split partitions considering the occurrence of network partitioning. Besides, the high mobility of the participating node and fluctuating network size provoke the partitioning of the network over MANETs. So, in case of network partitioning, it is almost impossible to fetch the mapping details considering the existing methodologies like 3DRP, and 3DDR. The reason behind this inaccessibility to mapping details is the existence of logical anchor participant in the split partition. Additionally, in the proposed approach, the replica preservation around the critical connectivity down the path towards logical anchor participant enhances the reliability of 3DcPR. The other reason behind this reliability enhancement of 3DcPR is significant decrease in routing traffic overhead. Consequently, at the media access layer, the collision of packets is notably minimized. The proposed methodology i.e., 3DcPR preserves this scenario for long which results in significant improvement in delivering the packets towards the destination. Moreover, the existing schemes exhibits the mismatch issue among the physical and logical network topologies. Consequently, the existing methodologies suffer from lengthy paths towards destination and unnecessary routing traffic in the network. On the other hand, the proposed methodology i.e., 3DcPR considers three-dimensional construction in a clustering environment to perfectly map the physical linkage in the logical network and avoiding the mismatch issue.

V. Conclusion and Future Directions

One of the critical problems concerning the DHT-oriented routing over mobile ad hoc networks is the partitioning of networks. Specifically, network partitioning phenomenon becomes worsen in case of high mobility and with the rising number of the participating nodes in the network. Consequently, the mapping details are inaccessible, the logical identifier construction is dissolved, and lengthy lookup delays are observed considering the occurrence of network partitioning. As a result, no communication within the mobile ad hoc network. So, the prompt discovery of network partitioning over MANETs is a demanding task considering the high mobility and increasing network size of the participating nodes in the network. In this research article, the issue of network partitioning over MANETs is tackled by developing a novel mechanism in a distributed fashion. A novel partition discovery mechanism alongside the effective clone management strategy is recommended to successfully deal with the partitioning of network over MANETs. Also, a distinct and effective replication mechanism is proposed which results in minimized lookup and routing delays considering the DHT-oriented mobile ad hoc networks. The prompt preventive actions e.g., clone management and neighbourhood connectivity knowledge of the participating nodes are considered for discovering the network partitioning in time and effectively handling the issues after network partitioning. As a future work, the implementing of the proposed methodology for secure data dissemination in mobile ad hoc networks is under consideration.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Kaviani, S., Ryu, B., Ahmed, E., Larson, K. A., Le, A., Yahja, A., & Kim, J. H. (2021). Robust and Scalable Routing with Multi-Agent Deep Reinforcement Learning for MANETs. arXiv preprint arXiv:2101.03273.

[2] Kang, D., Kim, H. S., Joo, C., & Bahk, S. (2018). ORGMA: Reliable opportunistic routing with gradient forwarding for MANETs. Computer Networks, 131, 52-64.

[3] Patel, S., & Pathak, H. (2021). A mathematical framework for link failure time estimation in MANETs. Engineering Science and Technology, an International Journal.

[4] Pathan, M. S., Zhu, N., He, J., Zardari, Z. A., Memon, M. Q., & Hussain, M. I. (2018). An efficient trust-based scheme for secure and quality of service routing in MANETs. Future Internet, 10(2), 16.

[5] Kinge, P., & Ragha, L. Comparative Analysis of Methodologies used to Address the MANET Partitioning Problems.

[6] Tarasov, M., Seitz, J., & Artemenko, O. (2011, October). A network partitioning recovery process in Mobile Ad-Hoc Networks. In 2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (pp. 32-36). IEEE.

[7] Shah, N., & Qian, D. (2010, December). Cross-layer design for merging of unstructured P2P networks over MANET. In 2010 Proceedings of the 5th International Conference on Ubiquitous Information Technologies and Applications (pp. 1-7). IEEE.

[8] Li, Y., & Wang, X. (2019). A novel and efficient address configuration for MANET. International Journal of Communication Systems, 32(13), e4059.

[9] Mutanga, M. B., TarwireyI, P., & Adigun, M. (2015, November). Handling network merging and partitioning in MANETs. In 2015 First International Conference on New Technologies of Information and Communication (NTIC) (pp. 1-6). IEEE.

[10] Shah, N., Qian, D., & Wang, R. (2015). Merging of P2P Overlays Over Mobile Ad Hoc Network: Evaluation of Three Approaches. Adhoc & Sensor Wireless Networks, 25.

[11] Datta, A., & Aberer, K. (2006). The challenges of merging two similar structured overlays: A tale of two networks. In Self-Organizing Systems (pp. 7-22). Springer, Berlin, Heidelberg.

[12] Abid, S. A., Othman, M., Shah, N., Sabir, O., Khan, A. U. R., Ali, M., ... & Ullah, S. (2015). Merging of DHT-based logical networks in MANETs. Transactions on Emerging Telecommunications Technologies, 26(12), 1347-1367.

[13] Abid, S. A., Othman, M., & Shah, N. (2014). A survey on DHT-based routing for large-scale mobile ad hoc networks. ACM Computing Surveys (CSUR), 47(2), 1-46.

[14] Kanemitsu, H., & Nakazato, H. (2021, June). KadRTT: Routing with network proximity and uniform ID arrangement in Kademlia. In 2021 IFIP Networking Conference (IFIP Networking) (pp. 1-6). IEEE.

[15] Shukla, N., Datta, D., Pandey, M., & Srivastava, S. (2021). Towards software defined low maintenance structured peer-to-peer overlays. Peer-to-Peer Networking and Applications, 14(3), 1242-1260.

[16] Arunachalam, A. (2021). A Survey of Search Algorithms for Peer-to-Peer File Sharing Applications in Mobile Computing Infrastructure.

[17] Zhu, Y. (2020, October). Supernode selection mechanism based on location information. In Journal of Physics: Conference Series (Vol. 1650, No. 3, p. 032055). IOP Publishing.

[18] Zear, A., Ranga, V., & Gola, K. K. (2024). Network partition detection and recovery with the integration of unmanned aerial vehicle. *Concurrency and Computation: Practice and Experience*, *36*(13), e8048.

[19] Fayyaz, S., Rehman, M. A. U., Khalid, W., & Kim, B. S. (2023). SHM-NDN: A seamless hybrid mobility management scheme for named data mobile ad hoc networks. *Internet of Things*, *24*, 100943.

[20] Zear, A., Ranga, V., & Bhushan, K. (2023). Coordinated network partition detection and bi-connected inter-partition topology creation in damaged sensor networks using multiple UAVs. *Computer Communications*, *203*, 15-29.

[21] Krcmaricic-barackov, P., Ilicin, B., Idalene, K., Llobet-calaf, D., & Raskovic, N. (2024). Ad-hoc Network. U.S. Patent No. 12,021,988. Washington, DC: U.S. Patent and Trademark Office.

[22] Latif, S., Fang, X., Mohsin, S. M., Akber, S. M. A., Aslam, S., Mujlid, H., & Ullah, K. (2023). An enhanced virtual cord protocol based multi-casting strategy for the effective and efficient management of mobile ad hoc networks. Computers, 12(1), 21.

[23] Zahid, S., Abid, S. A., Shah, N., Naqvi, S. H. A., & Mehmood, W. (2018). Distributed partition detection with dynamic replication management in a DHT-based MANET. IEEE Access, 6, 18731-18746.

[24] Abid, S. A., Othman, M., Shah, N., Ali, M., & Khan, A. R. (2015). 3D-RP: A DHT-based routing protocol for MANETs. The Computer Journal, 58(2), 258-279.

[25] Abid, S. A. (2014). An application of 3D logical structure in a DHT paradigm for efficient communication in MANETs (Doctoral dissertation, University of Malaya).

[26] Kousar, R., Alhaisoni, M., Akhtar, S. A., Shah, N., Qamar, A., & Karim, A. (2020). A Secure Data Dissemination in a DHT-Based Routing Paradigm for Wireless Ad Hoc Network. Wireless Communications and Mobile Computing, 2020.

[27] Tahir, A., Abid, S. A., & Shah, N. (2017). Logical clusters in a DHT-Paradigm for scalable routing in MANETs. Computer Networks, 128, 142-153.

[28] Shin, S., Lee, U., Dressler, F., & Yoon, H. (2016). Motion-MiX DHT for wireless mobile networks. IEEE Transactions on Mobile Computing, 15(12), 3100-3113.

[29] Caesar, M., Castro, M., Nightingale, E. B., O'Shea, G., & Rowstron, A. (2006). Virtual ring routing: network routing inspired by DHTs. ACM SIGCOMM computer communication review, 36(4), 351-362.

[30] Awad, A., Sommer, C., German, R., & Dressler, F. (2008, September). Virtual cord protocol (VCP): A flexible DHT-like routing service for sensor networks. In 2008 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems (pp. 133-142). IEEE.

[31] Jain, S., Chen, Y., Zhang, Z. L., & Jain, S. (2011, April). Viro: A scalable, robust and namespace independent virtual id routing for future networks. In 2011 Proceedings IEEE INFOCOM (pp. 2381-2389). IEEE.

[32] Wirtz, H., Heer, T., Hummen, R., & Wehrle, K. (2012, June). Mesh-DHT: A locality-based distributed look-up structure for wireless mesh networks. In 2012 IEEE International Conference on Communications (ICC) (pp. 653-658). IEEE.

[33] Ritter, H., Winter, R., & Schiller, J. (2004, October). A partition detection system for mobile ad-hoc networks. In 2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004. (pp. 489-497). IEEE.

[34] Derhab, A., Badache, N., & Bouabdallah, A. (2005, January). A partition prediction algorithm for service replication in mobile ad hoc networks. In Second Annual Conference on Wireless On-demand Network Systems and Services (pp. 236-245). IEEE.

[35] Wang, K. H., & Li, B. (2002, June). Efficient and guaranteed service coverage in partitionable mobile ad-hoc networks. In Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (Vol. 2, pp. 1089-1098). IEEE.

[36] Hauspie, M., Carle, J., & Simplot, D. (2003). Partition detection in mobile ad-hoc networks using multiple disjoint paths set. In International Workshop on Objects models and Multimedia technologies (p. 15).

[37] Milic, B., Milanovic, N., & Malek, M. (2005, January). Prediction of partitioning in location-aware mobile ad hoc networks. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences (pp. 306c-306c). IEEE.

[38] Sivakumar, B., & Varaprasad, G. (2012). Identification of critical node for the efficient performance in Manet. Editorial Preface, 3(1), 20-25.

[39] Sampaio, S., Souto, P., & Vasques, F. (2015). DCRP: a scalable path selection and forwarding scheme for IEEE 802.11 s wireless mesh networks. EURASIP Journal on Wireless Communications and Networking, 2015(1), 1-22.

[40] Eriksson, J., Faloutsos, M., & Krishnamurthy, S. V. (2007). DART: Dynamic address routing for scalable ad hoc and mesh networks. IEEE/ACM transactions on Networking, 15(1), 119-132.

[41] Caleffi, M., & Paura, L. (2011). M-DART: multi-path dynamic address routing. Wireless communications and mobile computing, 11(3), 392-409.

A Proposed λ_Mining Model for Hierarchical Multi-Level Predictive Recommendations

Yousef S. Alsahafi[1], Ayman E. Khedr[2], Amira M. Idrees[3]
University of Jeddah, College of Computing and Information Technology at Khulais,
Department of Information Technology, Jeddah, Saudi Arabia[1]
University of Jeddah, College of Computing and Information Technology at Khulais,
Department of Information Systems, Jeddah, Saudi Arabia[2]
Faculty of Computers and Information Technology, Future University in Egypt, Egypt[3]

***Abstract*—Delivering the most suitable products and services essentially relies on successfully exploring the potential relationship between customers and products. This immense need for intelligent exploration has led to the emergence of recommendation systems. In an environment where an immense variety exists, it is vital for buyers to own an intelligent exploratory map to guide them in finding their choices. Personalization has proven to be a successful contributor to recommenders. It provides an accurate guide to explore the users' preferences. In the field of recommendation systems, the performance of the systems has been continuously measured by their success in accurate, personalized recommendations. There is no argument that personalization is one key success; however, this research argues that recommendation systems are not only about personalization. Other success factors should be considered in targeting optimality. The current research explores the hierarchy map representing the strengths and dependencies of the recommendation systems pillars associated with their influence level and relationships. Moreover, the research proposes a novel predictive approach that applies a hybrid of content and collaborative filtering recommendation systems to provide the most suitable customer recommendations effectively. The model utilizes a proposed features selection approach to detect the most significant features and explore the most effective associations' schemes for the recommendations label feature. The proposed model is validated using a benchmark dataset by extracting direct and transitive associations and following the identified schematic for the required recommendations. The classification techniques are applied, proving the model's applicability with an accuracy ranging from 96% to 99%.**

***Keywords*—Recommendation systems; data mining; features selection; associations rules mining; classification techniques**

I. Introduction

Delivering the most suitable products and services essentially relies on successfully exploring the potential relationship between customers and products. This immense need for intelligent exploration has led to the emergence of recommendation systems [1]. The recommendation systems continuously apply different intelligent [2] as well as statistical [3] [4] techniques targeting accurate recommendations which directly lead to grasp the customer attraction and gain his positive interaction. The recommendation system types are content-based, knowledge-based, and collaborative filtering [5]. Each type has its own perspective in the analysis task to provide the most suitable recommendations. Content-based type focuses only on the product or service data, while collaborative filtering focuses on users' behavior. Moreover, knowledge-based recommendation systems provide recommendations based on the gathered knowledge of both of them, which could be considered the most challenging type as it should successfully force a homogeneous environment in which knowledge of both products and users could be mutually interpreted.

Although collaborative filtering is the most widely applied approach [6] [7], however, the current study highlights the competence of the knowledge-based approach over other types either the content-based or the collaborative filtering. Knowledge-based approach relies on both users' and products' data as well as these products' ratings. With the applicability to include this divergence of knowledge, the current study argues for higher and satisfied performance of the recommendation systems based on the knowledge-based approach. Although the recommendation systems field has had exceptional success, there are many challenges that should be considered. One of the challenges is to identify suitable methods for dealing with data sources and their issues, such as sparse data. Improving the scalability and efficiency of the recommendations is a continuous challenge in the field [8]. Many research papers have tackled these enhancements for one of the approaches; however, the integration of different approaches is not considered. The idea of collaboration has succeeded in different research to overcome one of the approach challenges with the other approach benefits. In addition, neighborhood techniques such as mining techniques, machine learning, and natural language processing contribute efficiently in the same direction. One of the neighborhood techniques is association rules mining. When these techniques are embedded in the recommendation systems, one of the substantial factors is the quality level of the explored associations. This quality level is heavily affected by the significance of the features [9]. The quality evaluation metrics, including support and confidence, may provide the same measures. However, this study argues that the recommendations could be strongly affected positively by the associations of the significant features. Therefore, one of the current research objectives is to explore the significant features association schemes with the label feature. In addition, considering the significant feature dependencies, even extend the associations' schemes targeting to provide more comprehensive and efficient recommendations. The associations' dependencies could be considered the pillar of building the transitive associations, contributing to more exploration for efficient recommendations.

In this study, the recommendations were based on a set of stages; each of these stages applies an intelligent technique and provides more exploration to the knowledge in-hand. The proposed recommender has three main stages. First, exploring the significant features through weighting methods with the collaboration of statistical methods. Second, exploring the direct association rules, then a second level of associations are explored for more accurate recommendations. Finally, build the semantic recommendation road map through the explored relationship between the features. The evaluation is performed by applying the classification algorithms to confirm the accuracy of the provided recommendations. In addition to the collaboration of the previously mentioned intelligent techniques, the contribution of the paper could be extended to the fact that the extracted map could be applied to different fields. The model is built on processing the provided data and is not restricted to a certain scheme. The remaining of the paper discusses the related work in Section II The proposed model details are discussed in Section III. The experimental study and the evaluation results are presented in Section IV. Finally, the discussion and the conclusion are presented in Section V and Section VI.

II. Related Work

Data mining and machine learning techniques have contributed to a wide range of developing intelligent recommendation systems over the years [10]. Many of the developed models achieved significant advancement in the field. These techniques can contribute from the start of the process [11] in the data pre-processing stage to the final stage of data presentation [12]. It has been proven by many research that the contribution of different models could provide more success to the recommendation process. One successful example is [13], whose aim was to recommend the price of the used cell phones based on the phone status using image processing techniques. Another research [14] followed the collaborative approach, which is also adopted in the current research, to recommend the targeted customers to grant bank loans. A survey was developed [15], which discussed the research for risk assessment and the customers' recommendation models. Another research in the banking sector [16], which reviewed the machine learning techniques that are utilized in detecting fraud in using credit cards. In the field of construction, an accurate budget recommendation was the objective of the research [17], which proposed a model for predicting the overhead cost for commercial projects. Most of the existing recommendation systems follow the seller-centric approach [18]. This approach is restricted to the seller's product availability, which could be more beneficial when extending this perspective to include a wider range of information to the user. While some research focuses on the products that are already bought by the user [19], however, extending the user focus with unintended yet related products could be a step forward in the recommendations.

A research in study [20] has presented recommendation systems using both data of the products and users' information. The proposed models utilized similarity measures for document analysis. However, the models suffered from the non-homogeneity of the items on focus during the analysis of the documents. There was also an issue in the model accuracy due to the non-dealing for the missing values. Another research in [21] depended on the users' response to the products' ratings. These models also applied similarity techniques as well as k-nearest neighbor mining algorithm. The model also suffered from the existence of synonyms while not manipulated. As the model depended on the user rating, the false rating was also considered a serious issue in the accuracy. Moreover, the research [22] focused only on demographic data such as age and gender. The model applied group recommendations using clustering techniques. However, the complete dependence on the demographic data is a major issue that hinders the accuracy. The research in study [23] proposed a content-based model for recommendations, which also depended on the user input. Although the model also highlighted the limited context input with a recommendation to the immense need for including other data perspectives and working with various types of data sources.

The research in study [24] and [25] highlighted the issue of including new items in the data sources whose ratings are not specified. This issue is called the cold start issue. The current research proposed the approach of the collaborative environment between both user and item data with pre-processing mining steps to solve this issue. Additionally, the issue of sparsity is also highlighted in the model proposed in the research [26] [27]. The prior research recommended a clustering approach to resolve this issue. While it could provide more accurate recommendations, however, the current research follows the approach of multi-leveling to further reach higher level of accuracy. The issues of scalability are highlighted in the research [28] and [29], which may be a result of the lack of user trustiness [30] and [31]. Both issues could be viewed as a dilemma situation. Therefore, working on the system performance by leveling up the users' confidence in the system recommendations is one of the motivations in the current research.

III. Proposed Model

The main target of the proposed model is to build the recommendation road map for the business on focus. The model presents the user characteristics, the possible multi-level recommendations for the users on focus as well as an exploratory map for potential users to enlarge the users' segment. The proposed model has four main stages. The following subsections discuss each stage and the included steps in detail.

A. Data Preparation Stage

1) Gather data: Data sources vary. It could be gathered from databases, users' responses, or real-time data from devices that measure and register data in a timely manner. Data could be static or streaming. These different sources of data are processed in different methods due to the difference in nature for these resources. The primary goal of this step "gather data" is the ability to collect the available data that could efficiently contribute to the recommendation main task. Data could be categorized as primary or secondary. The definition of both types varies according to the surrounding environment. One definition is the fact that the primary data is collected directly from the source itself such as the measuring devices while the secondary data is previously gathered for other reasons and could be utilized in the current task. Moreover, secondary data could be previously processed before delivery. In this section, general definition will be presented for both types as there is no restriction in the proposed model on either the data type or category. However, in the experiment section, the types of data will be explicitly highlighted.

Gathering data starts with identifying the potential data sources. The set of data sources (DS) is defined as follows:

$DS = \{ds1, ds2, \ldots.dsj | j \in \mathbb{N} \}$

DS represents the set of data sources that includes all the potential data sources. These sources could be primary or secondary. The sets of primary data sources (PDS) and secondary data sources (SDS) could be defined as follows:

$PDS = \{pds1, pds2, \ldots.pdsi | i \in \mathbb{N} , pdsi \in DS \}$

$SDS = \{sds1, sds2, \ldots.sdsk | k \in \mathbb{N} , sdsk \in DS \}$

$DS = PDS \cup SDS$

The general definition of the set of features for the data source dsk is as follows:

$Features(dsk) = \{fk1, fk2, \ldots. fks \mid k, s \in \mathbb{N}, dsk \in DS\}$

The general definition of the data sources with defining their types is as follows:

$TDS = \{<ds1, typet1>, <ds2, typet2>, \ldots.dsj | j \in \mathbb{N} \}$

2) Integrate multiple resources: The goal of data integration is enriching the model with the data from different data sources. Integrating data leads to a uniform data sources' access. Although the same feature could be a part of different data sources description, however, it could have different types. For example, the gender feature could be nominal in one source (male and female) and integer in another sources (1 and 2). Therefore, the unification process should consider a unifying pre-step for the features type. The following algorithmic steps should be applied.

```
h=1
Repeat
  For each dsh+1 in DS
        For each feature fhs in Features(dsk)
        If  fhs∈ fhs+1  & type (fhs) <> type (fhs+1)
                    unify features' types (fhs, fhs+1)
until h = k-1
```

Then, integrating the data sources requires unifying the descriptive features to allow a single view for all data. The proposed model focuses on transactional data sources. Therefore, the set of the unified criteria for all data sources after integration is as follows:

$Features(DS) = \{fks, fyt, \ldots. fwh \mid k,y,s,t,w,h \in \mathbb{N} \}$

$Features(DS) = Features(ds1) \cap Features(ds2)\ldots.. \cap Features(dsk)$

$Features(DS) = \cap_{l=1}^{k} \, Features\ (ds_l)$

3) Data cleaning: Also called "data wrangling", one of the vital steps in the data analytics process is data cleaning. The hygiene of data heavily affects the whole analytics process. The inconsistencies and existence of errors directly affect results and cause flawed data which could not be configured and leads to unreliable results. The following cleaning steps are performed.

Remove direct redundancies: two records are directly redundant when they have the same values for all of their attributes.

Consequently, the dataset dsk` represents dsk after removing the redundancies. dsk` could be defined as follows:

$dsk` = \{tn,\ldots tm\} | dsk` \subseteq dsk \mid \exists\ tg, tg \in dsk, tg \notin dsk`, tg = tt, tt \in dsk`$

Remove outlier: a record is defined to be a direct outlier if the values of its attributes do not match the standard values. It could be due to incorrect calculation or incorrect data entry. For example, a student record is an outlier if its id does not follow the faculty standards such as to be 4 digits instead of 6 digits and does not have the year as required.

The set of outliers' records of the set dsk could be identified as follows:

$Outlier(dsk`) = \{tn,\ldots tm\} | Outlier(dsk`) \subseteq dsk`$

Consequently, the dataset dsk`` represents dsk` after removing the outliers. It could be defined as follows

$dsk`` = dsk` \cap Outlier(dsk`)$

Remove contradictions: two records are contradictory if they provide conflicting data. For example, one record may include the birth date while the age does not match with this date, or the salary of the employee is less than the required salary taxes.

The set of contradictions' records of the set dsk could be identified as follows:

$Contradiction(dsk``) = \{tn,\ldots tm\} | Contradiction\ (dsk``) \subseteq dsk``$

Consequently, the dataset dsk``` represents dsk`` after removing the Contradictions. It could be defined as follows

$dsk``` = dsk`` \cap Contradiction(dsk``)$

Adapt missing values: different methods could be applied for missing values detection. A simple solution is to remove the record with the missing value while a more informative solution is to predict this value either with statistical calculation or with intelligent techniques.

B. Explore the Features' Significant Level

The goal of this stage is exploring the most informative features in the dataset. Different methods could contribute to this stage. Filtering, wrapper, and embedded categories are introduced in different research [32] [33]. The proposed model does not restrict the contribution to a certain category or method, rather, the selection of the contributing methods is based on the applying experiment. In this research, the contributing methods will be stated in the experiment section.

1) Set weighting measures: The set of the types of weighting methods is defined as follows:

WT = {T1, T2, T3}

The set of the weighting methods that belong to a one of these types is defined as follows

WM (T1) = {wT11, wT12…, wT1x}

WM (T2) = {wT21, wT22…, wT2y}

WM (T3) = {wT31, wT32…, wT3z}

The set of all weighting methods are determined which is defined:

W = {w1, w2, ….wz| z $\in \mathbb{N}$, wz $\in$ [WM (T1) | WM (T2) | WM (T3)]}|

W = WM (T1) ∪ WM (T1) ∪ WM (T1)

The weighting threshold is identified to be equal 50% (0.5)

2) Apply weighting techniques: Each weighting measure is applied on the dataset and the features weightings are determined. The set of weighted features for each weighting technique is defined as

Wgt (wz) = {<fwh, wgt_valwhz>| wz $\in$W, fwh$\in$ Features (DS), wgt_valwhz>=0, wgt_valwhz <=1}

Reversibly, each attribute is weighted by a set of weighting measures, the set of all weights for a defined attribute is as follows:

Fgt (fwh) = {<wz, wgt_valwhz>| wz $\in$W, fwh$\in$ Features (DS), wgt_valwhz>=0, wgt_valwhz <=1}

The matrix of all features' weights with respect to the weighting measures is as follows:

FT-WT = $\bigcup_{k=1}^{j}\bigcup_{l=1}^{i} \mathrm{AttWt}\,(\mathrm{j},\mathrm{i})$ where j is the feature index and i is the weighting measure index

3) Explore features' significant level: Exploring the significance of the datasets features is performed through the following steps. The exploration task is the pillar to identify the suggested associations' schemes which relies on the features according to their strength in affecting the prediction task performance. This section clarifies the steps for exploring the significance level of the features while the following sections continue to discuss how the prediction process is performed.

The first step is building the matrix for each feature with respect to the label feature y (FT-WT). This step results in identifying x-1 matrices where x is the number of features.

FT-WT (x) = $\begin{bmatrix} \mathrm{ft}_{z1} & \cdots & \mathrm{ft}_{y1} \\ \vdots & \ddots & \vdots \\ \mathrm{ft}_{zj} & \cdots & \mathrm{ft}_{yj} \end{bmatrix}$ where z ≠ y, z, y ≠ j

The matrix of each of the contributing weighting measures h highlights the contribution of a feature x on the consistency of feature y. This impact is determined by the weight of x that is determined by h given that y is the label attribute. The matrix of the weighting measure h is constructed as follows:

After applying weighting technique h on feature x with label feature y, the weights ranges are normalized by setting the value of 1 for the weights of range from 0 to <0.3, 2 for weights of 0.3 to <0.6, 3 for the weights of range 0.6 to 1, and 4 otherwise. The matrix wtxy element is the determined normalized value while the value of wtyx equals to the multiplicative inverse of the normalized value 1/ wtxy. The following matrix represents the normalized matrix of the weighting measure i for all attributes weights with considering y as the label attribute.

WTy (h) = $\begin{bmatrix} \mathrm{Nwt}_{11} & 1/\mathrm{Nwt}_{12} & \cdots & \mathbf{1}/\mathrm{Nwt}_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ & \mathrm{Nwt}_{1x} & \mathrm{Nwt}_{2x} & \\ & \mathrm{Nwt}_{1j} & \mathrm{Nwt}_{2j} \cdots & \mathrm{Nwt}_{jj} \end{bmatrix}$

Following the same approach, a set of matrices are built for each attribute while considering the significance determinants (SD). The current research follows the same approach of [12]. The research adopted only three measures for λ with an acceptable accuracy percentage which confirmed the applicability of the proposed approach. However, the current research extends the contributing measures seeking targeting to raise the consistency accuracy level. The current research extends the significance determinants for each feature to include the minimum and maximum weighting value among all the weighting measures, the mean and median value of all weightings, the upper and lower inter quartile range, the mean of these determined upper and lower values of all the weightings, and the value of mean for λ. Finally, the features' level of consistency is determined according to the number of consistent determinants. In case two features have the same count of consistent determinants, then the weighting value is considered. The following rule applies for constructing the features consistency ordered set.

The set of SD for each feature ftx is as follows:

SD (ftx) = {sdt,…sdr}

The ordered set of features based on consistency level = {ftx,… fty} where (|SD (ftx)| >= |SD (fty)| & min(Fgt (ftx)) > min(Fgt (fty)))

C. Associations Exploration Stage

The core stage of the proposed model is the mining phase. Mining phase employs different methods with a variety of nature and analysis procedures targeting to recognize the non-trivial embedded patterns. Data mining is continuously contributing to the business field for many aspects. One of these aspects which is the objective of the current research is the customer. Data mining could apply collaborating techniques targeting for customers beneficiary such as intelligent advertising, recommendations, detect fraudulent, and others. In the proposed model, associations rules mining, which is one of the data mining techniques, is employed to detect the embedded schemes in data. These schemes will then be utilized to explore both direct and indirect level associations. Then, the explored associations are employed to support the following recommendation phase. The following subsections discuss the main steps of the mining stage.

1) Determine direct associations: In this step, associations rules mining algorithm is applied to the dataset in order to govern the customers' perceptions and willingness. Many algorithms are introduced for this task; therefore, the following definition is generic. However, a defined algorithm will be stated in the experimental study with a clear justification for the reason for this selection. At this step, there are no restrictions in the associations' generation. The following steps prune the generated rules to explore the first step in detecting the most vital associations to the assigned task which is based on the metrics' threshold.

∃ Assc = Ass (ftxi → ftwi) ∧ fxi , ftwi ∊ Features(DS) ∧ Support (Assc) >= SU_Threshold ∧ Confidence (Assc) >= Con_Threshold → Add_ Assc (ftwi, Pre-Accept (ci))

2) Identify direct associations schemes: Identifying the targeted schemes follows the explored significance level of the dataset features. The first step is prioritizing the influencing features with respect to the required label feature based on the influencing level. The level of influence is determined according to the consistency degree of the feature on focus with the label feature based on the previous λ measures. The second step is identifying the associations schemes to have the influencing features as premise and the label feature as conclusion. The final step is pruning the detected schemes with considering only the schemes that have the influencing features with consistency above the determined threshold. These steps are formally described as follows:

∀ ftx, ftx ∈ Features (DS), label (ftx) → ∃ fty, influence (fty, ftx)

∀ ftz, ftx, influence (fty, ftx), influence (ftz, ftx), Value (significance (fty)) > Value (significance (fty)) → ordered_significance (ftx) = {ftz, ftz,….. | fty ≺ ftz }

∀ ftz, ftx, ftz ∈ ordered_significance (ftx), significance (ftz) > consistency_Threshold →

Schemes (ftx) = Schemes (ftx) ∪ association (ftz → ftx)

3) Filter direct associations: Based on the assigned schemes, this step focuses only on a subset of the associations that satisfy the required schemes as well as the minimum threshold of the performance metrics, support and confidence. The filtering step is described as follows:

Target = { ftx,,…., fty} where ftx ∊ Features (DS), fty ∊ Features (DS)

∀ ftx, ftx, ∈ Target, ASS = association (ftz → ftx), Support (association (ftz → ftx)) >= Min_Support, Confidence (association (ftz → ftx)) >= Min_Confidence →

4) Identify indirect associations schemes Based on significant level: In this step, an argument is highlighted for the ability to gain higher recommendation accuracy by considering transitive associations. A feature may be indirectly significant to the label feature is the following rule applies.

∀ ftx, fty, ftx, ∈ Target, ASS = association (ftz → ftx), Support (association (ftz → ftx)) >= Min_Support, Confidence (association (ftz→ ftx)) >= Min_Confidence , ASS = association (fty→ ftz), Support (association (fty→ ftz)) >= Min_Support, Confidence (association (fty→ ftz)) >= Min_Confidence → Indirect_ASS(ftx) = Indirect_ASS(ftx) ∪ ASS (fty→ ftz → ftx)

Therefore, extending the previous identifying the associations schemes step is performed in the current step. After the influencing features are identified, then they are considered as label features and the same previous steps are performed. Each influencing feature is considered a label, then schemes are determined, and associations are detected with the same previous description and pruning rules. The idea of this step is extending the opportunities for higher performance recommendations based on the existence of indirect influencing features. A feature x has an influence on feature y when there exist a third feature z that is influenced by x and, at the same time, influencing y. although the origin definition of the indirect relation states that this influence exists only through the feature z, however, this relation restriction is not considered in the current relation definition. This is because the main objective of extending the relationship is to consider the related factors for highlighting more opportunities in exploring different opportunities for the most available as well as applicable recommendations with highest accuracy.

5) Extract direct & indirect associations: The final step in the mining stage is extracting the relevant associations to the assigned task. based on the task, the association scheme is determined, and its corresponding associations rules are explored. These rules are then utilized in the recommendation stage.

D. Explore Recommendations Road Map

The deliverable of the proposed model is constructed in this stage. The road map of the recommendations is developed by applying a set of steps as follows.

1) Set recommendations threshold: One of the recommendations threshold (RTH) determination methods could be a direct method as it could be decided by the responsible. Other more intelligent methods could be applied such as the proposed method in [34]. It varies according to the applied domain. Some domains require a high level of accuracy such as healthcare or tourism. Other domains do not require accuracy to be critical such as restaurants or groceries. In this stage, according to the determined recommendations threshold, the associations are pruned while the recommendation threshold matches with the associations' minimum consistency level of the premises in the association rules. the following applied in this step

∀ ass in ASS, association (ftz → ftx), | consistency (ftz) > RTH → RecASS = RecASS ∪ association (ftz → ftx)

2) Detect recommendations peaks: The recommendation peak is defined as the point that has either the maximum or minimum features values with all other features with various values. Determining the recommendations peaks of a determined features ftx is performed by capturing the set of spike points of the feature x Spike (ftx). A value is identified as a spike point of a feature fx if it is associated with most of the values of the significant feature fy in the association rules set of fx and fy. Formally speaking, Spike (ftx) is constructed by the following rule.

∀ ftx, ∀ fty , ∀ Vali(fty), ∃ association (Vali (fty) → Vali (ftx)) ∈ ASS, → ∀ Vali (fty), feature(x,y) = feature(x,y) ∪ {<Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> }

Max_Count(association (Vali (fty) → Vali (ftx))) = Vali (ftx) where ∃ <Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> and ∀ Val(fty), ∀ j, j >=0, j < | ASS| , Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> > Count (association (Valj(fty) → Valj (ftx))), Valj (fty) , Valj (ftx)>

The spike of the feature ftx with respect to the significant feature fty is determined as: Spike(fty, ftx) = Vali (ftx)

The set of all spike points of the feature ftx is determined as: Spike (ftx) = $\cup_{y=0}^{|\text{Features(dsk)}|}$ Spike (ft_y , ft_x)

IV. Experimental Study

This section confirms the applicability of the proposed model. The section discusses the applied experiments and the evaluation results. The experiments dataset includes books data which aim is for books recommendations. The dataset describes the ratings of 271379 books from 278858 users with a total of 1048755 rating records [35]. The total number of features describing the dataset are fourteen features. As mentioned in the dataset source, the dataset was collected from Amazon and the books were purchased online. The dataset was distributed over three files. The first file describes the books data (ISBN, Title, Author, Publication Year, and Publisher). The total records were a total of 271289 records. The file included incomplete records, a set of these records were missing most of the features' values and the remaining set had only one of the features with a missing value. The first set of records was removed from the dataset, this set included 38876 incomplete records which were removed. The second set included 30341 records. These values were simply retrieved as most of the features' values exist. The second file represents the users' data which are described by the user id, location, district, country, and age. The file included the data of 276271 users. A set of 4563 records had incomplete residence data while a set of 109163 records had incomplete age data. This means that a set of 162545 records had complete data of users. According to this distribution, the experiment has been limited to 162545 users and 232413 books with a total rating records equal to 645212. After preparing the data in their original files, a simple code was developed to integrate the data into a unified structure.

The next step is identify the contributing weighting measures. Eight weighting measures are included in the experiments, they are Principle Component Analysis (PCA), Information Gain (IG), Information Gain Ratio (IGR), Support Vector Machine (SVM), and Correlation. The measures were applied and Table I presents the weighting results for the features above the threshold for each measure.

TABLE I. Weighting Measures Results for the Features above the Threshold

IG		IGR		Correlation	
rating	0.896	age	0.788	rating	0.959
rating	1.134	Pub. year	0.584	publisher	1.019
author	1.107	year	0.620	age	0.912
age	0.830	author	0.560	Pub. year	0.746
Pub. year	0.783	rating	0.812	country	0.798
				author	0.794

SVM		PCA	
country	0.762	publisher	0.959
author	1.000	author	1.019
age	0.889	age	0.912
Pub. year	0.739	Pub. year	0.746
		rating	0.798
		country	0.794

The next step is to detect the attributes' consistency. Each of the weighting technique contributed in all attributes' consistency detection. Five iterations are performed for each attribute with a total of seventy iterations. The percentage matrix is calculated by determining the product values and the Eigen vector value, then at least six of the λ determents determines should be consistent in order to consider the attribute to be consistent. Table II and Table III presents the Eigen vector and λ determents for the Author attribute respectively.

TABLE II. Eigen Vector for Author's Feature

	SVM	CRL	IGR	DEV	IG	P	EV
SVM	1.000	0.900	0.250	0.500	0.100	0.007	0.019
CRL	0.657	1.000	4.475	0.650	0.140	0.228	0.679
IGR	0.253	0.224	1.000	0.850	0.900	0.051	0.152
DEV	0.899	0.900	0.980	1.000	0.100	0.050	0.149
IG	0.100	0.146	0.655	0.106	1.000	0.001	0.700

TABLE III. Λ Statistics for Author's Feature

		L	C.I	C.R	C/NC
Lmax		5.9890	0.2473	0.2208	C
Lmean		4.8490	0.0378	0.0337	C
Lmedian		5.8240	0.2060	0.1839	C
Lmin		1.0550	0.9863	0.8806	NC
Range		4.6350	0.0913	0.0815	C
Standard Deviation		1.2990			
Inter-Quartile range	0.1770	1.2058	-1.0766	0.167	C
	0.4830	1.1293	1.0083	0.413	NC
Mean (rang,UpperQ)		5.2620	0.0655	0.0585	C
Mean (rang,LowerQ)		3.9420	0.2645	0.2362	NC
Mean (range, mean)		4.5620	0.1095	0.0978	C

The same calculations are performed for the eleven feature. Four attributes are considered consistent, they are (Author, Country, Age, Year, and rating). As a final step, with considering the weights of these features to the five weighting measures, it is detected that four of the consistent attributes are significant, they are (Author, Country, Age, and rating). Therefore, these are the attributes that will contribute to the associations' exploration stage. The next phase is identifying the required associations' schemes. Following the selected features, the associations' schemes are as follows: Age → Author, Author → Age, Country → Author, Author → Country, and Author → rating. As the main label feature is the rating, therefore, the indirect associations' schemes are detected to be Age→ Authors → rating, Country → Author → rating. The final step is to identify the recommendations map. The explored map has been discussed with professors in the literature field who confirmed the strong associations and dependencies. A second evaluation step is applying a set of classification techniques over the dataset with considering only the features that are contributing in the association. The first experiment is applying five classification techniques namely K nearest Neighbor (KNN), Naïve Bayes (NB), Enhanced ID3 (EID3) and the contributing features are the age and authors with the rating as the label feature. The second experiment includes the country and authors with also the rating as the label feature. Moreover, a third experiment is conducted that includes the three features. The results reveals the high accuracy percentage which confirms the applicability of the explored associations and consequently the proposed model. Table IV demonstrates the average evaluation results for the classification experiments.

TABLE IV. CLASSIFICATION RESULTS

Criteria	**KNN**	**NB**	**EID3**
Accuracy	95.96 %	97.98 %	98.99
Classification Error	4.04 %	2.02 %	1.01
Kappa Statistic	0.937	0.969	0.984
Weighted mean recall	96.17 %	97.94 %	99.28
Weighted mean precision	95.56 %	97.52 %	98.85
correlation	0.928	0.956	0.972

V. DISCUSSION

According to the results, the proposed approach succeeds to perform effective feature selection with a generality to be applied in different fields. The presented performance measures revealed the success of the approach, however, the missing percentage in the classification task could be due to the limited number of attributes in the dataset. Datasets with higher number of attributes could results to higher performance.

VI. CONCLUSION

This research proposed an intelligent model that aims to build the recommendation road map for business. The model explores the possible multi-level recommendations for the users on focus as well as an exploratory map for potential users to enlarge the users' segment. The proposed model had four main stages which applies mining techniques in the preprocessing and explorations stages, features selection techniques for exploring the significant features and statistical techniques to more illumination for these features in order to provide the most accurate recommendations base on these features associations. The proposed model has been evaluated by applying the detailed stage on a dataset of books. The results have been evaluated by experts and by applying classification models and evaluate the classification results. The results revealed a success in the classification that ranged from 96% to 99% which ensures the model applicability. More future work could be further applied to the proposed model. Another future direction to conduct more experiments could be conducted from different fields. Investigating more methods for feature selection could be further developed. Another enhancement direction is including the users' comments and apply text mining techniques with embedding the keywords as features.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-23-DR-13). Therefore, the authors thank the University of Jeddah for its technical and financial support.

REFERENCES

[1] A. Al Mazroi, A. E. Khedr and A. M. Idrees, "A Proposed Customer Relationship Framework based on Information Retrieval for Effective Firms' Competitiveness," Expert Systems With Applications, vol. 176, 2021.

[2] Y. Helmy, A. E. Khedr, S. Kholief and E. Haggag, "An Enhanced Business Intelligence Approach for Increasing Customer Satisfaction Using Mining Techniques," International Journal of Computer Science and Information Security, vol. 17, no. 4, 2021.

[3] A. M. Idrees and W. H. Gomaa, "A Proposed Method for Minimizing Mining Tasks' Data Dimensionality," International Journal of Intelligent Engineering and Systems, vol. 13, no. 2, 2020.

[4] M. Atia, M. Mahmoud, M. Farghally and A. M. Idrees, "A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data," Future Computing and Informatics Journal, vol. 6, no. 2, 2021.

[5] A. M. Idrees, A. E. Khedr and A. A. Almazroi, "Utilizing Data Mining Techniques for Attributes' Intra-Relationship Detection in a Higher Collaborative Environment," International Journal of Human-Computer Interaction, 2022.

[6] A. M. Idrees and F. K. Alsherif, "A Collaborative Evaluation Metrics Approach for Classification Algorithms," Journal of Southwest Jiaotong University, vol. 55, no. 1, pp. 1-14, 2020.

[7] F. Abogabal, S. M. Ouf and A. M. Idrees, "Proposed framework for applying data mining techniques to detect key performance indicators for food deterioration," Future Computing and informatics journal, vol. 7, no. 2, 2022.

[8] M. Tamer, A. E. Khedr and S. Kholief, "A Proposed Framework for Reducing Electricity Consumption in Smart Homes using Big Data Analytics," Journal of Computer Science, vol. 15, no. 4, 2019.

[9] M. Attia, M. A. Abdel-Fattah and A. E. Khedr, "A proposed multi criteria indexing and ranking model for documents and web pages on large scale data," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, 2022.

[10] A. E. Khedr, A. I. El Seddawy and A. M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS," International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 2, no. 6, pp. 111-118, 2014.

[11] D. H. A. Hassouna, A. E. Khedr, A. M. Idrees and A. I. ElSeddawy, "Intelligent Personalized System for Enhancing the Quality of Learning," Journal of Theoretical and Applied Information Technology, vol. 98, no. 13, pp. 2199-2213, 2020.

[12] A. M. Idrees, A. I. ElSeddawy and M. O. Zeidan, "Knowledge Discovery based Framework for Enhancing the House of Quality," International

Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 7, pp. 324-331, 2019.

[13] A. M. Idrees and S. Taie, "Online Price Recommendation System for Shopping Used Cell Phones," Research Journal of Applied Sciences, Engineering and Technology, vol. 13, no. 1, pp. 15-23, 2016.

[14] A. M. Idrees and A. E. Khedr, "A Collaborative Mining-Based Decision Support Model for Granting Personal Loans in the Banking Sector," International Journal of E-Services and Mobile Applications (IJESMA), vol. 14, no. 1, pp. 1-23, 2022.

[15] M. sharaf, S. Ouf and A. M. Idrees, "Risk Assessment Approaches in Banking Sector-A Survey," Future Computing and Informatics Journal, vol. 8, no. 1, 2023.

[16] N. S. Elhusseny, S. M. Ouf and A. M. Idrees, "Credit Card Fraud Detection Using Machine Learning Techniques," Future Computing and Informatics Journal, vol. 7, no. 1, 2022.

[17] A. H. Z. Hassan, A. M. Idrees and A. I. Elseddawy, "Neural Network-Based Prediction Model for Sites' Overhead in Commercial Projects," International Journal of e-Collaboration, vol. 19, no. 1, 2023.

[18] S. Fazeli, H. Drachsler, M. Bitter-Rijpkema, F. Brouns, W. van der Vegt and P. B. Sloep, "User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg," IEEE Transactions on Learning Technologies, vol. 11, no. 3, 2018.

[19] A. M. Idrees and E. Shaaban, "Reforming home energy consumption behavior based on mining techniques a collaborative home appliances approach," Kuwait Journal of Science, vol. 47, no. 4, 2020.

[20] Z. Yang, B. Wu, K. Zheng, X. Wang and L. Lei, "A survey of collaborative filtering-based recommender systems for mobile internet applications," IEEE Access, vol. 4, 2016.

[21] C. He, D. Parra and K. Verbert, "Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities," Expert Systems with Applications, vol. 56, 2016.

[22] M. Elahi, F. Ricci and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," Computer Science Review, vol. 20, no. C, 2016.

[23] N. Pereira and S. L. Varma, "Financial planning recommendation system using content-based collaborative and demographic filtering," Smart Innovations in Communication and Computational Sciences, Part of the Advances in Intelligent Systems and Computing, vol. 669, 2018.

[24] V. R. Revathy and A. S. Pillai, "A proposed architecture for cold start recommender by clustering contextual data and social network data," Computing, Communication and Signal Processing, Part of the Advances in Intelligent Systems and Computing, vol. 810, 2018.

[25] Y. Zhu, J. Lin, S. He, B. Wang, Z. Guan, H. Liu and D. Cai, "Addressing the item cold-start problem by attribute-driven active learning," IEEE Transactions on Knowledge and Data Engineering, vol. 32, 2020.

[26] S. Ahmadian, M. Afsharchi and M. Meghdadi, "A novel approach based on multi-view reliability measures to alleviate data sparsity in recommender systems," Multimedia Tools and Applications, vol. 78, 2019.

[27] A. K. Sahu and P. Dwivedi, "User profile as a bridge in cross-domain recommender systems for sparsity reduction," Applied Intelligence, vol. 49, no. 7, 2019.

[28] H. Varudkar, S. M. Deosthale and J. Mehta, "Collaborative recommendation system based on Hadoop," Global Journal for Research Analysis, vol. 6, no. 1, 2016.

[29] V. Koshti, N. Abhilash, K. S. Gill, N. Nair, M. B. Christian and P. Gupta, "Online partitioning of large graphs for improving scalability in recommender systems," Computational Intelligence: Theories, Applications and Future Directions, vol. 2, 2019.

[30] H. Feng and T. Tran, "Context-aware approach for restaurant recommender systems," Encyclopedia of Information Science and Technology, Fourth Edition, 2019.

[31] M. Singh, H. Sahu and N. Sharma, "A personalized context-aware recommender system based on user-item preferences," Data Management, Analytics and Innovation, vol. 140, 2019.

[32] N. Hegazy, M. Khafagy and A. E. Khedr, "Proposed Approach for Academic Paper Ranking Based on Big Data and Graph Analytics," Journal of Theoretical and Applied Information Technology, vol. 101, no. 2, 2023.

[33] H. Elmasry, A. E. Khedr and H. M. Abdelkader, "Enhancing the Intrusion Detection Efficiency using a Partitioning-based Recursive Feature Elimination in Big Cloud Environment," International Journal of Advanced Computer Science and Applications,, vol. 14, no. 1, 2023.

[34] E. Hikmawati, N. U. Maulidevi and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," Journal of Big Data, vol. 8, no. 146, 2021.

[35] O. Monk, "https://www.kaggle.com/datasets/saurabhbagchi/books-dataset," 2004. [Online]. [Accessed 9 2023].

Proposal of OptDG Algorithm for Solving the Knapsack Problem

Matej Arlović, Tomislav Rudec, Josip Miletić, Josip Balen
Faculty of Electrical Engineering-Computer Science and Information Technology
J. J. Strossmayer University of Osijek, Osijek, Croatia

***Abstract*—In a computational complexity theory, P, NP, NP-complete and NP-hard problems are divided into complexity classes which are used to emphasize how challenging it is to solve particular types of problems. The Knapsack problem is a well-known computational complexity theory and fundamental NP-hard optimization problem that has applications in a variety of disciplines. Being one of the most well-known NP-hard problems, it has been studied extensively in science and practice from theoretical and practical perspectives. One of the solution to the Knapsack problem is the Dantzig's greedy algorithm which can be expressed as a linear programming algorithm which seeks to discover the optimal solution to the knapsack problem. In this paper, an optimized Dantzig greedy (OptDG) algorithm that addresses frequent edge cases, is suggested. Furthermore, OptDG algorithm is compared with the Dantzig's greedy and optimal dynamically programmed algorithms for solving the Knapsack problem and performance evaluation is conducted.**

***Keywords*—Dynamic programming; Dantzig algorithm; greedy algorithm; knapsack problem; linear programming; NP-Problem; optimization; OptDG**

I. Introduction

In computational complexity theory P, NP, NP-complete and NP-hard terminology is used to denote how difficult it is to find solutions to a specific type of problem. To begin, it has to be determined whether the problems are P, NP, NP-complete, or NP-hard. A problem is P if it can be resolved in polynomial time [1]. If the suggested solution can be checked in polynomial time, the problem is NP [2]. A problem is NP-complete if it is in NP, and all other NP problems can be reduced to it in polynomial time [1]. This implies that if a solution is discovered for an NP-complete problem, it can be applied to all other NP problems. A problem is NP-hard if it is at least as difficult as the most difficult NP problems. This implies that if a solution is discovered for an NP-hard problem, it can be applied to all NP-complete problems [3]. The Knapsack problem is a well-known computational complexity theory and fundamental NP-hard optimization problem that has applications in a variety of disciplines, including operations research, cryptography, and combinatorial optimization. If a collection of items is given, each of which has a weight and a profit, and a knapsack that can only carry a certain quantity of weight, the Knapsack problem is present. The objective is to determine which subset of items can enhance the total profit of the knapsack without exceeding its carrying capacity. The 0-1 Knapsack is a variant of the Knapsack problem in which a knapsack with a specific capacity *c* and *n* items with weights $t_1, t_2, ..., t_n$ and profits $p_1, p_2, ..., p_n$ is given. The main goal is to find the item combination that maximizes the total profit of the knapsack while adhering to the weight limit, with the additional constraint that each item can only be turned on or off once, i.e., each item can only be used once. The theory can be formulated and Integer Linear Program (ILP) shown below:

$$maximize \sum_{i=1}^{n} p_i x_i \tag{1a}$$

$$subject\,to \sum_{i=1}^{n} t_i x_i \leq c \tag{1b}$$

$$x_i \in \{0,1\}, \forall i \in \{1,2,3,...,n\} \tag{1c}$$

The *n* binary variables $x_1, x_2, ..., x_n$ are decision variables that determine whether or not an item is placed in the knapsack. As part of the input, the problem parameters $c, n, t_1, t_2, ..., t_n$ and $p_1, p_2, ..., p_n$ are given. The 0-1 knapsack problem is NP-hard, indicating that exact solutions to problems with enormous inputs may be intractable. To tackle the NP-hard knapsack problem, several algorithms, including dynamic programming, branch and bound, and genetic algorithms, have been proposed. However, these algorithms only perform well with smaller quantities, i.e., in particular types of problems. As the numbers increase, it becomes nearly impossible to solve the problem.

The 0-1 Knapsack problem is a major optimization problem strongly related to a number of other major optimization problems. For example, by solving an equivalent 0-1 Knapsack problem instance, instances of the binary integer programming, and the bounded and unbounded knapsack problems can be solved. Furthermore, the 0-1 Knapsack problem arrises as a column generation subproblem of the cutting stock problem and is a particular instance of a variety of problems such as the knapsack problem with conflicts [4], [5], the traveling thief problem [6] and the multidimensional knapsack problem [7].

II. Related Work

The Knapsack problem is among the most actively researched topics in combinatorial optimization. In recent years, a vast number of the Knapsack problem variants have been addressed. The 0-1 Knapsack is the most well-known Knapsack problem, and it has been intensively studied for decades. These studies have yielded an abundance of theoretical, practical, and algorithmic findings that have, to some extent, saturated this particular field [8]. According to research conducted by study [9], the Knapsack problem is listed as one of the most popular algorithmic problems, as well as the and the second most popular problem in the NP-hard category. In this section, a brief summary of the literature regarding the 0-1 Knapsack problem is presented.

The most effective algorithms for solving the 0-1 Knapsack problem employ branch-and-bound, dynamic programming, or hybrid approaches that combine the both. Over time, a sequence of advancements resulted in the creation of MT1 [10], MT2 [11], Expknap [12], Minknap [13], and Combo [14] algorithms. The Combo algorithm is the most significant algorithm among several others. Despite being published over twenty years ago, it remains the best-known and most effective algorithm. It can solve the majority of problem instances within a few seconds. Although it shares similarities with the Minknap algorithm, Combo introduces new techniques when faced with a large number of dynamic programming states. The Combo algorithm uses a number of interesting methods, one of which is adding valid inequalities (cardinality constraints) to the formulation of integer programming, which are then relaxed to make more accurate dual bounds [15]. Previous research [12], [14], [16], [17] consistently demonstrates that Combo is typically the fastest algorithm among the five algorithms studied. In this paper, an empirical hardness model, using Combo's running time as an indicator of problem instance difficulty, is adopted. Given Combo's ability to efficiently solve most large problem instances within seconds and the (weak) NP-hardness of the Knapsack problem, researchers have expressed interest in identifying more challenging instances. While it is generally believed that such hard instances exist, the process of discovering them and understanding the key features contributing to their difficulty remains unclear. Research on instances of the 0-1 Knapsack problem often focuses on instances with considerably large coefficients, such as those exhibiting exponential growth relative to the number of items (n). Evaluating the practical difficulty of these instances does not involve using existing algorithm implementations like Combo, as most implementations support only 32-bit or 64-bit integers. Instead, these instances are examined against hypothetical sets of algorithms, considering different assumptions about the strength of the bounds employed by these algorithms. Noteworthy papers in this category include [18], [19], and [20], which describe challenging problem instances with large coefficients designed for various hypothetical algorithms. [21] and [22] introduced new branching strategies for Branch-and-Bound (B&B) methods. The analysis of how item profits or weights respond to perturbations was explored by [23], [24] and [25]. [26] focused on a specific sensitivity analysis called tolerance analysis, which can be performed in amortized time $O(c\, log\, n)$ for each item. Recent advancements in enhancing existing Fully Polynomial Time Approximation Schemes (FPTAS) were made by [27] and [28]. A sensitivity study of greedy heuristics was provided by [29] for the 0-1 knapsack problem and the subset sum problem. [30] addresses the discounted 0-1 knapsack problem (DKP), an extension of the classical knapsack problem where items are grouped into threes, and only one can be picked from each group. They suggest preprocessing and reducing the problem size before using dynamic programming using exact and heuristic fixation methods. These strategies greatly reduce DKP instance solution time, and the authors provide a new collection of more difficult instances for further evaluation.

III. Methods to Solve 0-1 Knapsack Problem

Dynamic programming is a potent methodology and optimization technique in mathematics and computer programming that allows problems to be broken down into smaller subproblems and their solutions to be saved to prevent redundant computations [31]. Dynamic programming yields accurate results for locating optimal solutions, but its computational complexity frequently renders it impractical, necessitating the use of heuristics as an alternative method. In experiments conducted in this research, adapted [32] and Dantzig's greedy algorithm, are used.

A. Dantzig's Greedy Algorithm

The Dantzig's greedy algorithm can be used to solve the 0-1 Knapsack problem. The simplex method is a method for resolving linear programming issues, and the Dantzig's greedy algorithm is a particular kind of linear programming algorithm based on it. The algorithm, which can be expressed as a linear program, seeks to discover the optimal solution to the knapsack problem. The fundamental idea behind Dantzig's method is to build a table with columns that indicate the weight capacity of the knapsack and rows that represent items. The profit of each cell in the table represents the maximum profit that can be obtained for a specific weight capacity using the items utilized up until that point. The table is filled from bottom to top. The initial step of the technique is to add the profit of the first item to the table's first row. The algorithm chooses the maximum profit for each cell based on a comparison between the weight capacity profits of the knapsack with the current item and the knapsack without it for each subsequent row. The maximum possible profit for the specified weight restriction and item count is the last number in the last row of the table. Using a straightforward dynamic programming technique, Dantzig's greedy algorithm can be implemented with a time complexity of $O(nC)$, where n is the number of items, and C is the maximum weight the knapsack can carry. Due to the need to store a table with n rows and C columns, the technique has an $O(nC)$ time complexity. Due to its assumption that each item can only be used once, Dantzig's approach can only be used to solve the 0-1 Knapsack problem. As a result, the unbounded knapsack problem cannot be resolved using it [33]. Algorithm 1 displays Dantzig's greedy algorithm.

B. OptDG Algorithm

According to Danzig's heuristic approach, it has been assumed that items from largest to smallest are placed in a knapsack, based on the $\frac{p_i}{t_i}$ (profit-over-weight) ratio. The main challenge is to increase the total profit of placing items into the knapsack in this manner by replacing a specific item or more items, if possible. The following theorem is demonstrated: if an item with profit p_i and weight t_i is removed and then one or more items are inserted to the right of the removed item in the sorted sequence, a higher weight than the deleted item should be chosen in order to to attain a higher profit. In other words, if an item is removed and replaced with another item to its right while maintaining or decreasing the total weight in the knapsack, the overall worth will either remain the same or it will drop. This theorem will be useful in two situations: first while looking for a better profit, it will be obvious that either one of the previously skipped items must be included

Algorithm 1 Dantzig's greedy algorithm for solving 0-1 Knapsack problem

```
Require: A set of items $items$, and a knapsack capacity $capacity$
Ensure: A knapsack that is packed to its full carrying capacity with items that have the greatest possible profit

  function SolveKnapsackGreedy(items, capacity)
      Sort $items$ by item $p/t$ in decreasing order
      $knapsack\_items \leftarrow []$
      $total\_p \leftarrow 0$
      $total\_w \leftarrow 0$

      for each $item$ in $items$ do
          if $total\_w + item\_weight \leq capacity$ then
              $total\_p \leftarrow total_p + item\_val$
              $total\_w \leftarrow total_w + item\_weight$
              Add $item$ to $knapsack\_items$
          end if
      end for
  return $knapsack\_items$, $total\_p$, $total\_w$
  end function
```

(i.e., the item to the left of the deleted item in the sorted list) or items with a weight greater than the weight of the removed item. Furthermore, when developing algorithms to find the optimal profit or searching for a heuristic that provides a better profit than a pre-determined one when the knapsack is full, replacements involving items with lower profit-to-weight ratios and of equal or lower total weight should not be considered. In other words, there is no need to investigate these item replacement possibilities.

Theorem 1: If $\alpha, \beta, \gamma, and \delta$ are all greater than zero, and if $\frac{\alpha}{\beta} \geq \frac{\gamma}{\delta}$, then the following holds: $\frac{\alpha}{\beta} \geq \frac{\alpha+\gamma}{\beta+\delta} \geq \frac{\gamma}{\delta}$.

Proof:

Given equation $\frac{\alpha}{\beta} \geq \frac{\gamma}{\delta}$ gives:

$$\frac{\alpha}{\beta} = \frac{\frac{\alpha}{\beta} \cdot (1 + \frac{\delta}{\beta})}{1 + \frac{\delta}{\beta}} = \frac{\frac{\alpha}{\beta} + \frac{\alpha\delta}{\beta^2}}{1 + \frac{\delta}{\beta}} = \frac{\left(\frac{\alpha}{\delta} + \frac{\alpha}{\beta}\right) \cdot \frac{\delta}{\beta}}{\left(\frac{\beta}{\delta} + 1\right) \cdot \frac{\delta}{\beta}} =$$

$$\frac{\frac{\alpha}{\delta} + \frac{\alpha}{\beta}}{\frac{\beta}{\delta} + 1} \geq \frac{\frac{\alpha}{\delta} + \frac{\gamma}{\delta}}{\frac{\beta}{\delta} + 1} = \frac{\frac{\alpha+\gamma}{\delta}}{\frac{\beta+\delta}{\delta}} = \frac{\alpha + \gamma}{\beta + \delta},$$

$$\frac{\alpha + \gamma}{\beta + \delta} = \frac{\frac{\alpha+\gamma}{\beta}}{\frac{\beta+\delta}{\beta}} = \frac{\frac{\alpha}{\beta} + \frac{\gamma}{\beta}}{1 + \frac{\delta}{\beta}} \geq \frac{\frac{\gamma}{\delta} + \frac{\gamma}{\beta}}{1 + \frac{\delta}{\beta}} = \frac{\left(\frac{\gamma}{\delta} + \frac{\gamma}{\beta}\right) \cdot \frac{\beta}{\delta}}{\left(1 + \frac{\delta}{\beta}\right) \cdot \frac{\beta}{\delta}} =$$

$$\frac{\frac{\gamma\beta}{\delta^2} + \frac{\gamma}{\delta}}{\frac{\beta}{\delta} + 1} = \frac{\frac{\gamma}{\delta} \cdot \left(\frac{\beta}{\delta} + 1\right)}{\frac{\beta}{\delta} + 1} = \frac{\gamma}{\delta}.$$

■

Theorem 2: If in the Danzig sequence, the term $\frac{p}{t}$ is to the left of the terms $\frac{p_1}{t_1}, \frac{p_2}{t_2}, ..., \frac{p_n}{t_n}$, that is, $\frac{p}{t} \geq \frac{p_i}{t_i}$ for every $i = 1, 2, ..., n$, and if $t = t_1 + t_2 + ... + t_n$ (the weight of the item inserted in the knapsack is equal to the weight of those that will be inserted instead), then it follows that $v \geq p_1 + p_2 + ... + p_n$. That is, the profit of the item initially placed in the knapsack and then removed is greater than the total profit of the items that will be placed in its place.

Proof: If instead of an item with profit p and weight t, we insert an item to its right in the sequence of items sorted in descending order by the profit-to-weight ratio, that is, if we insert an item with ratio $\frac{p_1}{t_1}$ where $\frac{p}{t} \geq \frac{p_1}{t_1}$, then it is evident that if the denominators are equal, $p \geq p_1$. That is, $\frac{p}{t} \geq \frac{p_1}{t_1}$, and $t = t_1$ implies $p \geq p_1$. Hence $\frac{p}{t} \geq \frac{p_1}{t_1}$ and $t = t_1$ implicates $p \geq p_1$. ■

Assume that instead of an item with profit p and weight t, n items are inserted to its right in the sequence of items sorted in descending order by the profit-to-weight ratio, i.e., items with ratios $\frac{p_1}{t_1}, \frac{p_2}{t_2}, \cdots, \frac{p_n}{t_n}$, and $t = t_1 + t_2 + \cdots + t_n$ are inserted, which implies:

$$\frac{p}{t} \geq \frac{p_1}{t_1}, \frac{p}{t} \geq \frac{p_2}{t_2}, \cdots, \frac{p}{t} \geq \frac{p_n}{t_n}$$

Without loss of generality, it can be assumed that the ratios $\frac{p_i}{t_i}$ are sorted in descending order. Due to lemma 1 (all weights and profits are positive), it holds that:

$$\frac{p_1}{t_1} \geq \frac{(p_1 + p_2)}{(t_1 + t_2)} \geq \frac{p_2}{t_2} \geq \frac{p_3}{t_3}$$
$$\frac{p_1}{t_1} \geq \frac{p_1 + p_2 + p_3}{t_1 + t_2 + t_3} \geq \frac{p_3}{t_3} \geq \frac{p_4}{t_4}$$
$$\cdots$$
$$\frac{p}{t} \geq \frac{p_1}{t_1} \geq \frac{p_1 + p_2 + \cdots + p_n}{t_1 + t_2 + \cdots + t_n} \geq \frac{p_n}{t_n}$$

that is,

$$\frac{p}{t} \geq \frac{p_1 + p_2 + \cdots + p_n}{t_1 + t_2 + \cdots + t_n} \, and \, t = t_1 + t_2 + \cdots + t_n$$

which implies that Eq. 4, is true, as claimed.

$$p \geq p_1 + p_2 + \cdots + p_n \tag{4}$$

In particular, it follows that when replacing an item of weight t and profit p with a set of items that have a total weight smaller than the removed item ($t > t_1 + t_2 + \cdots + t_n$), and satisfy the condition that the ratio $\frac{p}{t} \geq \frac{p_i}{t_i}$ for every $i = 1, 2, ..., n$, then it also holds that $p \geq p_1 + p_2 + \cdots + p_n$. By employing the Dantzig's greedy algorithm to fill the knapsack initially and subsequently attempting to improve the situation by replacing one item with others, an item that is heavier than the one being removed or an item that was previously overlooked or skipped during the initial filling, has to be selected. Specifically, if the knapsack has been filled to its maximum capacity C using the Danzig method, it ensures the optimal selection of items. To enhance the results further, OptDG algorithm is proposed. The algorithm removes the heaviest item from the knapsack and adds other items until equation 5 is met.

$$\sum_{n}^{i=1} t_i \leq C \tag{5}$$

Algorithm 2 OptDG algorithm for solving the 0-1 Knapsack problem

```
Require: A set of items $items$, total knapsack capacity
    $capacity$, current $total\_profit$, and current $total\_weight$
Ensure: A knapsack that is packed to its full carrying
    capacity with items that have the greatest possible profit

  function SolveKnapsackOptimized(items, C,
  knapsack_items, total_p, total_w)
      Sort $items$ by item's weight in decreasing order
      $unique\_items \leftarrow$ Items not contained in the knapsack
      $heavy\_item \leftarrow$ Heaviest item in a knapsack

      $total\_w \leftarrow total\_w - heavy\_weight$
      $total\_p \leftarrow total\_p - heavy\_profit$
      Remove heaviest item in a knapsack

      for $item$ in $unique\_items$ do
          if $total\_w + item\_weight \leq C$ then
              $total\_p \leftarrow total\_p + item\_val$
              $total\_w \leftarrow total\_w + item\_weight$
              Add $item$ to $knapsack\_items$
          end if
      end for
      return $knapsack\_items$, $total\_p$, $total\_w$
  end function

  function SOLVE_KNAPSACK(items, capacity)
      $knapsack\_items$, $total\_p$, $total\_w$, $new\_items \leftarrow$
  SolveKnapsackGreedy($items$, $C$)
      if size of $knapsack\_items$ == 0 then
          return $[], 0, 0$
      end if
      $opti\_knapsack$, $opti\_p$, $opti\_w \leftarrow$
  SolveKnapsackOptimized($new\_items$, $C$,
  $knapsack\_items$, $total\_p$, $total\_w$)
      if $opti\_p > total\_p$ then
          return $opti\_knapsack$, $opti\_p$, $opti\_w$
      else
          return $knapsack\_items$, $total\_p$, $total\_w$
      end if
  end function
```

The resulting knapsack must meet an Eq. 6, and have a higher profit than the previous knapsack. If the improvement in knapsack profit is not achieved, the algorithm reverts to the Dantzig's solution. The optimization aims to maximize the knapsack's profit and approach the outcome of the optimal algorithm. OptDG algorithm, depicted in Algorithm 2, has a time complexity of $O(nC)$, where n represents the number of items and C represents the maximum weight capacity of the knapsack.

$$C - \sum_{i=1}^{n} t_i \geq 0 \tag{6}$$

IV. Experimental Results

This section provides an evaluation of the OptDG algorithm's performance in relation to Datzing's greedy algorithm (implemented using dynamic programming). Table I to Table VII present results of an experiment that benchmarked both Datzing's greedy algorithm and OptDG algorithm with different item generation strategies.

V. Discussion

This section discusses research outcomes on the optimal (using dynamic programming) algorithm, Datzing's greedy algorithm, and newly proposed OptDG algorithm.

The maximum number of items that can be placed in the knapsack remains at 20, but the total capacity of the knapsack, as well as the profits and weights of the items, vary depending on the case. In order to observe the speed of execution of a specific test case, the number of iterations was set to 100,000. This means that we generated new items 100,000 times and attempted to fill the knapsack. The number of cases and the duration of the algorithm are mean averages. In the first test case, we generated random items with weights ranging from 1 to the knapsack's carrying capacity C, and their profits ranged from 1 to 100. The outcomes for the first test case are displayed in Table I. In 4% of cases, the OptDG algorithm outperformed Dantzig's greedy algorithm, but it took longer to execute in those cases than Dantzig did in other situations. Table II shows the results of the methods that were tested, with weights equal to the knapsack's total capacity and random profits between 1 and 100. Because the heaviest item in the knapsack could not be found when all things were of the same weight, the optimized procedure in this observed scenario was never better than conventional Dantzing. Table III displays how items are organized in a knapsack, with weights equal to 5 and profits chosen at random from 1 to 100. Even in this test case, Dantzing's greedy algorithm was still superior because, like in the previous observation, it is unable to identify the heaviest item in the knapsack when all of the profits are equal. The improved approach takes twice as long as the standard Dantzig's greedy algorithm due to the time spent looking for the heaviest item, which cannot be discovered. The configuration of items in the knapsack is shown in Table IV, where the profits are equal to 5, and the weights of the items are distributed at random from 1 to the knapsack's maximum capacity. Because all items have fixed profits and the OptDG algorithm also sought the items with the highest profit, it was unable to determine the best arrangement of the items in the knapsack and did not outperform Dantzing's greedy algorithm in this case. Table V shows the combination of items when their squared profits and randomly generated weights between 1 and T are used. The OptDG algorithm exhibited the highest efficiency in the observed case in point, but since it must place items in the knapsack twice, it is frequently slower than the standard Dantzig's greedy algorithm. Table VI displays how items are arranged when the weights are created at random between the ranges of 1 and T and the profits are powers of 10. In this instance, the Dantzig's greedy algorithm and the optimal algorithm were nearly identical, and the OptDG algorithm also added to the Dantzig's greedy algorithm's strengths. As a result, the accuracy of the algorithm has increased to 100%. If the OptDG algorithm cannot find a better result, it will

TABLE I. Case 1: The Items in the Knapsack are Randomly Generated, Where Both the Profits and the Weights are Randomly Generated in the Range of 1 to 100

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	60.264%	4.509%	231050	4700	4830
100	20	100000	60.127%	4.342%	478580	3780	5830
150	20	100000	60.261%	4.380%	686380	4880	4560
200	20	100000	60.197%	4.316%	897600	4230	6540
250	20	100000	60.182%	4.293%	1108070	2480	5950
300	20	100000	60.183%	4.400%	1343440	3270	5470
350	20	100000	60.216%	4.344%	1566680	3400	7210
400	20	100000	60.246%	4.318%	1818600	4580	5620

TABLE II. Case 2: The Items in the Knapsack are Randomly Generated Where the Profits are Randomly Generated in the Range from 1 to 100, and the Weights are Equal to the Total Capacity of the Knapsack C

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	100%	0%	171210	2820	4860
100	20	100000	100%	0%	342810	2170	4180
150	20	100000	100%	0%	527060	3440	4490
200	20	100000	100%	0%	688170	4390	4850
250	20	100000	100%	0%	833560	2970	4010
300	20	100000	100%	0%	999790	3270	5760
350	20	100000	100%	0%	1183110	3920	3270
400	20	100000	100%	0%	1323160	3580	4950

TABLE III. Case 3: The Items in the Knapsack are Randomly Generated Where the Profits are Randomly Generated in the Range of 1 to 100, and the Weights are Equal to the Number 5

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	100%	0%	254390	3590	6240
100	20	100000	100%	0%	503960	4210	7220
150	20	100000	100%	0%	718510	4090	6660
200	20	100000	100%	0%	950920	3760	7660
250	20	100000	100%	0%	1191520	4040	5180
300	20	100000	100%	0%	1444620	5890	5300
350	20	100000	100%	0%	1706450	5930	8460
400	20	100000	100%	0%	2001300	5530	6250

TABLE IV. Case 4: The Items in the Knapsack are Randomly Generated Where the Profits are Equal to the Number 5, and the Weights are Randomly Generated Ranging from 1 to 100

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	100%	0%	195300	2190	5010
100	20	100000	100%	0%	430430	3140	4540
150	20	100000	100%	0%	626020	3570	4490
200	20	100000	100%	0%	849670	4090	5870
250	20	100000	100%	0%	1060400	3730	5580
300	20	100000	100%	0%	1283480	5000	5750
350	20	100000	100%	0%	1529420	6100	5010
400	20	100000	100%	0%	1786890	4240	3960

TABLE V. Case 5: The Items in the Knapsack are Randomly Generated Where the Profits are Squared, and the Weights are Randomly Generated Ranging from 1 to 100

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	64.006%	5.439%	226020	4740	4530
100	20	100000	64.037%	5.148%	452930	4830	5590
150	20	100000	64.065%	5.077%	680830	5470	3720
200	20	100000	63.641%	5.249%	897590	4460	5100
250	20	100000	63.872%	5.182%	1118750	4090	5170
300	20	100000	63.959%	5.097%	1347190	4030	6530
350	20	100000	64.040%	4.990%	1597040	3780	5000
400	20	100000	63.969%	5.115%	1864040	4820	7990

TABLE VI. Case 6: Items in the Knapsack are Randomly Generated Where the Profits are Powers of 10, and the Weights are Randomly Generated Ranging from 1 to C

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	99.869%	0.027%	234560	6570	6420
100	20	100000	99.847%	0.029%	473490	5280	6400
150	20	100000	99.834%	0.031%	692280	4900	6760
200	20	100000	99.843%	0.042%	918950	4990	7000
250	20	100000	99.840%	0.021%	1147130	5170	8740
300	20	100000	99.844%	0.025%	1407410	4990	7480
350	20	100000	99.841%	0.028%	1666750	5780	7180
400	20	100000	99.809%	0.041%	1902380	7320	7960

TABLE VII. Case 7: The Items in the Knaspack are Randomly Generated Where the Profits are Randomly Generated Ranging from 1 to 100, and the Weights are Powers of 10

Capacity [C]	Items	Iters.	Dantzing closer to optimal	OptDG better than Dantzing	Optimal timings [ns]	Danzing timings [ns]	OptDG timings [ns]
50	20	100000	100.000%	0.000%	187830	6880	4840
100	20	100000	98.609%	1.351%	375880	7030	5960
150	20	100000	100.000%	0.000%	556060	6560	7840
200	20	100000	99.989%	0.002%	723580	8410	5520
250	20	100000	100.000%	0.000%	915100	7980	9490
300	20	100000	100.000%	0.000%	1118070	7950	5790
350	20	100000	100.000%	0.000%	1279060	8640	8800
400	20	100000	100.000%	0.000%	1492240	7160	6690

return the Dantzing result. Table VII displays the results when the weights are powers of 10, and the profits are randomly generated in the range of 1 to 100.

VI. Conclusion

In this paper, the binary knapsack (KP01) problem is addressed, which is one of the most complex problems in the theory of computational complexity. Although dynamic programming yields an exact answer, it is computationally expensive for large numbers. Heuristic algorithms are used to quickly approximate results in order to speed up the process at the expense of solution precision. In this research, we looked at Dantzig's greedy algorithm, which efficiently places items in the knapsack but falls short of the dynamic programming approach's outcome in some circumstances.

In order to address this issue, OptDG algorithm is proposed. The proposed algorithm, after arranging the knapsack using the Dantzig's greedy algorithm, removes the heaviest item and attempts to add other items that are not currently in the knapsack. If the total profit of the knapsack increases, the function returns the new solution, otherwise it returns the original solution. Theoretical foundations of the proposed algorithm is described and mathematically presented. Furthermore, a benchmark for performance evaluation of the speed and accuracy metrics of the proposed OptDG algorithm, optimal dynamically programmed algorithm and Dantzig's greedy algorithm in various scenarios is conducted. In the majority of instances, the proposed OptDG algorithm outperformed the conventional Dantzig's greedy algorithm. According to the performance evaluation results, due to the additional knapsack filling, the proposed OptDG algorithm requires an additional minor time for execution which is not considered as a drawback in majority of possible applications.

The aim of our forthcoming research is to integrate further benchmark studies to enhance our comprehension of how to optimize the OptDG algorithm in regard to Dantzing's greedy model.

Acknowledgment

This work is supported by Croatian Science Foundation under project number DOK-2021-02-7509.

References

[1] W. Tian, G. Li, X. Wang, Q. Xiong, and Y. Jiang, "Transforming NP to P: An Approach to Solve NP Complete Problems," Apr. 2015.

[2] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, ser. STOC '71. New York, NY, USA: Association for Computing Machinery, May 1971, pp. 151–158.

[3] M. Sipser, *Introduction to the Theory of Computation*, third edition, international edition ed. Australia Brazil Japan Korea Mexiko Singapore Spain United Kingdom United States: Cengage Learning, 2013.

[4] S. Coniglio, F. Furini, and P. San Segundo, "A new combinatorial branch-and-bound algorithm for the Knapsack Problem with Conflicts," *European Journal of Operational Research*, vol. 289, no. 2, pp. 435–455, Mar. 2021.

[5] U. Pferschy and J. Schauer, "The Knapsack Problem with Conflict Graphs," *Journal of Graph Algorithms and Applications*, vol. 13, no. 2, pp. 233–249, 2009.

[6] M. R. Bonyadi, Z. Michalewicz, and L. Barone, "The travelling thief problem: The first step in the transition from theoretical problems to realistic problems," in *2013 IEEE Congress on Evolutionary Computation*, Jun. 2013, pp. 1037–1044.

[7] A. Fréville, "The multidimensional 0–1 knapsack problem: An overview," *European Journal of Operational Research*, vol. 155, no. 1, pp. 1–21, May 2004.

[8] V. Cacchiani, M. Iori, A. Locatelli, and S. Martello, "Knapsack problems — An overview of recent advances. Part I: Single knapsack problems," *Computers & Operations Research*, vol. 143, p. 105692, Jul. 2022.

[9] S. S. Skiena, "Who is interested in algorithms and why?: Lessons from the Stony Brook algorithms repository," *ACM SIGACT News*, vol. 30, no. 3, pp. 65–74, Sep. 1999.

[10] S. Martello and P. Toth, "An upper bound for the zero-one knapsack problem and a branch and bound algorithm," *European Journal of Operational Research*, vol. 1, no. 3, pp. 169–175, May 1977.

[11] ——, "A New Algorithm for the 0-1 Knapsack Problem," *Management Science*, vol. 34, no. 5, pp. 633–644, May 1988.

[12] D. Pisinger, "A minimal algorithm for the Bounded Knapsack Problem," in *Integer Programming and Combinatorial Optimization*, ser. Lecture Notes in Computer Science, E. Balas and J. Clausen, Eds. Berlin, Heidelberg: Springer, 1995, pp. 95–109.

[13] ——, "A Minimal Algorithm for the 0-1 Knapsack Problem," *Operations Research*, vol. 45, no. 5, pp. 758–767, 1997.

[14] S. Martello, D. Pisinger, and P. Toth, "Dynamic Programming and Strong Bounds for the 0-1 Knapsack Problem," *Management Science*, vol. 45, no. 3, pp. 414–424, 1999.

[15] J. Jooken, P. Leyman, and P. De Causmaecker, "Features for the 0-1 knapsack problem based on inclusionwise maximal solutions," *European Journal of Operational Research*, Apr. 2023.

[16] K. Smith-Miles, J. Christiansen, and M. A. Muñoz, "Revisiting where are the hard knapsack problems? via Instance Space Analysis," *Computers & Operations Research*, vol. 128, p. 105184, Apr. 2021.

[17] J. Jooken, P. Leyman, and P. De Causmaecker, "A new class of hard problem instances for the 0–1 knapsack problem," *European Journal of Operational Research*, vol. 301, no. 3, pp. 841–854, Sep. 2022.

[18] V. Chvátal, "Hard Knapsack Problems," *Operations Research*, vol. 28, no. 6, pp. 1402–1411, Dec. 1980.

[19] Z. Gu, G. L. Nemhauser, and M. W. P. Savelsbergh, "Lifted Cover Inequalities for 0-1 Integer Programs: Complexity," *INFORMS Journal on Computing*, vol. 11, no. 1, pp. 117–123, Feb. 1999.

[20] S. Jukna and G. Schnitger, "Yet harder knapsack problems," *Theoretical Computer Science*, vol. 412, no. 45, pp. 6351–6358, Oct. 2011.

[21] F. A. Morales and J. A. Martínez, "Analysis of Divide-and-Conquer strategies for the 0–1 minimization knapsack problem," *Journal of Combinatorial Optimization*, vol. 40, no. 1, pp. 234–278, Jul. 2020.

[22] Y. Yang, N. Boland, and M. Savelsbergh, "Multivariable Branching: A 0-1 Knapsack Problem Case Study," *INFORMS Journal on Computing*, vol. 33, no. 4, pp. 1354–1367, 2021.

[23] M. Hifi, H. Mhalla, and S. Sadfi, "Sensitivity of the Optimum to Perturbations of the Profit or Weight of an Item in the Binary Knapsack Problem," *Journal of Combinatorial Optimization*, vol. 10, no. 3, pp. 239–260, Nov. 2005.

[24] ——, "An adaptive algorithm for the knapsack problem: Perturbation of the profit or weight of an arbitrary item," *European Journal of Industrial Engineering*, vol. 2, no. 2, pp. 134–152, Jan. 2008.

[25] T. Belgacem and M. Hifi, "Sensitivity analysis of the optimum to perturbation of the profit of a subset of items in the binary knapsack problem," *Discrete Optimization*, vol. 5, no. 4, pp. 755–761, Nov. 2008.

[26] D. Pisinger and A. Saidi, "Tolerance analysis for 0–1 knapsack problems," *European Journal of Operational Research*, vol. 258, no. 3, pp. 866–876, May 2017.

[27] T. M. Chan, "Approximation Schemes for 0-1 Knapsack," p. 12 pages, 2018.

[28] C. Jin, "An Improved FPTAS for 0-1 Knapsack," Apr. 2019.

[29] D. Ghosh, N. Chakravarti, and G. Sierksma, "Sensitivity analysis of a greedy heuristic for knapsack problems," *European Journal of Operational Research*, vol. 169, no. 1, pp. 340–350, Feb. 2006.

[30] C. Wilbaut, R. Todosijevic, S. Hanafi, and A. Fréville, "Heuristic and exact reduction procedures to solve the discounted 0–1 knapsack problem," *European Journal of Operational Research*, vol. 304, no. 3, pp. 901–911, Feb. 2023.

[31] T. H. Cormen, Ed., *Introduction to Algorithms*, 3rd ed. Cambridge, Mass: MIT Press, 2009.

[32] J. Dong, "Dynamic Programming — 0/1 Knapsack (Python Code)," Sep. 2020.

[33] G. B. Dantzig, "Discrete-Variable Extremum Problems," *Operations Research*, vol. 5, no. 2, pp. 266–277, 1957.

A Relevant Feature Identification Approach to Detect APTs in HTTPS Traffic

Abdou Romaric Tapsoba, Tounwendyam Frédéric Ouédraogo
UFR Sciences et Technologies, Université Norbert Zongo, Koudougou, Burkina Faso

***Abstract*—This study addresses the significant challenges posed by Advanced Persistent Threats (APTs) in modern computer networks, particularly their use of DNS to establish covert communication via command and control (C&C) servers. The advent of TLS 1.3 encryption further complicates detection efforts, as critical data within DNS over HTTPS (DoH) traffic remains inaccessible, and decryption would compromise user privacy. APTs frequently leverage Domain Generation Algorithms (DGAs), necessitating real-time detection solutions based on immediately accessible features within HTTPS traffic. Current research predominantly focuses on system-level behavioral analysis, often neglecting the proactive potential offered by Cyber Threat Intelligence (CTI), which can reveal malicious patterns through Techniques, Tactics, and Procedures (TTPs) and Indicators of Compromise (IoCs). This study proposes an innovative approach utilizing the MITRE ATT&CK framework to identify relevant features in the face of encryption and the inherent complexity of APT activities. The primary objective is to develop a robust dataset and methodology capable of detecting APT behaviors throughout their lifecycle, emphasizing a lightweight, cost-effective solution through passive monitoring of network traffic to ensure real-time detection. The key contributions of this research include an in-depth analysis of the encryption challenges in detecting DNS-based APTs, a thorough examination of APT attack strategies using DNS, and the integration of CTI to enhance detection capabilities. Moreover, this study introduces the KAPT 2024 dataset, generated by the KExtractor tool, and demonstrates the effectiveness of the detection model through experiments with a variety of machine learning algorithms. The results underscore the potential for this approach to significantly improve APT detection in encrypted network environments.**

Keywords*—*DNS over HTTPS; advanced persistent threats; machine learning; cyber threat intelligence; MITRE ATT&CK; domain generation algorithms

I. Introduction

Advanced Persistent Threats present significant challenges to security, representing a serious threat that demands thorough research and rigorous evaluation of effective detection techniques. These malicious actors, driven by various objectives ranging from espionage to service disruption, exploit sophisticated communication channels to establish connections with their Command and Control servers. Recent observations indicate that APT actors are increasingly leveraging DNS, even when encrypted, to establish these communications, thereby evading traditional detection methods.The use of DNS by APT actors as a communication channel can be detected through traffic analysis. However, existing machine learning-based methods for detecting malicious domains face significant challenges with HTTPS traffic because key features, such as textual and lexical domain information, NXDomain volumes, and other relevant data, are encrypted in TLS 1.3. The analysis of directly exploitable information within DoH traffic is hindered by the inaccessibility of a large amount of meaningful data, while decryption methods would compromise privacy. Another challenge, due to the evolving and stealthy nature of APT threats using DGA algorithms, is the responsiveness of malicious domain detection, which requires real-time analysis based on immediately accessible features in HTTPS traffic. The dynamic and evolving nature of these attacks necessitates immediate responsiveness, as the effectiveness of any security measure could be compromised without real-time detection [1].

Several techniques have been proposed in the literature to counter APT threats in general. Most current research on APT detection, based on machine and deep learning, focuses on behavioral analysis of attacks at the system level, thus neglecting crucial adversarial intelligence that could proactively contribute to threat prevention [2], [3]. Cyber Threat Intelligence has emerged as a potential solution to help organizations address the complex and stealthy nature of cyber threats [4]. The exploration of intelligence platforms involves extracting Techniques, Tactics, and Procedures and Indicators of Compromise on threats. TTPs and IoCs play a crucial role in identifying malicious behaviors and attack patterns specific to APTs. The term “tactics” refers to the method used by the APT to carry out the attack from start to finish. The “techniques” used by the APT during its attack describe its technological strategy to achieve its goals. Finally, the “procedure” of an APT describes the steps used by the attacker to achieve its objectives [5]. Many researchers have used machine learning to detect APT threats. However, these proposed methods do not consider detection at all levels of the APT lifecycle [6]. Additionally, the lack of datasets thoroughly exploiting the TTPs and IoCs provided by intelligence platforms does not proactively promote the detection of C&C domains submerged by DGAs, constituting a current research challenge [7]. Publicly available datasets can detect several levels of the cycle, but not the entirety of the phases, as many works have unfortunately not mentioned the use of persistence and stealth tools such as DGAs in the lifecycle. We aim to develop a system to analyze and detect malicious domains at every stage of the APT lifecycle using DNS, even when encrypted, as a communication channel. Our method must meet several key constraints, including ensuring privacy by avoiding the use of any decryption techniques and relying solely on clear-text features directly accessible from the traffic. Additionally, the solution should be lightweight and low-cost, requiring no equipment or installation on endpoints, and based on passive network traffic monitoring. Given the nature of APT threats, the method must also ensure efficiency and responsiveness, enabling quick reaction with real-time detection.

This study makes several significant contributions to the field of cybersecurity, particularly in the detection of APTs exploiting DNS. It provides a comprehensive analysis of the challenges posed by encryption in threat detection, addressing the limitations of current methods for detecting DGA attacks, which are increasingly complicated by the widespread encryption of communications. The research offers an in-depth examination of APT attack strategies, detailing the TTPs and IoCs employed by APTs to exploit DNS. Furthermore, it highlights the role of CTI platforms in enhancing detection capabilities by integrating relevant data sources and enriching detection features. The originality of this work lies in its innovative methodological approach, which combines Artificial Intelligence and Threat Intelligence to create a robust dataset, namely KAPT 2024. This dataset, coupled with importance level indicators, has been rigorously tested with various machine learning algorithms, demonstrating the effectiveness of the proposed architectural model in real-time, multi-class detection scenarios, thereby contributing significantly to the advancement of cybersecurity research.

The remainder of this paper is organized as follows. Section II provides a literature review on the subject. Section III discusses the contribution of threat intelligence in APT detection. Section IV details the steps of our proposed methodology, including the data sources used, feature extraction, and the multi-class classification module. Section V implements and evaluates the methodological approach. Finally, Section VI concludes the article and suggests future research directions.

II. Related Work

The literature review on Advanced Persistent Threats highlights the ongoing evolution of sophisticated attacks targeting computer systems and networks. Various techniques have been developed to counter APTs, with recent research emphasizing the predominant use of machine learning techniques in detecting these threats [5]. These studies underline the importance of continuous research to enhance APT detection capabilities and mitigate cybersecurity risks as threats evolve. For instance, Weiwu Ren *et al.* have analyzed the effectiveness of deep learning for precise and real-time APT detection [8], while Nkiruka Eke *et al.* proposed a hybrid model combining deep and machine learning techniques for more effective detection [9]. Manuel Miguez *et al.* contributed by proposing a cyber kill chain model and early detection methodology for APTs, stressing the importance of a strategic defense approach [10]. APT attacks, meticulously planned by malicious actors, generally involve several stages. While these actors may exhibit distinct characteristics, the phases of their attacks are typically similar, differing mainly in the tactics and techniques used in each phase. Alshamrani *et al.* categorized APT attacks into five stages: Reconnaissance, Establish Foothold, Lateral Movement/Stay Undetected, Exfiltration/Impediment, and Post-Exfiltration/Post-Impediment [11]. The authors argue that these stages can represent any APT attack, regardless of the objective. Several studies [2], [7], [9] in the literature have drawn inspiration from this schema presented in [11]. By leveraging this cycle, as done in the present study, it is possible to detect and prevent APT attacks by understanding the techniques and tactics employed by attackers [2].

CTI plays a crucial role in compiling a comprehensive database on APT behavior. CTI helps manage indicators of compromise, techniques, tactics, and protocols associated with various APT groups. Building on this foundation, Abir Dutta *et al.* proposed the integration of machine learning with a threat intelligence platform [12]. Other works [13], [14] have also emphasized the importance of CTI in enhancing enterprise resilience. Researchers like Yinghai Zhou *et al.* [15] and Nan Sun *et al.* [3] have explored the use of CTI to counter APT attacks by employing automatic extraction and analysis methods for CTI information. Additional studies [16], [17] have presented a framework for sharing CTI that incorporates machine learning models. The challenges of effective threat intelligence sharing are explored in [18], while the works mentioned in [19] propose a trust taxonomy for sharing threat information. Among the most widely used intelligence platforms is the MITRE ATT&CK matrix, which catalogues a comprehensive set of TTPs used by adversaries in each phase of their attacks [3], [5]. The matrix's database is continuously updated with contributions from the research community, making it a cornerstone for APT threat intelligence. Certain studies underscore the increasing importance of integrating threat intelligence and machine learning for defending against cyberattacks [20], [21].

While existing research underscores the importance of machine learning and CTI in detecting APT threats, it also reveals significant limitations that need to be addressed. The necessity of improving machine learning models by exploring CTI platforms for enhanced resilience is evident. However, the exploitation of TTPs and IoCs remains insufficient, as the publicly available training datasets do not comprehensively cover all stages of the APT lifecycle. Consequently, many of the proposed methods fall short of detecting APT threats across all lifecycle stages, particularly when it comes to identifying C&C domains obfuscated by DGAs. This gap highlights the need for further research and the development of innovative methodologies that can overcome these challenges and provide more effective APT detection systems.

III. Contribution of CTI in Detecting APTs

Cyber Threat Intelligence has emerged as a potential solution for companies to address the complex and stealthy nature of cyber threats [13]. Exploring intelligence platforms involves extracting TTPs and IoCs about threats. TTPs and IoCs play a crucial role in identifying malicious behaviors and attack patterns specific to APTs. TTPs describe how attackers achieve their objectives, while IoCs provide concrete evidence of an intrusion, such as IP addresses, file hashes, or malicious domain names. Various tactics, techniques, and procedures are used at each stage of an APT attack, which progresses to the next stage. Given that the TTP attribute allows profiling an APT actor, it is relevant to consider it as a constituent element of a specific detection technique, and it can be used to anticipate and identify APT attacks early [22].

By anticipating the tactics used by attackers, security teams can strengthen their defenses, identify weak points, and develop appropriate detection and response strategies. In-depth knowledge of TTPs also allows for better targeting of security investments and maximizing the effectiveness of protection tools and technologies. In this context, the use of the MITRE ATT&CK matrix proves to be a valuable asset for security professionals. The ATT&CK matrix provides a comprehensive view of the TTPs used by attackers, categorized by attack stage and target platform. By integrating ATT&CK matrix data into their security strategies, organizations can gain a deep understanding of the tactics employed by APTs, as well as the IoCs associated with each stage of the attack. MITRE ATT&CK has been used to represent APT TTPs as it provides extensive knowledge of adversary tactics and techniques based on real-world observations. The ATT&CK knowledge base has been widely used as a foundation for developing specialized threat models and techniques in the business, government, and cybersecurity sectors worldwide [5]. The choice of MITRE as a source of information on TTPs and IoCs stems from its reputation for excellence in the field of cybersecurity. In addition to the ATT&CK matrix, MITRE offers a multitude of free and paid resources, such as technical reports, analysis tools, and training programs. This wealth of resources makes MITRE an essential partner for organizations seeking to strengthen their security posture and protect against APT threats. Compared to methods proposed in the literature, the MITRE platform is better suited to APT threat intelligence, which describes TTPs in a canonical form and can more accurately extract the TTPs summarized in CTI reports [15].

One of the major aspects influencing the accuracy of Machine Learning models is the search for discriminatory features [25]. Indeed, the features designated in one APT detection solution are not necessarily applicable to another solution. The MITRE platform, by providing TTPs and IoCs about APT threats, helps us designate relevant features on the behaviors and techniques used by attackers. By analyzing this data, we can identify the specific patterns and signals associated with APT attacks, allowing us to determine which features are most likely to be relevant for detecting these attacks during each phase. For example, if an IoC indicates that an APT attack used a specific technique to compromise a system, we might select features related to that technique to strengthen our detection model. Similarly, TTPs can guide us toward the features that are most representative of the attack methods used by APTs, enabling us to better target our analysis and defense efforts. The feature selection process begins by identifying representative techniques by exploring the different phases of the APT attack (Initial Access, Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Lateral Movement, Collection, Command and Control, Exfiltration, Impact) and identifying the specific techniques at each phase relevant to the type of targeted attacks. Each phase of the cycle identified on the platform is then translated into measurable features in network data. For example, the "Data Exfiltration" phase will have measurable characteristics such as statistics related to the volume of data transferred, the frequency of outgoing connections, etc. There has been immense interest in exploiting CTI, specifically for proactive cybersecurity defense. By exploring this unique dimension of cybersecurity, this research provides new insights and opens avenues for innovation in combating persistent and sophisticated threats, thereby making a significant contribution to the advancement of the field. In the following section, we will delve into the architectural model of our approach for identifying relevant features in detail.

IV. Proposed Methodology

We have already explored intelligence platforms to identify the most significant features that can contribute to the detection of APT threats and designate data sources covering all phases of the APT lifecycle in the previous section. In this section, we present the architecture used in this study, illustrated in Fig. 1. We proceed with feature extraction to construct the KAPT-2024 dataset, leveraging the selected data sources. We conclude the process by training our models with several supervised learning algorithms.

A. Data Source

The accuracy of classification models inevitably relies on the quality of the data. Exploring intelligence platforms allows us to observe the various tools, TTPs, and IoCs used by APT actors during each phase of the APT lifecycle, thereby indicating the different data sources widely referenced in the literature. These data sources, in pcap format, are selected based on the tools used to generate the traffic. These tools enable us to map the threats from each public data source according to the APT lifecycle. This is essentially what will be discussed in this subsection. In total, four data sources will be analyzed to constitute our own dataset.

1) CSE-CIC-IDS2018 on AWS: CSE-CIC-IDS2018[1], developed in collaboration between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC), is designed to generate a diverse and comprehensive benchmark dataset for intrusion detection. Among the simulated attacks are scenarios such as network infiltration from the inside, HTTP denial-of-service attacks, web application attack collection, brute-force attacks, and attacks based on recent vulnerabilities such as Heartbleed. CSE-CIC-IDS2018 dataset mainly covers the initial compromise, lateral movement, and camouflage phases of the APT threat lifecycle. Included attacks such as brute-force attacks, web attacks, and port scans illustrate a strong focus on the initial compromise phase, crucial for gaining initial access to target systems. Although this dataset is useful for analyzing these critical stages, it does not cover the entire APT threat lifecycle, omitting phases such as persistence, privilege escalation, defense evasion, data exfiltration, and final impact [23].

2) UNSW-NB15: The UNSW-NB15[2] dataset was specifically designed to evaluate Intrusion Detection Systems, making it a valuable tool for threat detection and prevention. With a variety of source files and a wide range of simulated attacks, UNSW-NB15 provides a more realistic and comprehensive representation of modern network traffic compared to some previous datasets like NSL-KDD. The UNSW-NB15 dataset covers several phases of the APT threat lifecycle. Fuzzers and Analysis attacks are associated with the Reconnaissance phase, where attackers gather information about potential targets. Backdoors, DoS, Exploits, and Shellcode attacks fall under the Initial Compromise phase, where attackers exploit vulnerabilities to gain access to target systems. Generic and Worms attacks are linked to the Lateral Movement phase, allowing attackers to move within the compromised network. Although the dataset provides a good representation of critical

[1] https://www.unb.ca/cic/datasets/ids-2018.html

[2] https://research.unsw.edu.au/projects/unsw-nb15-dataset

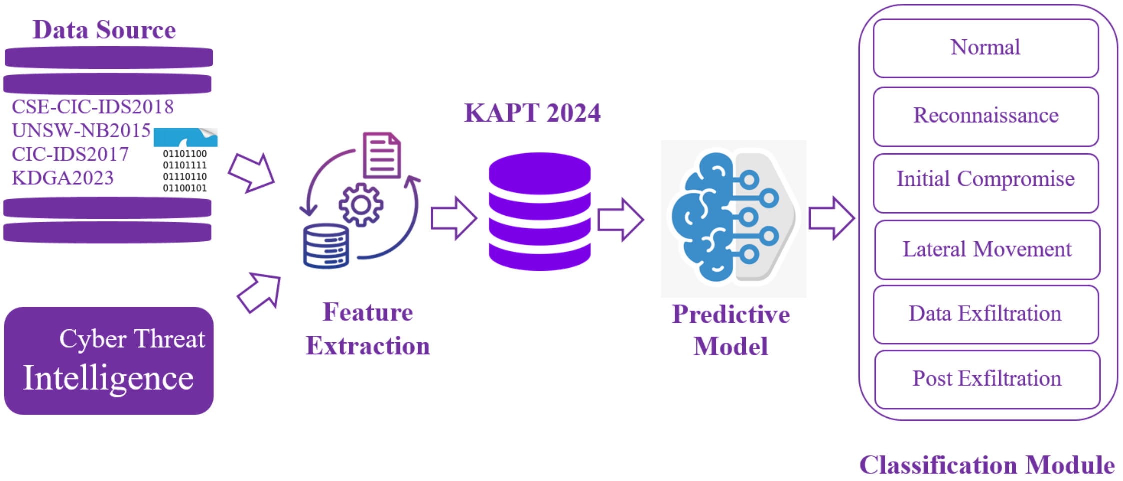


Fig. 1. The methodological structure of the proposed method.

phases, it does not explicitly cover all phases, such as data exfiltration and post-exfiltration [24].

3) CIC-IDS 2017: The CIC IDS 2017[3] dataset is designed for the evaluation of Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS). It contains both benign network traffic and common attacks, thus providing a realistic representation of real-world data. The attacks included in the dataset cover a wide range, including Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS. CIC-IDS 2017 covers the Initial Compromise phase with web attacks such as XSS (Cross-Site Scripting) and SQL injection attacks, as well as brute-force attacks on services like FTP and SSH. The Lateral Movement phase is represented by infiltration attacks and the use of botnets to move laterally within the network, with port scans (PortScan) being possible. The Camouflage phase includes DoS attacks like Hulk, GoldenEye, slowloris, slowhttptest, and DDoS, which seek to conceal malicious activities. However, the dataset does not seem to explicitly cover the data exfiltration phase, which is also an essential component of the APT attack lifecycle [23].

4) KDGA-Insight23: The KDGA-Insight23[4] dataset [25] is specifically designed for real-time analysis of DNS traffic, focusing on detecting malicious activities such as Domain Generation Algorithms, particularly in the context of using DNS over HTTPS. It includes 36 features extracted from pcap files, which are used to distinguish different types of DNS traffic, including DoH and non-DoH traffic, DoH-Tunnel and non-Tunnel traffic, as well as DGA and non-DGA traffic. This dataset can contribute to APT threat detection, especially regarding DGA attacks. In the context of DGA attacks, the initial compromise of a host machine is a fundamental step, corresponding to the first two stages of the APT cycle: acquiring initial access and establishing an initial foothold. Subsequently, setting up a tunnel through the Command and Control corresponds to a later stage of the APT cycle, usually associated with lateral movement within the network and exfiltration of sensitive data. This step aims to establish secret communication between the compromised machine and the C&C server, thereby allowing the attacker to access and control the network more broadly. As for the use of DGA, it may be considered in the camouflage or persistence phase, where attackers deploy sophisticated techniques to evade detection and maintain their network access over the long term. The KDGA-Insight23 dataset provides valuable information for APT threat detection, focusing particularly on DGA attacks and the use of tunnels via C&C. These aspects are closely related to several stages of the APT attack lifecycle, thereby enhancing its relevance in the cybersecurity context.

In conclusion, the use of the four datasets CSE-CIC-IDS2018, CIC-IDS2017, UNSW-NB15, and KDGA-Insight23 in our project is of crucial importance for several reasons. Firstly, each dataset offers a unique perspective on threats and potential attacks encountered in the modern cybersecurity landscape. CSE-CIC-IDS2018 and CIC-IDS2017 provide a variety of real-world attacks, allowing for the testing and evaluation of Intrusion Detection Systems effectiveness in detecting common attacks such as DoS, brute-force attacks, and web attacks. On the other hand, UNSW-NB15 focuses on detecting malicious activities related to DNS traffic, offering valuable insight into detecting DNS-based attacks, including DGA attacks. Lastly, KDGA-Insight23 specifically focuses on detecting DGA attacks in the context of using DNS over HTTPS, making it particularly relevant for detecting camouflage and persistence activities associated with APT attacks. By combining these four datasets, we are able to broadly cover the entire lifecycle of APT attacks, from reconnaissance to long-term persistence in the network. Each dataset contributes to filling the gaps of the others in terms of coverage of specific attack types and camouflage techniques used by attackers. The combined use of these four datasets allows us to benefit from a comprehensive and balanced overview of potential threats in the cybersecurity domain, thereby enhancing our ability to develop and evaluate robust and effective Intrusion Detection Systems against APT attacks. Table I represents our dataset, encompassing all phases of the APT threat lifecycle and enabling threat detection at each stage of the cycle. By extracting these features from the data sources explored in this

[3] https://www.unb.ca/cic/datasets/ids-2017.html
[4] https://github.com/artapsoba/KDGA-Insights

TABLE I. APT CYCLE DESCRIPTION AND ATTACK/TOOLS

APT Cycle	**Description**	**Attack/Tools**
Reconnaissance	Network reconnaissance, identifying vulnerabilities	PortScan
Initial compromise	Establishing a foothold in the network through various techniques	Brute Force, Sql Injection, XSS, FTP-Patator, SSH-Patator
Lateral Movement	Discovering the internal network through compromised systems and taking control of critical devices	Infiltration attack, Bot ARES
Data Exfiltration	Transferring data from local machines in the network to C&C servers, locations, or remote users	Iodine, Dnscat2, Dns2tcp
Post Exfiltration	Persisting the exfiltration process, disabling other critical components, and destroying evidence to ensure clean removal from the organization's network	DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, DoS Heartbleed, DDoS LOIC, DGA

subsection, our dataset stands out for its ability to capture and analyze the various aspects of APT attacks through these 87 selected features.

B. Feature Extraction and Data Pre-Processing

Lexical and textual data have largely lost their relevance due to traffic encryption. Information related to DNS, HTTP, and TLS layers, which has been successfully used in the literature, is now encrypted. The widespread adoption of traffic encryption presents fewer opportunities for security professionals and represents one of the major challenges today. This study addresses the issue of respecting user privacy while maintaining an optimal level of security. It seeks to demonstrate the effectiveness of Machine Learning methods using only the information directly accessible from DoH traffic. Feature extraction from network packets is a crucial step in network data analysis, particularly for APT detection. In this process, packets are grouped by flow to capture network interactions between specific IP addresses and ports, thereby defining forward (incoming) and backward (outgoing) flows. This grouping allows for detailed and granular characterization of data flows, facilitating the analysis of suspicious network behaviors. Forward and backward flows are used to distinguish communication directions, which is essential for identifying potential attack patterns such as unusual data transfers or suspicious responses. By analyzing features such as packet lengths, inter-arrival times (IAT), and TCP flags (PSH, URG), traffic patterns can be better understood, and anomalies indicating a threat can be detected. This approach enables the identification of not only the overall characteristics of flows but also the directional nuances that might indicate malicious activities, making the analysis more precise and relevant for APT detection. The dataset is labeled based on the tools used to generate the traffic. We extract a total of 87 features that comprehensively cover the phases of the APT lifecycle. Depending on the tools used to generate the data sources utilized in this study, we have grouped the raw data into six categories (Normal, Reconnaissance, Initial Compromise, Lateral Movement, Data Exfiltration, and Post Exfiltration) as outlined in Table I. Preprocessing involves cleaning the dataset of all its outliers. The transformations applied to this dataset include digitization, normalization, imputation of missing values, and feature selection. Normalization involves changing the range of values from a large range to a smaller one, typically [0, 1] or [-1, 1] [26]. In this study, data normalization was conducted using the *MinMaxScaler()* method from the Sklearn library, followed by data imputation, which involved removing rows with missing values. The next step focused on selecting the most significant features. This process began with analyzing the correlation between features using a heatmap, which helped identify and remove highly correlated, redundant columns, thereby simplifying the model and improving its performance. The study then employed Recursive Feature Elimination (RFE) to optimize feature selection, retaining only the most relevant variables. This approach reduced noise, improved predictive accuracy, and stabilized performance with around 55 features, leading to a more efficient and accurate model.

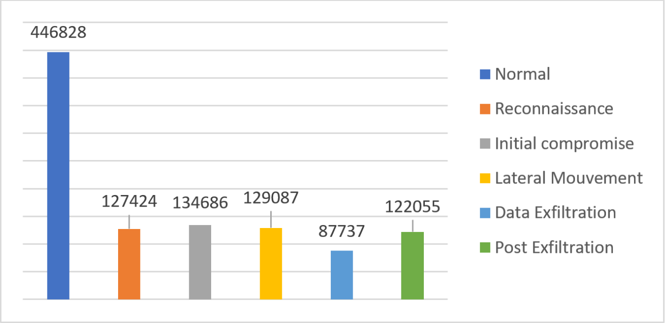


Fig. 2. Collected dataset.

C. Dataset KAPT 2024

The KAPT24 dataset is designed to address the challenges posed by APT threats. It is structured around the complete lifecycle of APT threats, covering the phases of Reconnaissance, Initial Compromise, Lateral Movement, Data Exfiltration, and Post-Exfiltration. The primary objective of this dataset is to provide a comprehensive solution for detecting APT threats by leveraging features extracted directly from HTTPS traffic. To construct this dataset, we utilized intelligence platforms such as MITRE ATT&CK to identify relevant Techniques, Tactics, and Procedures and Indicators of Compromise. This information was crucial in selecting features that effectively capture malicious behaviors. This approach allows for the detection of suspicious activities in a non-intrusive manner, making the dataset valuable for research and the development of new threat detection techniques. The data is collected and classified into different phases of the APT threat lifecycle, as illustrated in Fig. 2. This dataset includes the following categories: Normal (446,828 samples), Reconnaissance (127,424 samples), Initial Compromise (134,686 samples), Lateral Movement (129,087 samples), Data Exfiltration (87,737 samples), and Post Exfiltration (122,055 samples).

A thorough analysis is conducted to select the most relevant features, those that demonstrate a significant ability to discriminate between normal traffic and malicious traffic related to APTs. The choice of features and the method of grouping by flow are motivated by the need to capture the complex dynamics and abnormal behaviors that characterize APT attacks. The forward and backward flows provide a detailed view of network interactions, thus facilitating the identification of anomalies typical of different phases of an APT attack. The features of the KAPT24 dataset, summarized in Table II, include essential information for network flow analysis and APT threat detection.

TABLE II. EXTRACTED FEATURES

Feature Group	**Features**
Flow Identification	F01: FlowID, F02: SrcIP, F03: DstIP, F04: SrcPort, F05: DstPort, F06: Protocol, F07: Timestamp
Flow Duration and TTL	F08: Fl_Duration, F09: TTL, F10: DistinctTTLValue
Packet Counts	F11: Tot_Fwd_Pkts, F12: Tot_Bwd_Pkts, F13: TotLen_FwdPkts, F14: TotLen_BwdPkts
Packet Length Statistics	F15: Fwd-Pkt_Len_Max, F16: Fwd-Pkt_Len_Min, F17: Fwd-Pkt_Len_Mean, F18: Fwd-Pkt_Len_Std, F19: Bwd-Pkt_Len_Max, F20: Bwd-Pkt_Len_Min, F21: Bwd-Pkt_Len_Mean, F22: Bwd-Pkt_Len_Std
Flow Rates	F23: Flow_Byts_sec, F24: Flow_Pkts_sec
Inter Arrival Times	F25: Flow_IAT_avg, F26: Flow_IAT_Std, F27: Flow_IAT_Max, F28: Flow_IAT_Min, F29: Fwd_IAT_Tot, F30: Fwd_IAT_avg, F31: Fwd_IAT_Std, F32: Fwd_IAT_Max, F33: Fwd_IAT_Min, F34: Bwd_IATTot, F35: Bwd_IAT_avg, F36: Bwd_IAT_Std, F37: Bwd_IAT_Max, F38: Bwd_IAT_Min
Flag Counts	F39: Fwd_PSH_Flags, F40: Bwd_PSH_Flags, F41: Fwd_URG_Flags, F42: Bwd_URG_Flags, F54: FIN_Flag_Cnt, F55: SYN_Flag_Cnt, F56: RST_Flag_Cnt, F57: PSH_Flag_Cnt, F58: ACK_Flag_Cnt, F59: URG_Flag_Cnt, F60: CWE_Flag_Cnt, F61: ECE_Flag_Cnt
Header and Byte Metrics	F43: Fwd_Header_Len, F44: Bwd_Header_Len, F45: Fwd_Byts_sec, F46: Bwd_Byts_sec, F47: Fwd_Pkts_sec, F48: Bwd_Pkts_sec
Ratios and Averages	F62: Down_Up_Ratio, F63: Pkt_Size_Avg, F64: Fwd_Seg_Size_Avg, F65: Bwd_Seg_Size_Avg, F66: Fwd_Byts_blk_Avg, F67: Fwd_Pkts_blk_Avg, F68: Fwd_Blk_Rate_Avg, F69: Bwd_Byts_blk_Avg, F70: Bwd_Pkts_blk_Avg, F71: Bwd_Blk_Rate_Avg
Subflows	F72: Subflw_Fwd_Pkts, F73: Subflw_Fwd_Byts, F74: Subflw_Bwd_Pkts, F75: Subflw_Bwd_Byts
Initial Window Metrics	F76: Init_Fwd_Win_Byts, F77: Init_Bwd_Win_Byts
Active and Idle Times	F78: Fwd_Act_Data_Pkts, F79: Fwd_Seg_Size_Min, F80: Active_Mean, F81: Active_Std, F82: Active_Max, F83: Active_Min, F84: Idle_Mean, F85: Idle_Std, F86: Idle_Max, F87: Idle_Min

D. Classification Module

To evaluate the performance of our classification model, we employed six state-of-the-art algorithms. Each of these algorithms was selected for its distinct capabilities to handle complex data and deliver accurate results in various contexts. Using these supervised learning algorithms, we constructed detection models based on the features extracted from the KAPT24 dataset. This dataset, rich in information and meticulously annotated, served as the foundation for training our models. Through this training, the models have acquired the ability to effectively predict APT threats at each stage of their lifecycle. Specifically, our classification models were designed to identify and categorize malicious activities into six distinct categories: Normal (Stage 0), representing benign or normal activities that pose no threat; Reconnaissance (Stage 1), involving information-gathering activities where the attacker searches for vulnerable entry points; Initial Compromise (Stage 2), which marks the phase where the attacker successfully compromises the target system initially; Lateral Movement (Stage 3), referring to movement within the network, allowing the attacker to navigate and extend access to other systems; Data Exfiltration (Stage 4), where the attacker extracts sensitive information from the target network; and Post Exfiltration (Stage 5), encompassing post-exfiltration activities that typically include attempts to cover up traces of the attack or maintain access to the compromised system.

The integration of these algorithms into our classification module has enhanced the accuracy and reliability of threat detection. Each algorithm brings a unique approach to data analysis, capturing various nuances of malicious behaviors. For example, Random Forests and Support Vector Machines provide robust perspectives in terms of classification, while ensemble algorithms like XGBoost, LGBM, and CatBoost optimize performance through sophisticated aggregation methods. The Multi-Layer Perceptron leverages neural networks' capabilities to model complex relationships between features. Similarly, Koala *et al.* employed a comparable approach by using multiple machine learning algorithms to analyze application behavior through event traces in detecting security vulnerabilities [**?**]. The use of these cutting-edge algorithms has enabled the creation of a sophisticated and effective classification model, capable of detecting and categorizing APT threats across all phases of their lifecycle, ensuring proactive and robust defense against advanced cyber threats. Building on the proposed methodology, the next section will focus on the practical implementation and evaluation of this approach, demonstrating its effectiveness in real-world scenarios through comprehensive experimentation and analysis.

V. IMPLEMENTATION AND EVALUATION OF THE APPROACH

In this section, we transition from the theoretical foundation provided in the proposed methodology to the practical implementation of our approach. The main objective of implementing this approach is to demonstrate its real-world applicability, especially given the evolving and adaptive nature of DGAs. To this end, we developed a Flask application using Python, capable of capturing real-time network traffic, extracting key features, and analyzing them through a pre-trained classification model integrated into the system. For packet manipulation, we utilized the Scapy library, which allows for efficient packet sniffing and feature extraction. The system is designed to operate in real-time, predicting whether the sniffed packet is benign or malicious, thus offering an immediate response to potential threats.

Additionally, we integrated a graphical user interface (GUI) using Tkinter to streamline user interaction, making the system more accessible and user-friendly for practical deployment. This real-time capability is crucial for staying ahead of the constantly evolving DGA tactics and bolstering cybersecurity defenses.

In this section, we will thoroughly evaluate the relevance of the extracted features using performance indicators to ensure the system's predictions are both accurate and reliable. Finally, we will conclude by presenting the results of various performance metrics that validate the robustness of our approach, highlighting its capacity to detect APT threats in real-time environments.

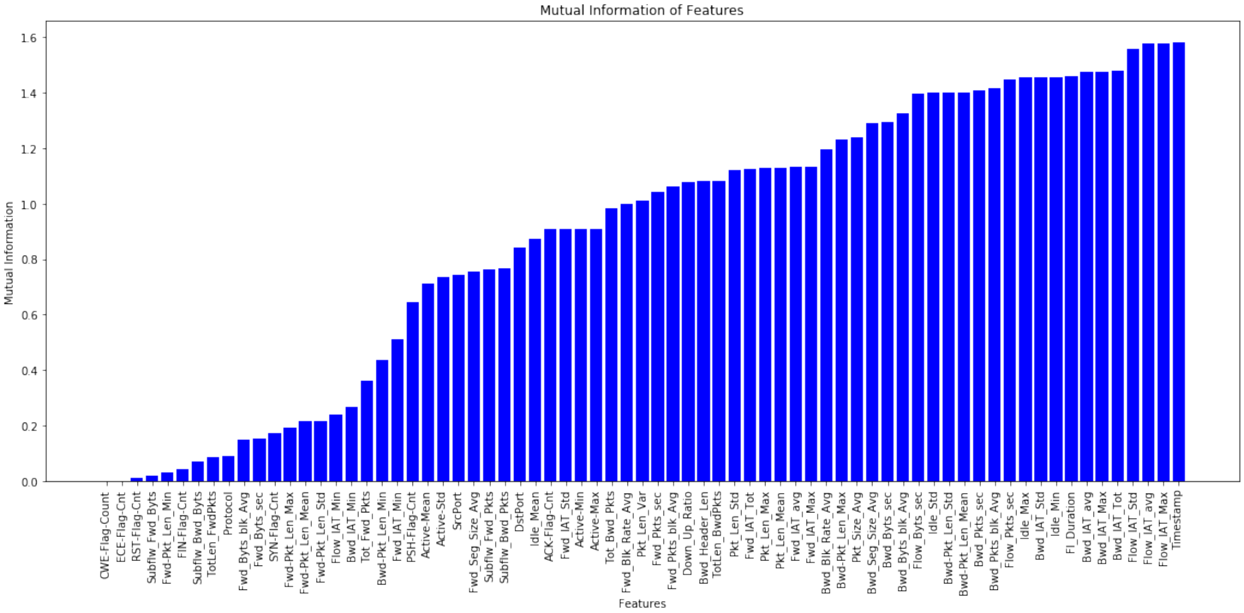


Fig. 3. Feature importance using mutual information.

A. Relevance of Features Analysis

Mutual information is a measure that captures nonlinear and independent relationships in data distribution, making it more flexible than methods like Pearson correlation or ANOVA. Unlike Pearson correlation, which only measures linear relationships, or chi-square tests, limited to categorical variables, mutual information is robust to monotonic transformations and useful for both continuous and categorical variables. It also outperforms model-based feature importance methods in machine learning by avoiding bias towards features with more levels. Fig. 3 presents the mutual importance of features in the KAPT 2024 dataset, used to address challenges posed by APT. The vertical bar chart displays features on the horizontal axis and their mutual importance values on the vertical axis. This visualization highlights that virtually all features have notable mutual importance, suggesting that the feature set is relevant for detecting APT threats.

In summary, we can conclude that the selected features contribute, to varying degrees, to the classification process of the models. This provides a solid starting point for training our models with supervised learning algorithms, whose very satisfactory results are detailed in the following subsection.

B. Results Analysis

In this subsection, we analyze the results obtained from applying various machine learning algorithms to our dataset. The confusion matrix is a crucial metric for evaluating classification models in machine learning, particularly in the context of multi-class classification problems [30]. It provides a detailed view of the model's performance by showing not only the number of correct predictions but also the types and quantities of errors made. From the confusion matrix, various performance metrics such as precision, recall, F1 score, and accuracy can be calculated for each phase individually, offering a more nuanced evaluation of the model's performance. It also helps identify if certain classes are systematically under-predicted or over-predicted, which is particularly useful in imbalanced datasets where some classes may dominate. By using ensemble algorithms such as XGBoost, CatBoost, and LightGBM, as well as algorithms like Random Forests, SVMs, and MLPs, the classification performances of these algorithms are compared and evaluated in a multi-class approach for each of the 6 stages of the APT lifecycle. To ensure the integrity of the evaluation, we followed a methodical approach by mixing samples from each class before splitting them into distinct training and test sets. The results of these evaluations are concisely summarized in Fig. 4, which represents the average metrics across all phases of the cycle [31].

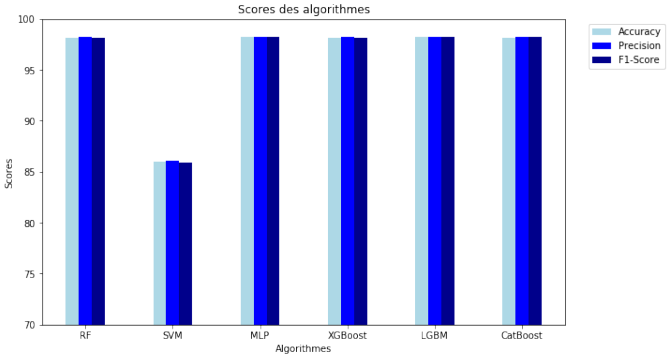


Fig. 4. Simulation results.

The graph shows that the MLP, LGBM, XGBoost, CatBoost and RF algorithms achieve very high and similar scores in terms of Accuracy, Precision, and F1-Score, indicating their effectiveness for this dataset.

Regarding the results from the confusion matrix analysis, let's focus on the MLP case, which offers very satisfactory

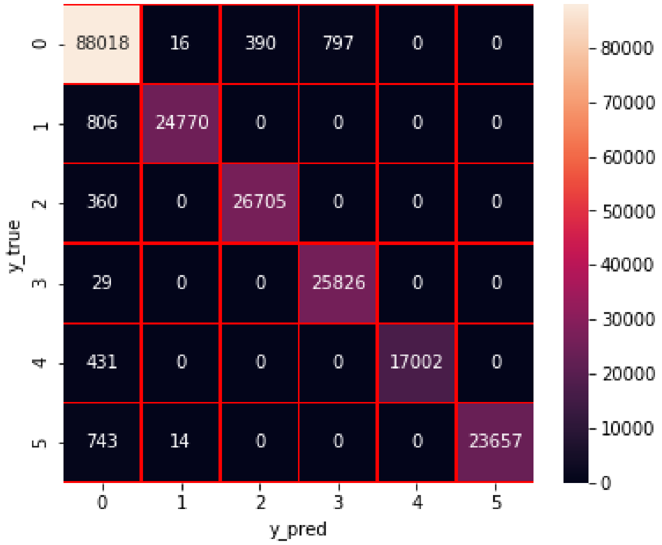


Fig. 5. Confusion matrix for MLP.

results. The confusion matrix for the MLP algorithm (Fig. 5) shows robust overall performance in classifying the different classes, with high numbers on the diagonal representing correct predictions. For example, class 0 is well predicted with 88,018 instances correctly classified, while classes 2 and 3 also show high classification rates, with 26,705 and 25,826 instances correctly classified, respectively. However, some classification errors, though minor, are present, as indicated by the off-diagonal values. False positives, such as the 797 instances of class 0 incorrectly classified as class 3, and false negatives, like the 806 instances of class 1 predicted as class 0, highlight the limitations of the algorithm.

The results obtained from our study have exceeded our expectations, confirming the effectiveness of our proposed solution. Our approach aims to provide a cost-effective and privacy-conscious method that is highly responsive to the evolving nature of APT threats targeting DNS. Initially, we identified 87 plaintext features that are directly accessible without the need for third-party equipment to decrypt data. Subsequently, we developed a lightweight application for capturing traffic, which demonstrated its capability to analyze network traffic in real-time and detect APT threats effectively. Furthermore, the integration of the MITRE framework has provided a comprehensive understanding of APT behaviors, enabling proactive threat detection. This makes our approach not only more efficient and effective but also lightweight and secure, addressing critical concerns in cybersecurity.

VI. Conclusion

The absence of significant features due to encryption in TLS, the limited exploitation of intelligence platforms in the search for proactive solutions, and the lack of training data covering all stages of the APT lifecycle underscore the importance of a balanced understanding of adversaries' behaviors, capabilities, and intentions for effective defense against APT threats targeting DNS. This study sought to develop a comprehensive approach for analyzing and detecting malicious domains throughout the entire APT lifecycle. The method had to meet several constraints: strict privacy compliance, lightweight and low operational cost, as well as optimal efficiency and responsiveness. To address this challenge, we developed a distinctive feature extraction module by analyzing TTPs and IoCs from APT threats using the MITRE ATT&CK matrix, thus contributing to the identification of features and data sources for building a dataset covering all phases of the APT lifecycle. Another major contribution of this study lies in the focus on detecting C&C servers and tools such as DGAs within the APT lifecycle.

Our experiments were conducted using six machine learning algorithms enabling a thorough evaluation of our approach's performance in a multi-class framework. This novel approach, which integrates intelligence platforms and importance indicators, has proven effective in detecting APTs throughout their entire lifecycle. The results obtained open promising perspectives for the continuous improvement of threat detection systems. Our future work will focus on establishing an intelligence platform aimed at sharing threat information within a trust circle for stakeholders with common challenges and strategies. Community sharing allows for alerting others about the occurrence of a probable attack and benefiting from feedback on how to counter a threat.

References

[1] A. R. Tapsoba, T. F. Ouédraogo, and W.-B. S. Zongo, "Analysis of Plaintext Features in DoH Traffic for DGA Domains Detection," in *Information Technology and Systems*, Á. Rocha, C. Ferrás, J. Hochstetter Diez, and M. Diéguez Rebolledo, Eds., Cham: Springer Nature Switzerland, 2024, pp. 127–138. doi: 10.1007/978-3-031-54235-0_12.

[2] A. Al Mamun, H. Al-Sahaf, I. Welch, et S. Camtepe, "Advanced Persistent Threat Detection: A Particle Swarm Optimization Approach", in *2022 32nd International Telecommunication Networks and Applications Conference (ITNAC)*, Wellington, New Zealand: IEEE, nov. 2022, pp. 1-8. doi: 10.1109/ITNAC55475.2022.9998358.

[3] N. Sun et al., "Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives", *IEEE Commun. Surv. Tutor.*, vol. 25, no 3, pp. 1748-1774, 2023, doi: 10.1109/COMST.2023.3273282.

[4] C. Gan, J. Lin, D.-W. Huang, Q. Zhu, et L. Tian, "Advanced Persistent Threats and Their Defense Methods in Industrial Internet of Things: A Survey", *Mathematics*, vol. 11, no 14, p. 3115, juill. 2023, doi: 10.3390/math11143115.

[5] N. I. Che Mat, N. Jamil, Y. Yusoff, et M. L. Mat Kiah, "A systematic literature review on advanced persistent threat behaviors and its detection strategy", *J. Cybersecurity*, vol. 10, no 1, p. tyad023, janv. 2024, doi: 10.1093/cybsec/tyad023.

[6] G. Yan, Q. Li, D. Guo, et B. Li, "AULD: Large Scale Suspicious DNS Activities Detection via Unsupervised Learning in Advanced Persistent Threats", *Sensors*, vol. 19, no 14, p. 3180, juill. 2019, doi: 10.3390/s19143180.

[7] J. Al-Saraireh et A. Masarweh, "A novel approach for detecting advanced persistent threats", *Egypt. Inform. J.*, vol. 23, no 4, pp. 45-55, déc. 2022, doi: 10.1016/j.eij.2022.06.005.

[8] W. Ren et al., "APT Attack Detection Based on Graph Convolutional Neural Networks", *Int. J. Comput. Intell. Syst.*, vol. 16, no 1, p. 184, nov. 2023, doi: 10.1007/s44196-023-00369-5.

[9] H. N. Eke et A. Petrovski, "Advanced Persistent Threats Detection based on Deep Learning Approach", in *2023 IEEE 6th International Conference on Industrial Cyber-Physical Systems (ICPS)*, Wuhan, China: IEEE, mai 2023, pp. 1-10. doi: 10.1109/ICPS58381.2023.10128062.

[10] M. Miguez et B. Sassani (Sarrafpour), "Feature-based Systematic Analysis of Advanced Persistent Threats", *AI Comput. Sci. Robot. Technol.*, vol. 2, mai 2023, doi: 10.5772/acrt.21.

[11] A. Alshamrani, S. Myneni, A. Chowdhary, et D. Huang, “A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities”, *IEEE Commun. Surv. Tutor.*, vol. 21, no 2, pp. 1851-1877, 2019, doi: 10.1109/COMST.2019.2891891.

[12] A. Dutta et S. Kant, “An Overview of Cyber Threat Intelligence Platform and Role of Artificial Intelligence and Machine Learning”, in *Information Systems Security*, vol. 12553, S. Kanhere, V. T. Patil, S. Sural, et M. S. Gaur, Éd., Lecture Notes in Computer Science, vol. 12553. Cham: Springer International Publishing, 2020, pp. 81-86. doi: 10.1007/978-3-030-65610-2_5.

[13] S. Saeed, S. A. Suayyid, M. S. Al-Ghamdi, H. Al-Muhaisen, et A. M. Almuhaideb, “A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience”, *Sensors*, vol. 23, no 16, p. 7273, août 2023, doi: 10.3390/s23167273.

[14] S. Kumar, B. P. Singh, et V. Kumar, “A Semantic Machine Learning Algorithm for Cyber Threat Detection and Monitoring Security”, in 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India: IEEE, déc. 2021, p. 1963-1967. doi: 10.1109/ICAC3N53548.2021.9725596.

[15] Y. Zhou, Y. Tang, M. Yi, C. Xi, et H. Lu, “CTI View: APT Threat Intelligence Analysis System”, Secur. Commun. Netw., vol. 2022, p. 1-15, janv. 2022, doi: 10.1155/2022/9875199.

[16] D. Preuveneers et W. Joosen, “Sharing Machine Learning Models as Indicators of Compromise for Cyber Threat Intelligence”, J. Cybersecurity Priv., vol. 1, no 1, p. 140-163, févr. 2021, doi: 10.3390/jcp1010008.

[17] X. Zhang, X. Miao, et M. Xue, “A Reputation-Based Approach Using Consortium Blockchain for Cyber Threat Intelligence Sharing”, Secur. Commun. Netw., vol. 2022, p. 1-20, août 2022, doi: 10.1155/2022/7760509.

[18] A. Ramsdale, S. Shiaeles, et N. Kolokotronis, “A Comparative Analysis of Cyber-Threat Intelligence Sources, Formats and Languages”, Electronics, vol. 9, no 5, p. 824, mai 2020, doi: 10.3390/electronics9050824.

[19] T. D. Wagner, E. Palomar, K. Mahbub, et A. E. Abdallah, “A Novel Trust Taxonomy for Shared Cyber Threat Intelligence”, Secur. Commun. Netw., vol. 2018, p. 1-11, juin 2018, doi: 10.1155/2018/9634507.

[20] W. Tounsi et H. Rais, “A survey on technical threat intelligence in the age of sophisticated cyber attacks”, Comput. Secur., vol. 72, p. 212-233, janv. 2018, doi: 10.1016/j.cose.2017.09.001.

[21] M. Gschwandtner, L. Demetz, M. Gander, et R. Maier, “Integrating Threat Intelligence to Enhance an Organization’s Information Security Management”, in Proceedings of the 13th International Conference on Availability, Reliability and Security, Hamburg Germany: ACM, août 2018, p. 1-8. doi: 10.1145/3230833.3232797.

[22] A. R. Tapsoba et T. Frederic Ouedraogo, “Evaluation of supervised learning algorithms in binary and multi-class network anomalies detection”, in 2021 IEEE AFRICON, Arusha, Tanzania, United Republic of: IEEE, sept. 2021, p. 1-6. doi: 10.1109/AFRICON51333.2021.9570886.

[23] I. Sharafaldin, A. Habibi Lashkari, et A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization”:, in Proceedings of the 4th International Conference on Information Systems Security and Privacy, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, p. 108-116. doi: 10.5220/0006639801080116.

[24] N. Moustafa et J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)”, in 2015 Military Communications and Information Systems Conference (MilCIS), nov. 2015, p. 1-6. doi: 10.1109/MilCIS.2015.7348942.

[25] A. R. Tapsoba, T. F. Ouédraogo, M. B. Diallo, et W.-B. S. Zongo, “Toward Real Time DGA Domains Detection in Encrypted Traffic”, in *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security*, in NISS ’24. New York, NY, USA: Association for Computing Machinery, August 2024, pp. 1-8. doi: 10.1145/3659677.3659684.

[26] A. R. Tapsoba, T. F. Ouédraogo, et A. E. Ouédraogo, “Relevance of the Gaussian classification on the Detection of DDoS Attacks”, in 2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Suzhou, China: IEEE, oct. 2022, p. 42-49. doi: 10.1109/CyberC55534.2022.00018.

[27] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning”, Neural Netw. IEEE Trans. On, vol. 5, p. 537-550, août 1994, doi: 10.1109/72.298224.

[28] H. Liu, L. Liu, et H. Zhang, “Feature Selection Using Mutual Information: An Experimental Study”, in PRICAI 2008: Trends in Artificial Intelligence, T.-B. Ho et Z.-H. Zhou, Éd., Berlin, Heidelberg: Springer, 2008, p. 235-246. doi: 10.1007/978-3-540-89197-0_24.

[29] Gouayon Koala, Didier Bassolé, Telesphore Tiendrebeogo, and Oumarou Sié. “Software Vulnerabilities’ Detection by Analysing Application Execution Traces.” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, January 2023. doi:10.14569/IJACSA.2023.01406136. Licensed under CC BY 4.0.

[30] M. Heydarian, T. E. Doyle, et R. Samavi, “MLCM: Multi-Label Confusion Matrix”, IEEE Access, vol. 10, p. 19083-19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[31] H. Suryotrisongko, Y. Musashi, A. Tsuneda, et K. Sugitani, “Robust Botnet DGA Detection: Blending XAI and OSINT for Cyber Threat Intelligence Sharing”, IEEE Access, vol. 10, p. 34613-34624, 2022, doi: 10.1109/ACCESS.2022.3162588.

Multiclass Fruit Detection Using Improved YOLOv3 Algorithm

Seema C. Shrawne, Jay Sawant, Omkar Chaubal, Karan Suryawanshi, Diven Sirwani, Vijay K. Sambhe
Department of CE and IT
Veermata Jijabai Technological Institute, H. R. Mahajani Marg, Matunga, Mumbai 400019.

Abstract—**Manual interventions continue to be used in fruit-picking and billing at large-scale fruit storage facilities. Recent advances in deep in learning approaches, such as one-stage detectors like You Only Look Once (YOLO) and Single Stage Detector (SSD), as well as two-stage detectors like Faster RCNN and Mask RCNN, aim to streamline the processes involved with fruit detection and enhance efficiency. However, these frameworks continue to suffer with multi-scale objects, in terms of performance and efficiency due to large parameter sizes. These problems increase when multi-class fruits are encountered. We propose an improved version of the one-stage detector framework YOLOv3 for multi-class fruit detection. Our proposed model addresses the challenges of multi-scale object detection and detection of different fruit types in an image by incorporating CNN, bottleneck, and Spatial Pyramid Pooling Fast (SPPF) modules in the Head, Neck, and custom backbone of the YOLOv3 framework. Optimization of learnable parameters for computational efficiency is achieved by concatenating features at different feature map resolutions. The proposed model incorporates fewer layers and parameters compared to YOLOv3 and YOLOv5 models. We performed extensive testing on three datasets downloaded from Roboflow and compared them with YOLOv3 and YOLOv5 models. Our model achieved mAP50 of 0.747 on Dataset 1 comprising images of apples, bananas, and oranges whereas Dataset 2 consisting of images of apples, oranges, lemon, and Pear, achieved mAP50 of 0.981. Testing the Mineapple dataset comprising on-tree images of apples of varied sizes, achieved an accuracy of 0.643. We observe that the performance of our model beats the performance of the YOLOv3 and YOLOv5 models.**

Keywords*—*Precision agriculture; yield estimation; fruit detection; YOLOv3; feature concatenation; spatial contexting

I. Introduction

With the growing population, providing food security is of utmost concern. Precision Agriculture comprises methods to optimize resources by automation of agricultural tasks like sowing, weeding, spraying, and harvesting driven by technology which helps in increasing food production [1]. Before harvesting, Yield estimation is necessary to avoid post-harvest losses of fruits caused by harvesting raw or overripe fruits [2], [3]. Accurate counting and effective invoicing of these fruits during harvest [12] and their storage in the warehouse are critical to orchard profitability. However, the current method of counting fruits, which involves physical labor, takes a long time, is prone to error, cannot keep up with the volume, and is negative to schedule management. Orchards with multiple types of fruits pose a challenge. Automating this process with robots is a viable answer [15], but these robots need powerful computer vision systems to detect and locate the different types of fruits in the orchard and warehouse environments. Adding to this, challenges, including various fruit sizes, colors, and dense foliage, make detection more difficult. The key feature of object identification is central to this vision system, allowing the robots to discern between different fruit types and perform appropriate picking and billing activities.

Fruit Detection models designed by many researchers prioritize certain properties of fruits to improve accuracy through specialized methods. Commonly used detectors include Mask RCNN[9], Faster RCNN[8], and different versions of YOLO (3, 4, 5, and 8), along with DenseNets and ResNets[11] as feature extractors. Visual attributes like color, texture, shape, and size are important properties for recognizing fruits. across different growth stages need to be considered to differentiate between fruit types. Detection of different types and sizes of fruits in an image is a challenging problem. Detecting inter-class similarities and intra-class variations is possible by a combination of low-level features and high-level semantics. In this study, we propose a fruit detection model with a custom backbone network for feature extraction at multiple levels. The YOLOv3 algorithm caught our attention, particularly through its simplicity coupled with precision without compromising speed. YOLOv3 is often used as a base model for modifications leading to continuous improvement, such as in [11], and has relatively lower training times to help achieve this. These reasons inspired us to make good use of it to develop an advanced one-stage fruit detection model. While deep convolutional networks have shown promise in fruit detection, we identified key challenges that serve as the primary objectives this research aims to solve: 1. Creation of lightweight models for practical use. 2. Effectively handling objects of different scales. 3. Achieving strong performance while maintaining efficiency rampant among fruit detection studies. 4. Training the model successfully on images such that each image has objects of different classes.

We decided to achieve this by constructing our own variant of the YOLOv3 [4], one that would take on all four challenges while providing much better results. Our proposed model addresses these challenges by incorporating special modules and optimizing parameters for computational efficiency. The model shall work on both single class as well as multi class fruit detection. We utilized three key datasets, one of which is a benchmark dataset, detailed further in Part C of the Methodology section. Detailed explanations of our novel methods are provided in the Proposed System section in this paper, showcasing our contribution to advancing fruit detection technology, while the next subsection is a tiny yet precise gist of why we chose our system in the first place.

A. Our Contributions

1. We modified the existing Darknet-53 Backbone of the YOLOv3 model by including multi-scale feature extraction and then arranging bottleneck layers to reduce the dimensionality of feature maps making the network more computationally efficient. In the head part of the model, high-level semantic features are concatenated with low-level details so that fruits of different sizes can be detected. With these modifications, our model can now facilitate the acquisition of more discriminative features by allowing gradients to flow directly during training. It also solves the challenge of detecting fruits of varying sizes.

2. Our model is trained on three datasets to prove effectiveness: Dataset 1 and Dataset 2, which consist of images of different fruits (mixed fruits), and Dataset 3, comprising of the Benchmark Mineapple Dataset consisting of apples in a dense orchard environment.

3. We have evaluated the model using various Performance metrics: Precision, Recall, mAP@50 & mAP@90 and then compared the results with YOLOv3 and YOLOv5l. Notably, running our model on Dataset 1 produced a mAP@50 of 0.747 , 0.981 on Dataset 2 and 0.643 on Dataset 3. The model achieved a higher mAP@50 on Dataset1 and Dataset3 and a higher Recall on all 3 datasets when compared with YOLOv3 and YOLOv5l models.

4. While designing the Backbone network, care was taken that the number of layers and number of parameters in our model are less then those in the standard YOLOv3 model and YOLOv5l model.

B. Organization of the Paper

The paper is organized as follows : The Section II is a detailed survey of existing research in the field, followed by Section III comprising of the dataset characteristics, our proposed system and subsequent model training. In Section IV are the results and discussions followed by the Conclusion.

II. Related Work

Before delving into our own model, we explored various research contributions, each shedding light on distinct advancements in real-time fruit detection. The study in [5] utilizes the YOLOv4 neural network to enhance real-time banana recognition in complex orchards. It addresses similarity, occlusion, and uncertainties by extracting complex features. The model, based on YOLOv4 with CSPDarknet53, includes the FPN+PAN module, SPP module, and Mish activation function. The DIOU_nms algorithm improves detection confidence. Comparisons show YOLOv4 surpasses YOLOv3 and traditional methods in accuracy and speed, with an average execution time of 0.171s and a detection rate of 99.29%. The YOLOMuskmelon model /citec2, blends speed and accuracy for enhanced fruit detection. It features a ResNet43 backbone with ReLU activation, SPP for improved accuracy, FPN for efficient feature extraction, and DIoU NMS for efficiency. With an AP of 89.6%, it outperforms YOLOv3 and YOLOResNet50 but slightly lags YOLOv4 at 91.6%. Notably, it operates at 96.3 fps, faster than YOLOv3, YOLOv4, and YOLOResNet50, highlighting its potential for real-time fruit harvesting robots due to its speed advantage over YOLOv4. A bottleneck network module C2f is the building block in the YOLOv8 model for feature extraction. In study [7] YOLOv8 model achieved mAP50 of 99.5% for ripeness detection of apples and pears. Results were compared with CenterNet model with ResNet50 backbone. A light-weight model based on YOLOv5 for real-time applications is proposed in study [8] to detect strawberry fruits. A detection speed of 7.30ms and average precision 89.7% is reported. An attention module integrated with YOLOv7 in study [9] to detect kiwi fruits. Channel and spatial features extracted by the attention module improve the accuracy of detecting small and overlapping fruits. Comparison results of YOLOv8 and Mask RCNN in [9]show that YOLOv8 is better than Mask RCNN in terms of accuracy and speed. An experiment to detect objects of two types, trunks and apple tree branches, and next to detect green apples in an orchard environment confirmed the suitability of YOLOv8 for real-time detection for applications in robotic harvesting. To enhance real-time fruit recognition speed and accuracy, [11] introduces YOLOv5s. It targets applications on low-power devices and fruit-harvesting robots [16]. Adjustments to the backbone network, adaptive image scaling, and computed anchor boxes were made using a dataset of 1,350 strawberry and 1,959 jujube photos. Improvements like Stem, AC, Maxpool, CBS, SPPF, and CAM enhance adaptability to low-power devices. Validation and test results show mAP values of 93.4% and 96.0%, respectively. Operating at 74 fps on videos, YOLOv5s outperforms models like YOLOv4-tiny, YOLOv7-tiny, and GhostYOLOv5s in robustness and efficiency. This study [12] introduces a multi-cluster green persimmon recognition approach using an enhanced Faster RCNN model. It utilizes a dataset of 9,300 images captured under diverse natural light conditions, including scenarios with multiple fruits, leaf shadows, and overlapping clusters. The upgraded model integrates a weighted ECA mechanism into three key feature layers and enhances the DetNet feature extractor to balance information levels. It incorporates multi-scale features and employs K-means clustering for bounding box clustering and anchoring. Achieving a mAP of 98.4%, the model surpasses the traditional Faster RCNN by 11.8%, demonstrating significant improvements in identifying green persimmons, especially in complex and obscure environments. An input comprising of RGB and HSV images of Oranges fed to MaskRCNN improved segmentation accuracy in [13] Next, [14] uses the VGG16 architecture and Faster R-CNN model to detect kiwifruits in orchard photos under varied lighting and time conditions. With a dataset of 2400 images at 2352x1568 resolution, each containing at least 30 kiwifruits, the model excels in recognizing kiwifruits despite occlusion, overlap, and lighting variations. Outperforming ZFNet, it proves effective in dynamic agricultural settings, ensuring high accuracy and minimal false negatives in kiwifruit identification.

In this study [11], authors enhance the YOLOv3 model for automated oil palm loose fruit identification, integrating DenseNet for feature reuse, swish activation, and multi-scale detection to improve small object accuracy. Diversifying a dataset from 700 UAV and mobile camera photos to 6300 images, they boost model performance across detection metrics. Outperforming YOLOv3, Faster R-CNN-ResNet101, YOLOv3 tiny, YOLOv2, and SSD-MobileNet, the model achieves superior average precision, overlap metrics, and F1-score while maintaining computational efficiency.

A study [12] introduces a real-time olive fruit detection system using advanced deep learning frameworks like YOLOv5x, YOLOv5s, and YOLOR. YOLOv5s combines YOLO Layer, PANet, and CSPDarknet for detection. With 40,834 annotated olive fruit images, the system achieves 62 FPS detection speed and the highest 0.75 mAP_0.5 precision, making YOLOv5s the optimal choice for automating olive harvesting challenges. The project [17] applies deep learning to enhance papaya recognition in natural orchard settings. The YOLOv5s-Papaya model integrates bidirectional weighted feature pyramid network, Ghost module, and coordinate attention module for dense multitarget detection. Utilizing mosaic data augmentation, adaptive anchor computation, and PANet framework ensures multiscale feature fusion. With 1,000 diverse photos, the model achieves 92.3% average precision, 83.4% recall, and 90.4% precision, surpassing previous YOLO versions. The working of Two-stage Detectors and YOLO architecture and its successors have been reviewed in [18] A study [19] extensively examines the YOLO series, assessing their designs, regression methods, and performance on MSCOCO and Pascal VOC datasets. YOLO models demonstrate superior detection accuracy and speed compared to two-stage detectors like RCNN, Fast-RCNN, and Faster-RCNN, making them ideal for real-time applications in machine learning and deep learning tasks. An attention mechanism to improve the localization of fruits is introduced in YOLOv5 architecture between the backbone and neck region in [20]. Results on fruits like apple, oranges, grapes on state-of-the-art models show that the proposed model has a better target detection and generalization ability.The bidirectional attention mechanism extracts features from horizontal and vertical directions, assigns weights followed by concatenation. The paper in study [21] presents the GCS-YOLOV4-Tiny model, which enhances the YOLOV4-Tiny architecture for faster fruit detection by integrating spatial pyramid pooling (SPP), squeeze and excitation (SE) modules, and group convolution. Evaluations on Mango YOLO, Rpi-Tomato, and F. margarita datasets demonstrate significant improvements over YOLOV4-Tiny, achieving a 17.45% increase in mean average precision and a 13.80% rise in F1-score. These enhancements optimize both accuracy and speed in fruit detection tasks. In [22], authors have proposed a system for selecting image regions based on features like LBP, HOG, color histograms, and shape features with a weighted score for combining features. Improvement of region proposals based on Edgeboxes are proposed. A dataset of 18,155 images like apples, pears, kiwis, and persimmon are trained on the system and then compared with DPM (Deformable Parts Model), CNN with SVM and Faster RCNN.A detection rate of 0.9632 and a FPPI (False positive per image) of 0.0682 is reported. In this study, multiple types of fruits within the same image are included in the dataset. It is found that almost all studies have performed detection on images having fruits of single type within an image. The study on muticlass fruit detection with multiple fruits within the same image is carried out in study [22] and study [10]. In study [22], a custom region selection method is proposed and has very good accuracy when compared with Faster RCNN and other models. Faster RCNN F1 score is also nearly equal to the new method. The images are multiclass but fruit instances are not overlapping or together. In study [10] multiple types of classes like trunk and branches are detected in an image.

III. Datasets and Methodology

Extensive studies over the backbone network of YOLOv3 model and family of YOLO object detectors went into the initial part of our research. Architectural patterns in the detector were interpreted to understand how the detector captures various features. The next step was to examine existing research like those seen in sSection II to understand modifications in standard YOLOv3 and YOLOv5 models and their weak points which would grant us clarity as to what novel changes we could make to the single stage detector to overcome challenges with fruit detection in particular. This section is a detailed account of the technicalities behind a standard YOLOv3 detector and then the step by step working of our proposed system. We also formally introduce each dataset with its contents and their specifications.

A. YOLOV3 Model

YOLOv3 (You Only Look Once, Version 3) emerges as a pivotal advancement in real-time object detection, brought to fruition by the collaborative efforts of Joseph Redmon and Ali Farhadi. Building upon the foundations laid by YOLO and YOLOv2, YOLOv3 represents a significant leap forward in accuracy and speed, setting new benchmarks in the field. Released in 2018, this iteration refines its predecessors' successes while introducing key architectural intricacies that elevate its performance.At its core, YOLOv3 utilizes a variant of Darknet, originally a 53-layer network trained on ImageNet, which has now evolved into a 106-layer fully convolutional architecture tailored specifically for object detection tasks. This expanded architecture enables YOLOv3 to process images comprehensively, integrating global context during inference and improving its accuracy in detecting various objects.

The algorithm operates as a Convolutional Neural Network (CNN), drawing inspiration from ResNet and FPN architectures. Darknet-53, YOLOv3's feature extractor, incorporates skip connections and three prediction heads, facilitating spatial compression and precise detections. These architectural enhancements contribute significantly to YOLOv3's ability to detect objects accurately and efficiently in real-time scenarios.

In comparative evaluations against other popular frameworks like Faster R-CNN and MobileNet-SSD, YOLOv3 consistently demonstrates its superiority. It achieves a remarkable 37 mAP on the COCO-2017 validation set at 608x608 resolution, outpacing Faster R-CNN architectures while maintaining a significant speed advantage—17 times faster, to be exact. This speed-to-accuracy ratio positions YOLOv3 as a leading choice for real-time object detection tasks, especially in scenarios requiring rapid and precise identification of objects.

Anchored on concepts like anchor boxes, k-means clustering, and a meticulously designed network structure, YOLOv3 excels in detecting small targets with exceptional accuracy, making it a preferred solution for a wide range of applications demanding robust and efficient object detection capabilities.

1) Bounding Box Prediction in YOLOv3: In the context of YOLOv3's object detection technique, bounding box prediction entails generating bounding box attributes such as coordinates and size. This operation is facilitated by 1 x 1 detection kernels, which have the following shape: 1 x 1 x (B x (5 + C)). Here, B specifies the number of bounding boxes per cell, '5' represents the box attributes (x, y, width, height, confidence), and C denotes the number of classes. This method allows YOLOv3 to correctly forecast multiple bounding boxes for each object class in an image. Object confidence is an important factor in assessing if an object is present within a predicted bounding box. This confidence measure is generated using binary cross-entropy, which assesses the likelihood of an object appearing within a certain bounding box region. A higher object confidence score implies a larger possibility of the object's presence, whereas a lower score indicates a lower probability.

Bounding Boxes Prediction:
Shape: $1 \times 1 \times (B \times (5 + C))$
Where B represents the number of bounding boxes per cell, 5 denotes the box attributes (x, y, width, height, confidence), and C signifies the number of classes.
Object Confidence:
Object Confidence = $p(\text{Object}) \times \text{IoU}_{\text{pred}}^{\text{truth}}$

B. Proposed System

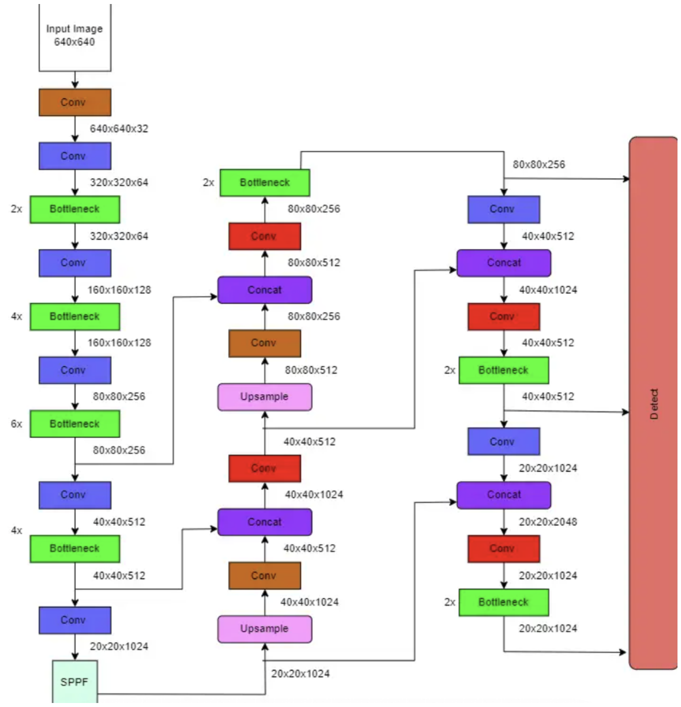


Fig. 1. Proposed system architecture.

The proposed system shown in Fig. 1 utilizes convolutional layers and bottleneck blocks in the backbone to extract hierarchical features, followed by a multiscale feature extraction process. In the head section, the system merges features from both the backbone and the multiscale extractor to enhance object detection capabilities. This integrated feature representation is then utilized for precise object localization and recognition, improving overall performance in detecting objects within images.

Our object detection system represents a sophisticated fusion of architectural components aimed at advancing object detection capabilities within deep learning frameworks. Once an image is passed, it undergoes resizing to 640 x 640 size and post this, enters the backbone of our network which contains a series of several convolutional layers, each strategically designed to capture intricate spatial patterns and hierarchical features from input feature maps.

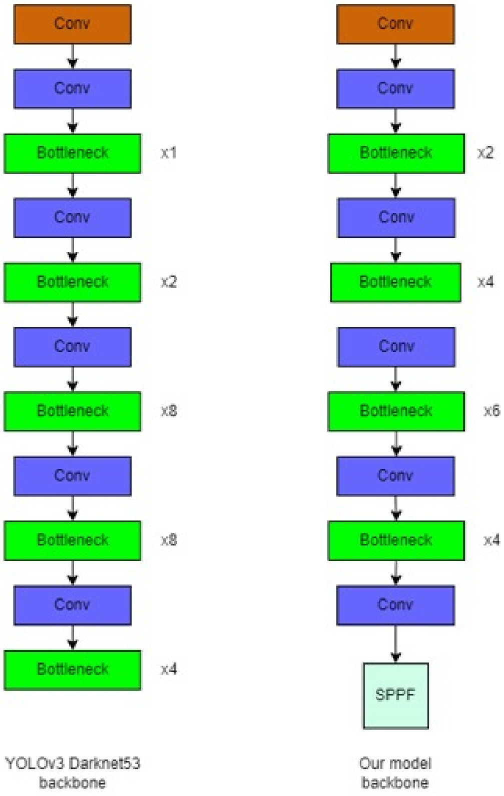


Fig. 2. Backbone of YOLOv3 and our model.

1) Custom Backbone: These convolutional layers are complemented by multiple bottleneck blocks, inspired by the modern architecture Darknet-53 as seen in Fig. 2. The bottleneck blocks in Fig. 4 play a crucial role in feature learning by incorporating residual connections, thereby facilitating the direct flow of gradients during training and mitigating the vanishing gradient problem. This approach not only enhances the model's ability to learn discriminative features but also reduces computational complexity, making the system more efficient and scalable, making it highly instrumental for deployment in resource-constrained environments like autonomous fruit-picking robots or edge computing devices in warehouses. What closely follows this is the integral part of the system's object detection prowess: an innovative inclusion of a multi-scale feature extractor called the SPPF (Spatial Pyramid Pooling Fast) layer. SPPF addresses the challenge of efficiently handling objects of varying sizes within input images. Traditional CNNs often require fixed-size input, limiting their efficacy in detecting objects at different scales. SPPF overcomes this limitation by enabling the network to operate on feature maps of arbitrary sizes, allowing for the detection of objects across multiple scales within the same image.

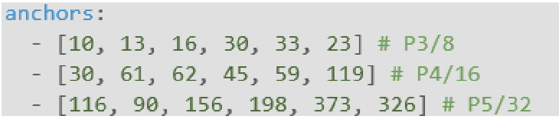
Through hierarchical feature fusion in the neck part via concatenation and multi-scale feature representation, the system achieves enhanced contextual understanding. This is crucial in a mixed fruit setting where for example the model must differentiate between a lemon and an apple. The SPPF layer's implementation of spatial pyramid pooling and factorization techniques contributes significantly to the system's efficiency, enabling it to handle variable input sizes while capturing multi-scale features effectively. What happens with this module is the network effectively partitions feature maps into progressively smaller segments. Segmentation enhances the model's ability to focus on and detect smaller objects by increasing both the resolution and receptive area, which is crucial in densely packed environments like orchards. This comprehensive architecture represents a substantial advancement in object detection methodologies, offering a scalable, efficient, and accurate solution for complex visual recognition tasks within the realm of computer vision and deep learning research.

2) Anchor Design Scheme: Anchors for different feature map sizes are shown in Fig. 3.

Anchors for feature map scale P3/8: These anchors are used with the feature map at a scale where the input image is downsampled by a factor of 8. For example, if the input image is 640x640, the feature map size would be 80x 80.

Anchors for feature map scale P4/16: These anchors are used with the feature map at a scale where the input image is downsampled by a factor of 16. For example, if the input image is 640x640, the feature map size would be 40x40.

Anchors for feature map scale P5/32: These anchors are used with the feature map at a scale where the input image is downsampled by a factor of 32. For example, if the input image is 640x640, the feature map size would be 20x20.



```
anchors:
  - [10, 13, 16, 30, 33, 23] # P3/8
  - [30, 61, 62, 45, 59, 119] # P4/16
  - [116, 90, 156, 198, 373, 326] # P5/32
```

Fig. 3. Anchors.

3) Feature Engineering Techniques: High-Level Features: These features are representations of abstract and semantic information about the input data. They typically capture complex patterns, object shapes, textures, and context within the image. High-level features are crucial for tasks such as object classification and scene understanding. As an example round form of the apple has a slight asymmetry that necessitates accurate feature detection because it can appear in varied sizes depending on the variety or maturity level. This also serves crucial when having to discern against a similar looking fruit such as in Dataset 2 we trained our model on, and most of this occurs under poor lighting or dense conditions.

Low-Level Features: These features represent more fine-grained details and local patterns in the input data. They typically capture simple structures such as edges, corners, textures, and colors. Low-level features are important for tasks that require precise localization or detection of specific visual elements. This would particularly help with oranges, bananas and pears wherefruit skin gradients and other minor colour changes can all be used to distinguish between different varieties and stages of maturity during the harvest cycle.

Fine-grained information refers to subtle, detailed, and specific visual characteristics present in the input data. It includes features such as small textures, intricate patterns, or subtle color variations.

Concat: Concatenating features from different layers of the backbone network serves several purposes- Hierarchical Feature Fusion: Features extracted from different layers of the backbone network capture information at various levels of abstraction. By concatenating these features, the model can leverage both low-level details and high-level semantic information simultaneously. This hierarchical feature fusion helps improve the model's ability to detect objects of different sizes and complexities.

Multi-Scale Feature Representation: Object detection often requires analyzing images at multiple scales to detect objects of different sizes. Features from different layers with varying receptive fields can effectively detect objects of varying sizes.This multi-scale feature representation enhances the model's robustness to scale variations in objects. The perfect example of this is the Minneapple Dataset we used (Dataset 3). Although Dataset 1 and 2 involved large scale fruits, Minneapple involved a densely packed apple orchard with several tiny fruits of very small scale, and yes, our model adapted well there too.

Enhanced Contextual Information: Concatenating features from different layers enriches the contextual information available to the model. Features from shallow layers provide fine-grained spatial details, while features from deeper layers offer more abstract semantic information. Combining these features allows the model to better understand the context of objects in the image and make more accurate predictions. Let's say in dataset 2 where both the shape of an orange as well as the understanding of its colour comes together to make accurate classification and not confuse it with a lemon.

SPPF: The Spatial Pyramid Pooling Fast layer, handles objects of various sizes within the input image. Handling Variable Input Sizes: One challenge in object detection is efficiently handling objects of different sizes within the input image. Traditional convolutional neural networks (CNNs) require fixed-size input images, which can be limiting when dealing with objects at different scales. The SPPF layer addresses this challenge by allowing the network to operate on feature maps of arbitrary sizes, enabling it to detect objects at multiple scales within the same image.

Spatial Pyramid Pooling: The SPPF layer implements a spatial pyramid pooling operation, which divides the input feature map into multiple regions of varying sizes and then pools features from each region separately. By using pooling operations with different window sizes, the SPPF layer captures features at multiple scales, allowing the network to be more robust to variations in object size.

Factorization: The term "factorization" in SPPF refers to the decomposition of the pooling operation into smaller, more manageable components. Instead of applying pooling operations directly to the entire feature map, the SPPF layer applies them to smaller regions or subregions of the feature

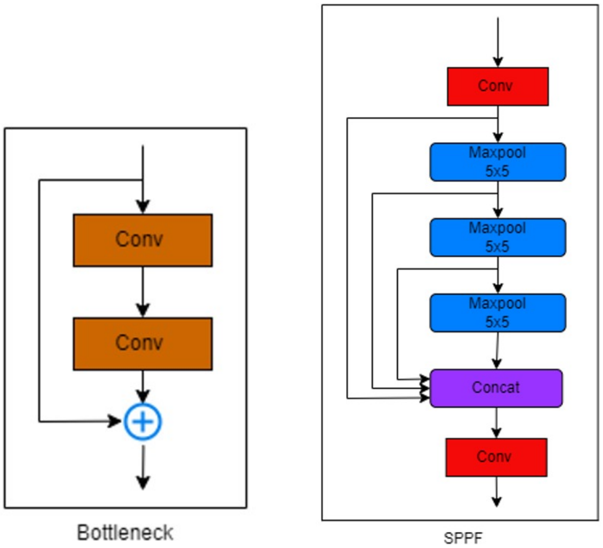


Fig. 4. Key system blocks.

map, reducing the computational complexity of the operation. This factorization process helps maintain the efficiency of the network while still capturing multi-scale features effectively.

Improved Spatial Context: By incorporating features from multiple spatial scales, the SPPF layer enhances the spatial context available to the network. This improved spatial context enables the network to better understand the spatial relationships between objects and their surroundings, leading to more accurate object detection results. A fruit picking robot working in a packed warehouse would be able distinguish between the fruit and non fruit background objects more efficiently, speeding up the process.

Detect: passing inputs of different scales with varying numbers of channels In object detection models like YOLOv3, the detection process often involves analyzing features at multiple scales to detect objects of different sizes. Features from deeper layers with larger receptive fields are better suited for detecting larger objects, while features from shallower layers with smaller receptive fields are more suitable for detecting smaller objects. The choice of having more channels in features from smaller scales and fewer channels in features from larger scales is often driven by the need to capture finer details in smaller objects. Smaller objects may require more spatial information and feature channels to be accurately detected, whereas larger objects may be adequately represented with fewer channels. By passing inputs from multiple scales with varying numbers of channels to the Detect layer, the model can effectively detect objects across a wide range of sizes. The model combines features from different scales and channels to generate bounding box predictions and class probabilities for objects present in the image.

C. Datasets

1) Dataset 1: The dataset comprises a curated collection of 150 high-resolution images, capturing three different fruits: apple, orange and banana. Each image is annotated with bounding boxes outlining the regions of interest containing the fruits. No preprocessing and Augmentation were performed on this dataset. Every model is trained for 100 epochs with a batch size of 13 and input image size of 640 x 640.

Fig. 5. Dataset 1.

2) Dataset 2: It contains images of mixed fruits categorized into four distinct classes, namely apples, oranges, lemon, and pear. Preprocessing steps specifically, RGB images were converted to grayscale using established color conversion algorithms. This transformation not only reduces computational complexity but also emphasizes shape and texture features essential for fruit detection, thereby enhancing model discriminative power. Augmentation techniques like random horizontal and vertical flips were introduced to simulate variations in fruit orientation, ensuring the model's ability to detect fruits irrespective of their spatial orientation. Additionally, rotation augmentation was employed, allowing images to be rotated by ±15 degrees around their center. This augmentation strategy introduces angular diversity, enabling the model to better generalize to fruits positioned at varying angles within the image frame. Here each model is trained for 20 epochs with a batch size of 13 and an input image size of 640 x 640.

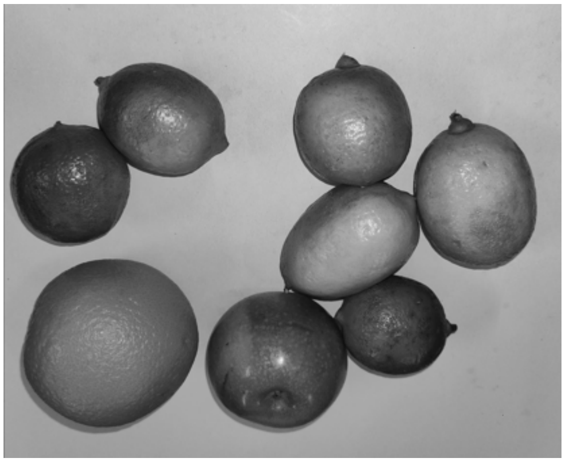


Fig. 6. Dataset 2.

3) Dataset 3: The dataset utilized is MinneApple, designed specifically for apple detection and segmentation within orchard environments, aiming to push the boundaries of fruit detection [6] technology. It focuses solely on a single class object detection: Apples. Acquired from Roboflow, an open-source computer vision tool, MinneApple has a split of 670 training images and 331 testing images totaling over 41,000

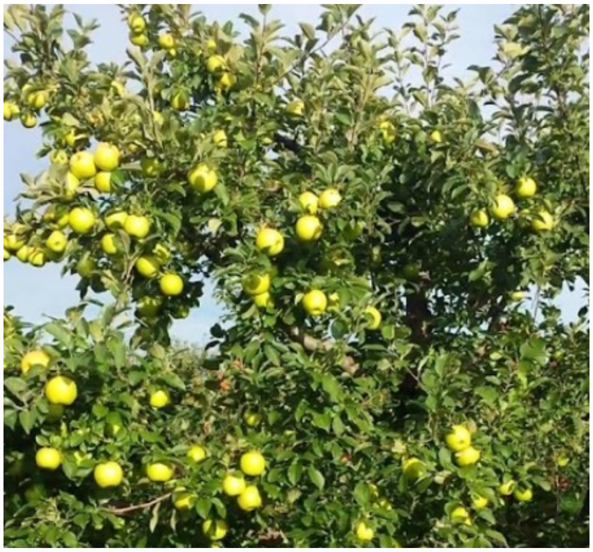


Fig. 7. Dataset 3.

annotated object instances across 1001 images. The data collection process spanned more than a year at the University of Minnesota's Horticultural Research Center, employing a standard Samsung Galaxy S4 cell phone to ensure real-world representativeness. Footage was captured at a controlled speed of 1 m/s to minimize motion blur, with images extracted at regular intervals to encompass diverse lighting, angles, and fruit ripeness stages. This diverse dataset, encompassing various fruit varieties, ripeness stages, and illumination conditions, is pivotal for training robust machine learning models capable of generalizing effectively. No data pre-processing or augmentation was applied to this dataset. All models in Dataset 3 were trained for 30 epochs with a batch size of 13 and input image size of 640 x 640.

IV. Results

In this study, we evaluated the performance of three object detection models **(YOLOv3, YOLOv51, and Our Model)** on three diverse datasets: **(Dataset 1, Dataset 2, and Dataset 3),** as displayed in Table I. We employed standard metrics including precision, recall, mAP@50, and mAP@50-95 to assess their detection accuracy.

TABLE I. Performance Metric Comparison

Dataset	Model	Precision	Recall	mAP	
				@50	@50-95
Dataset 1	YOLOv3	**0.715**	0.617	0.695	0.367
	YOLOV51	0.371	0.546	0.529	0.249
	Our Model	0.692	**0.664**	**0.747**	**0.392**
Dataset 2	YOLOv3	0.971	0.967	0.982	0.725
	YOLOV51	0.949	0.939	0.971	0.782
	Our Model	**0.978**	**0.968**	0.981	0.771
Dataset 3	YOLOv3	**0.700**	0.566	0.638	0.293
	YOLOV51	0.642	0.477	0.528	0.233
	Our Model	0.679	**0.594**	**0.643**	**0.301**

In terms of model size, we found that our model has less parameters than the YOLOv5l variant and the YOLOv3 version, as can be seen in Table II.

TABLE II. Model Details

Model	Layers	Parameters
YOLOV3	262	61,497,430
YOLOV51	368	46,119,048
Our Model	**225**	**45,403,880**

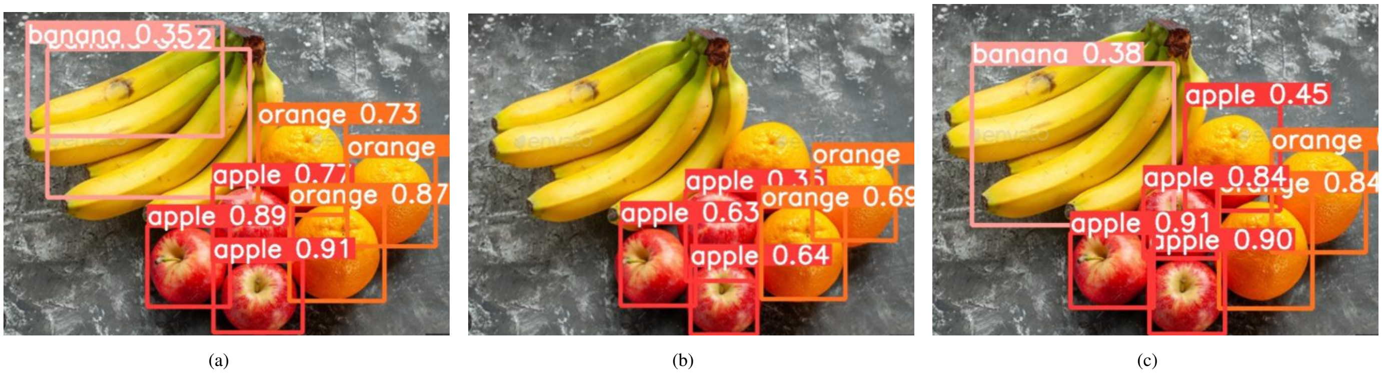
1) Dataset 1: Our model exhibits superior recall and mAP scores (both mAP50 and mAP50-95) compared to YOLOv3. While YOLOv3 has a slight edge in precision, our model's overall performance is better. When compared to YOLOv5, our model surpasses it in all measured metrics: precision, recall, mAP50, and mAP50-95. Thus, our model demonstrates a notable improvement over YOLOv3 in recall and mAP scores, and an overall superior performance across all metrics when compared to YOLOv5.

2) Dataset 2: Our model outperforms YOLOv3 in precision, mAP50, and mAP50-95, indicating superior accuracy and overall performance. However, YOLOv3 has better recall, likely due to its higher sensitivity in detecting a broader range of objects. Compared to YOLOv5, our model excels in all evaluated metrics: precision, recall, mAP50, and mAP50-95. Thus, our model demonstrates a more balanced and accurate performance overall, particularly in precision and mAP scores, while consistently surpassing YOLOv5 across all metrics.

3) Dataset 3: Our model surpasses both models, YOLOv3 and YOLOv5, in all metrics except precision, where YOLOv3 has a slight advantage. Specifically, our model demonstrates superior recall, mAP50, and mAP50-95 compared to YOLOv3.Our model excels in all evaluated metrics, including precision, recall, mAP50, and mAP50-95 compared with the YOLOv5 model.

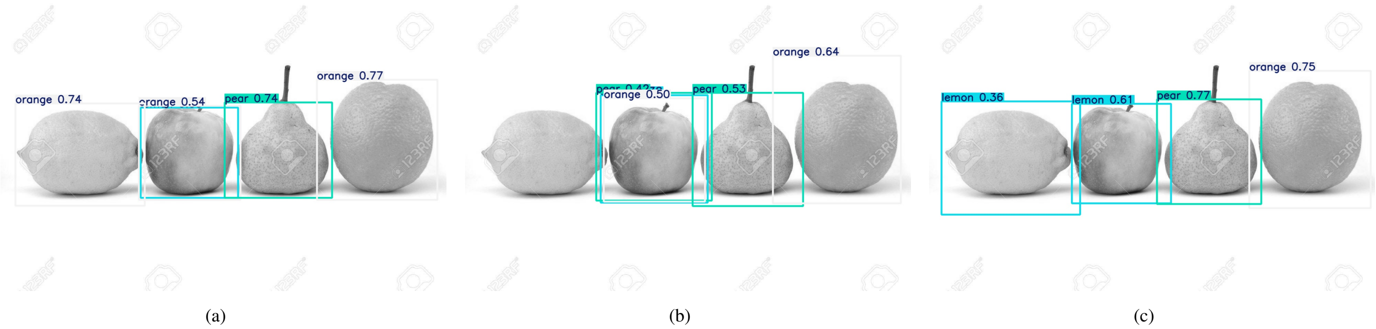
V. Discussion

The results across the datasets, when compared with standard detectors, provide valuable insights into the practical efficacy of our proposed system. For Dataset 1, the superior recall demonstrates the system's ability to detect a high number of fruit instances—apple, banana, and orange. This is particularly significant in dense orchard environments, where fruits are often obscured by leaves, twigs, and other elements within the tree canopy. The higher mAP@50 and mAP@50-95 values further highlight the model's superior performance compared to YOLOv5 and the base YOLOv3 variant in both accurately localizing and classifying fruits, particularly when predicted bounding boxes overlap with ground truth. In a multi-class dataset, this strong performance indicates the model's robustness in detecting multiple types of fruits. While standard YOLOv3 demonstrates a slight advantage in precision due to its effectiveness in reducing false positives, our model achieves a more balanced trade-off between precision and recall, which enhances the detection and classification of multiple fruits (apple, orange, banana) within a single image. The purpose Dataset 1 served was to evaluate the model's ability to detect large objects with occlusion. By using large fruit images, we can also evaluate the model's performance on objects that occupy a significant portion of the image. To perfect the system, a learning rate of 0.01 was applied for all the 100 epochs from scratch with no pre-trained weights. A split of 130 and 20 was chosen for train and validation respectively.



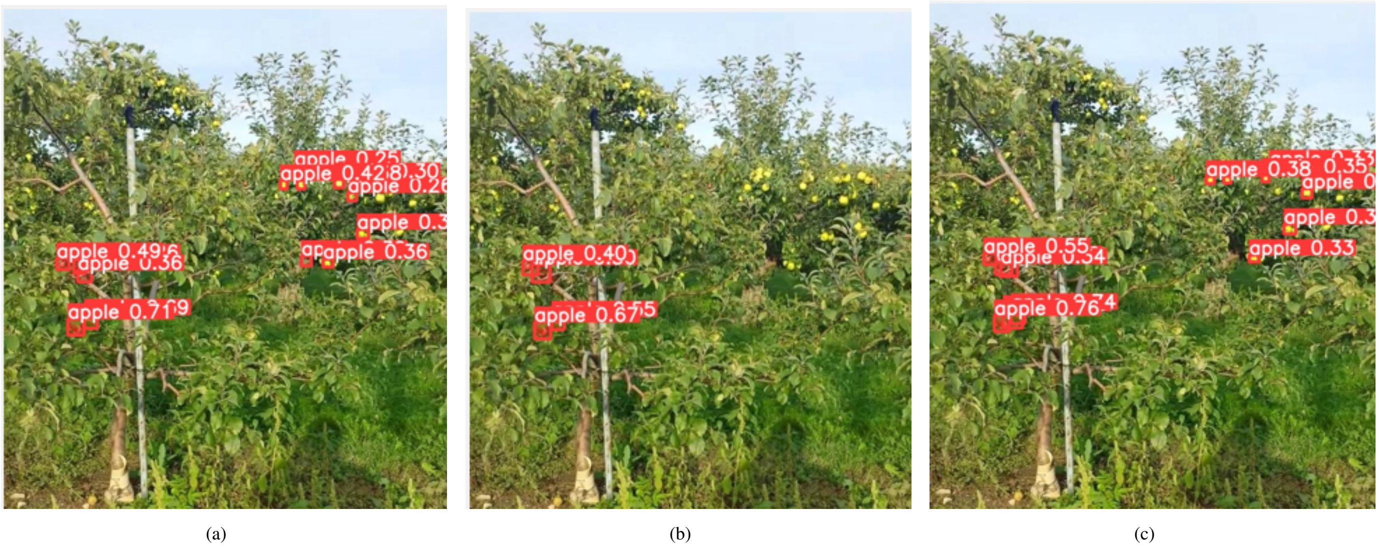
(a) (b) (c)

Fig. 8. Prediction on Dataset 1 using models YOLOv3 model(a) ,YOLOv5 model(b) and our model(c).



(a) (b) (c)

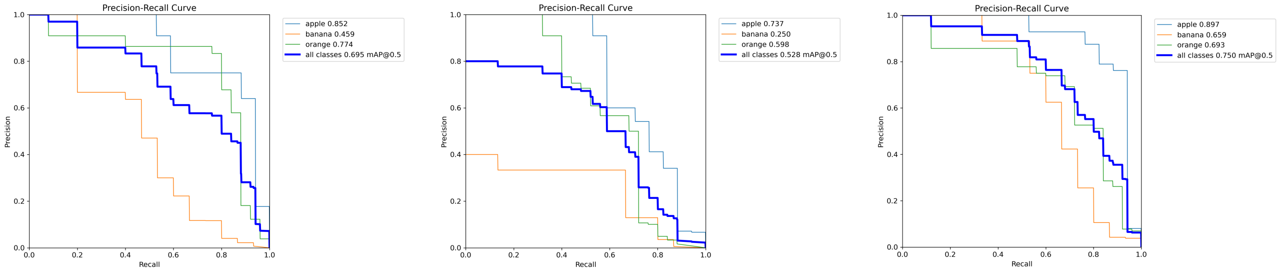
Fig. 9. Prediction on Dataset 2 using models YOLOv3(a), YOLOv5 model(b) and our model(c).



(a) (b) (c)

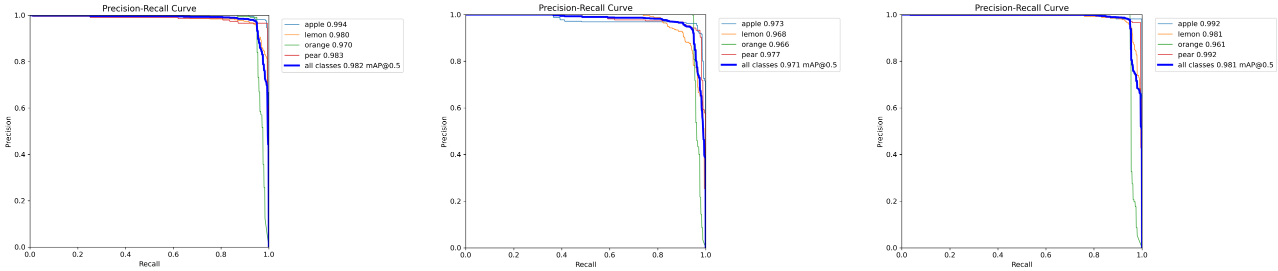
Fig. 10. Prediction on Dataset 3 by using models YOLOv3 model(a), YOLOv5 model(b) and our model(c).

In Dataset 2, our model exhibits superior performance in correctly identifying fruits without mislabeling background objects. This precision is particularly critical in controlled environments for distinguishing between similar objects or fruits of varying sizes. Although the mAP at a threshold of 50 reflects strong localization capabilities, YOLOv5l demonstrates a marginally better performance, likely due to the high similarity in shape between certain fruits, such as lemons and apples, or lemons and oranges. Despite standard YOLOv3 achieving higher recall, it compromises accuracy, as evident from the labels on the bounding boxes. This dataset had fruits of medium to large size. Further, the inclusion of grayscale



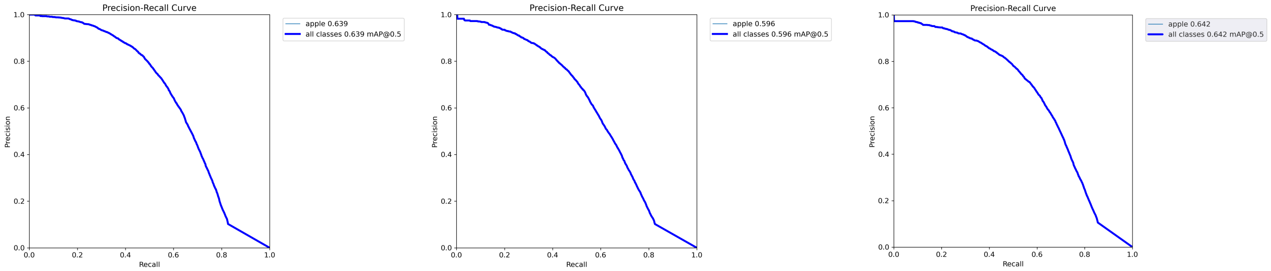
(a) (b) (c)

Fig. 11. Precision recall graph of Dataset 1 applied on YOLOv3 model(a), YOLOv5(b) model and our model(c).



(a) (b) (c)

Fig. 12. Precision recall graph of Dataset 2 applied on YOLOv3(a) model, YOLOv5 model(b) and our model(c).



(a) (b) (c)

Fig. 13. Precision recall graph of Dataset 3 applied onc YOLOv3 model(a), YOLOv5 model(b) and our model(c).

images and augmentation techniques (horizontal/vertical flips and rotation) helps increase the dataset's diversity and improve the model's generalization capabilities. A learning rate= of 0.01 was applied for all the 20 epochs with no pre-trained weights, and a split of 1450 and 230 for train and validation was chosen respectively.

Dataset 3 presents a slightly different trend. In dense orchard settings, precision is key in reducing false positives, such as branches, leaves, or occluded apples being incorrectly classified as apples. While standard YOLOv3 shows a slight advantage in minimizing these misclassifications, our model outperforms in terms of recall and mAP@50 and 90. This indicates that it detects a higher number of apples, even under challenging conditions where fruits may be partially hidden or closely packed. In such densely populated environments, the superior recall of our model is crucial, ensuring comprehensive apple detection, even when some fruits are obscured by branches. Furthermore, the higher mAP scores underscore the model's ability to accurately and consistently localize apples, particularly in cases where apples are tightly clustered, which is essential for effective yield estimation in dense orchards. We chose dataset 3 to specifically evaluate the model's performance on detecting very small objects. The MiniApple dataset provided a challenging benchmark for object detection tasks involving tiny objects. We applied a learning rate=0.01 over all 30 epochs, no pre-trained weights, and the 1001 images were kept at a split of 670 and 331 for train and validation respectively. It is noteworthy that high mAP@50 and mAP@50-95 scores proved powerful localization capabilities specially over the Minneapple dataset where similar research such as [23] which put to use MHT with YOLO and Faster RCNN, suffered at counting the fruits and had to rely on high velocity algorithms such as DeepSORT to improve performance while our system could correctly identify a very high number of tiny apples off the dense branches.

VI. Conclusion

Concerning model size, we observed that our model contained fewer parameters than both the YOLOv3 version and the YOLOv5l variant. Despite containing lower number of layers than YOLOv3, our model performed beter feature extraction. This shows that we can achieve good accuracy without a large and complex structure, making our model a great choice for tasks that need both speed and efficiency in object detection. This lightweight model can be effectively used in applications where efficiency is crucial, offering fast processing and reduced memory usage compared to larger models like YOLOv3 and YOLOv5. This certainly means we have got through with our first and third primary objectives listed at the start of this paper. In addition, the SPPF module working within our model provides a reliable solution to the issues of multi-scale fruit detection. Using pooling layers of varied sizes improves the model's capacity to detect fruits of vastly varying sizes, whether they are small in the Minneapple dataset's orchards or the large ones in the Mixed Fruit dataset. This method decreases sensitivity to variations in input resolution, resulting in consistent performance across different image qualities. Furthermore, by merging characteristics from several scales within a single feature map, SPPF enhances the fruit object representation with fine-grained details as well as global context, ultimately enhancing fruit recognition accuracy and reliability in a variety of situations. With this, we have achieved our second primary objective. In summary, our model showcases strong performance in multi-class fruit detection by taking a balanced approach and striking a worthy balance between precision and recall, as opposed to models like YOLOv3 and YOLOv5. The fact that we successfully tested over images with fruits of different classes hints at success with our fourth primary objective. Also, our model has superior overall detection accuracy, as indicated by the greatest mAP@50 score, suggesting its ability to recognize true positives despite minor differences in bounding box placement. Furthermore, our model performs consistently across IoU levels, achieving competitive mAP@50-95 values and assuring accurate fruit location pinpointing even under stringent bounding box overlap criteria. However, while our model demonstrates promising results, there is potential for further improvement. The model still does not perform well in cases of occlusion. Improvement in fusion techniques to better mix information from multiple layers is needed and for this, we can investigate sophisticated backbones such as EfficientNet. By exposing the model to a greater range of training data, data augmentation can also aid in enhancing the model's capacity for generalisation. Finally, testing the model against a variety of benchmarks can reveal the model's advantages and disadvantages as well as point out areas that still want improvement. We may improve our model's performance and attain even greater results in object detection tasks by implementing these strategies.

References

[1] Ritesh Kumar Singh, Rafael Berkvens, And Maarten Weyn, “AgriFusion: An Architecture For IoT And Emerging Technologies Based On A Precision Agriculture Survey”, September 2021, IEEE Access, PP(99):1-1, DOI: 10.1109/ACCESS.2021.3116814

[2] “Study to determine Post Harvest Losses of Agri Produce in India”,NABCONS. 2022

[3] Maheswari, Prabhakar and Raja, Purushothaman and Apolo-Apolo, Orly Enrique and Pérez-Ruiz, Manuel, “Intelligent Fruit Yield Estimation for Orchards Using Deep Learning Based Semantic Segmentation Techniques—A Review”, Frontiers in Plant Science, Vol 12, 684328, 2021

[4] Redmon, J. and Farhadi, A. (2018) YOLOv3 An Incremental Improvement. Computer Science, arXiv 1804.02767.

[5] L. Fu et al., “Fast and Accurate Detection of Banana Fruits in Complex Background Orchards”, IEEE Access, vol. 8, pp. 196835-196846, 2020, doi: 10.1109/ACCESS.2020.3029215

[6] O. M. Lawal, “YOLOMuskmelon: Quest for Fruit Detection Speed and Accuracy Using Deep Learning”, IEEE Access, vol. 9, pp. 15221-15227, (2021), doi: 10.1109/ACCESS.2021.3053167

[7] Xiao, B., Nguyen, M. Yan, W.Q., “Fruit ripeness identification using YOLOv8 model”, Multimed Tools Appl, 83, 28039–28056,2024. doi.org/10.1007/s11042-023-16570-9

[8] Lawal, O.M, “Study on strawberry fruit detection using light weight algorithm”, Multimed Tools Appl., 83, 8281–8293,2024. doi.org/10.1007/s11042-023-16034-0

[9] Xia, Y., Nguyen, M., Yan, W.Q., “A Real-Time Kiwifruit Detection Based on Improved YOLOv7”, Yan, W.Q., Nguyen, M., Stommel, M. (eds), ”Image and Vision Computing”, IVCNZ 2022. Lecture Notes in Computer Science, vol 13836. Springer, Cham.,2023 https://doi.org/10.1007/978-3-031-25825-1_4

[10] Ranjan Sapkota, Dawood Ahmed, Manoj Karkee, “Comparing YOLOv8 and Mask RCNN for object segmentation in complex orchard environments”, C.V.P.R., 2024. doi.org/10.48550/arXiv.2312.07935

[11] Lawal OM, Zhu S, Cheng K, “An improved YOLOv5s model using feature concatenation with attention mechanism for real-time fruit detection and counting”, Front. Plant Sci., 14:1153505, 2023. doi: 10.3389/fpls.2023.1153505

[12] Liu Y, Ren H, Zhang Z, Men F, Zhang P, WuDandFeng R, “Research on multi-cluster green persimmon detection method based on improved Faster RCNN”, Front. Plant Sci.14:1177114, 2023, PMID: 37346117; PMCID: PMC10279974

[13] P. Ganesh , K. Volle , T. F. Burks , S. S. Mehta (2019). ”Deep Orange: Mask R-CNN based Orange Detection and Segmentation”, IFAC-PapersOnLine,52-30,2019 70–75.Elsevier

[14] Zhenzhen Song, Longsheng Fu, Jingzhu Wu, Zhihao Liu, Rui Li, Yongjie Cui,“Kiwifruit detection in field images using Faster R-CNN with VGG16”,IFAC-PapersOnLine,52-30, 2019,76-81.

[15] M. H. Junos, A. S. Mohd Khairuddin, S. Thannirmalai, ”Automatic detection of oil palm fruits from UAV images using an improved YOLO model”,*Visual Computer*, vol. 38, no. 11, pp. 2341–2355, 2022. [Online]. Available: doi.org/10.1007/s00371-021-02116-3

[16] Ahmad Aljaafreh et.al., A Real-Time Olive Fruit Detection for Harvesting Robots Based on YOLO Algorithms, Acta Technologica Agriculturae, 2023,26,3,121-132. https://doi.org/10.2478/ata-2023-0017 ·

[17] Wang L, Zheng H, Yin C, Wang Y, Bai Z, Fu W. Dense Papaya Target Detection in Natural Environment Based on Improved YOLOv5s. Agronomy. 2023, 13(8):2019. https://doi.org/10.3390/agronomy13082019.

[18] Diwan, T., Anirudh, G.,Tembhurne J.V., Object detection using YOLO: challenges, architectural successors, datasets and applications, Multimed Tools Appl, 82, 9243–9275.2023. https://doi.org/10.1007/s11042-022-13644-y

[19] O M Lawal et al,Ablation studies on YOLOFruit detection algorithm for fruit harvesting robot using deep learning, (2021) IOP Conf. Ser.: Earth Environ. Sci., 922 012001.

[20] R. Yang, Y. Hu, Y. Yao, M. Gao, and R. Liu, “Fruit target detection based on BCo-YOLOv5 model,” *Mobile Information Systems*, vol. 2022, Article ID 8457173, 2022. [Online]. Available: https://doi.org/10.1155/2022/8457173

[21] Huang ML, Wu YS, GCS-YOLOV4-Tiny: A lightweight group convolution network for multi-stage fruit detection, Math Biosci Eng. 2023 Jan;20(1):241-268. doi: 10.3934/mbe.2023011

[22] Hulin Kuang and Cairong Liu and Leanne Lai Hang Chan and Hong Yan,“Multi-class fruit detection based on image region selection and improved object proposals”,Neurocomputing, Vol 283, 241-255,2018, doi https://doi.org/10.1016/j.neucom.2017.12.057,

[23] Juan Villacrés, Michelle Viscaino, José Delpiano, Stavros Vougioukas, Fernando Auat Cheein , “Apple orchard production estimation using deep learning strategies: A comparison of tracking-by-detection algorithms”, Computers and Electronics in Agriculture 204, [Online] Available: https://doi.org/10.1016/j.compag.2022.107513

A Meta-Heuristics-Based Solution for Multi-Objective Workflow Scheduling in Fog Computing

Gyan Singh, Vivek Dubey
Department of Computer Applications, Government Engineering College, Ajmer, India 305001

Abstract—**In recent years, there has been a significant increase in the volume of data generated by Internet of Things (IoT) applications, mostly driven by the rapid proliferation of IoT devices and advancements in communication technologies. The conventional cloud computing network was not specifically built to handle such a vast volume of data, leading to several issues, including increased processing time, higher costs, larger bandwidth usage, increased power usage, and communication delays. As a solution, conventional cloud servers have been expanded to additional layers of computing, storage, and network, termed as cloud-fog computing. The cloud-fog computing provides storage, processing, networking, and analytics capabilities in close proximity to IoT devices. The problem of scheduling workflow applications in cloud-fog environments to optimize several conflicting objectives is classified as computationally complex. Particle Swarm Optimization is the widely recognized evolutionary meta-heuristic and is the optimal method for implementing multi-objective solutions because of its user-friendly approach and quick converging capability. Despite its wide acceptance, it does have several drawbacks, such as early convergence and solution stagnation. In order to overcome these limitations, this paper establishes a comprehensive theoretical model to schedule workflow applications for cloud-fog systems. The proposed model employs various competing objectives, such as power usage, overall cost, and makespan. To achieve this, we introduce a novel algorithm, learning enhanced particle swarm optimization (LE-PSO), which incorporates an inverse tangent inertia weight policy and adaptive learning factor methods. The efficiency of the LE-PSO is subsequently assessed by employing an operational data set of scientific workflow applications within a cloudsim-based simulation and validated against GAMPSO, EMMOO, PSO, and GA state-of-the-art approaches. The workflow scheduling, we suggest achieves the substantial decrease in makespan and power usage while maintaining the total cost at an optimal level, in comparison to existing meta-heuristics.**

Keywords*—*Fog computing; DAG; workflow applications; makespan; energy; PSO

I. Introduction

In recent years, the combination of advancements in Information and Communication Technology and the rapid expansion of Big data has given rise to a new paradigm known as the Internet of Things (IoT). The Internet of Things (IoT) enables the connection of billions of tangible objects via Internet Protocol (IP). The Internet of Things (IoT) has a profound influence on diverse domains. The influence of IoT spans a diverse array of equipment, including sensors, automobiles, potable technology, cameras, patient observation devices, and numerous other applications. Moreover, the Internet of Things (IoT) has enabled the creation of a wide range of services and, applications including automotive communications, residential security, medical tracking, natural calamity prediction, scientific process automation, and roadway congestion management. [1], [2]. In accordance with predictions published by McKinsey Global Institute, the economy of IoT devices and applications will grow from $40 trillion in 2020 to $500 trillion by 2030. However, IoT end devices possess constrained capability in regards to processing, storage, and power capacities, despite generating substantial volumes of data [3]. Conversely, conventional data centers lack the capacity to efficiently handle massive volumes of data [4], [5].

In address to the aforesaid challenges Bonomi [6] proposed the concept of fog computing. Fog computing is paradigm that enhances the capabilities of current Cloud servers by extending them to network periphery, closer to IoT end devices. This improves the data center's capacity to manage workflow tasks which require increased real-time processing with agility. Hence, By establishing extra nodes in the cluster at the network periphery, referred to henceforth as cloud–fog environment.

Typically, The cloud-fog framework comprises several tiers. As shown by Fig. 1, The uppermost tier of architectural structure, known as the cloud tier, consists of multiple interconnected data centers. Optimally, the cloud tier is designed to handle and analyze jobs that need significant processing resources and can tolerate delays in data flow because of its long physical separation from the sources of data generators. The intermediate tier, known as the fog tier, serves a connection among the data centers and Internet of Things devices, usually located nearer to the IoT end-devices. Consisting of multiple machines strategically located at the network's periphery. Fog-tier machines have constrained computing, communication, energy, and data repository capacities. The fog tier machines efficiently handle time-critical operations from end-devices. If there is a task that can tolerate delays but requires a lot of computational power, it should be sent to the data centers. Fog tier nodes are hierarchically arranged to the nodes at the bottom levels. The lowest tier, referred to as the IoT end-device tier, comprises of moveable devices, sensors, cellphones, health monitoring systems, disaster prediction tools, computational science devices, and raw information generation sensors. Submission of tasks from lower tiers to upper tiers is necessary for processing requests. These devices possess extremely limited computing, repository, and energy capacities [6], [7].

Cloud-fog computing framework provides several benefits to address essential challenges. The framework exhibits increased complication and heterogeneity as a result of the various communication networks and processing nodes employed.

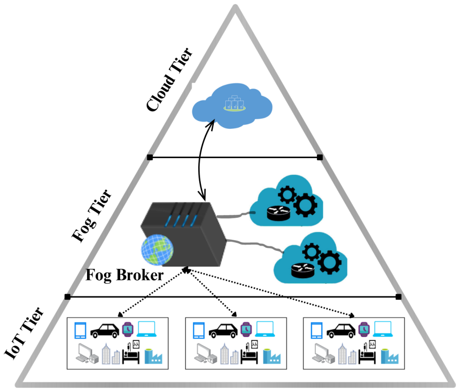


Fig. 1. Cloud-fog framework.

Workflow application scheduling is key challenge for those that need fast runtime, but poses with complex inter-dependent tasks in application and a huge volume of data moving between tasks [8].

Workflows represent the various applications that execute on end devices. The workflow applications are comprised of a set of tasks that have interdependencies. The task dependencies are arranged such that each job must wait for its predecessors to complete before it can be run. Domains such as astrophysics, patients monitoring, natural hazard prediction, computer imaging, and bio-informatics are described as workflow applications. A Directed Acyclic Graph (DAG) is a perfect data structure for representing workflow tasks and the relationships between them efficiently [9]. The aim of optimizing the scheduling of workflow applications is to provide mutually beneficial situations for cloud service providers and clients. This involves considering various parameters and constraints, such as cost, security, energy, deadlines, load balancing, budget, resource utilization, and makespan.

Fog and cloud nodes that are deployed across different geographical locations. Network technologies facilitate the interconnection of computing and storage servers. Enormous heat dissipation solutions are required to maintain essential functional temperatures of cloud data centers. Furthermore, the limited availability of resources in the end-devices tier and fog tier is contingent upon the use of uncertain power sources, like battery storage and sustainable power options. Therefore, the network of the entire cloud-fog system consumes huge amount of energy, with each activity often requiring an average of 15 megawatts. The Communication and information systems sector is responsible for emitting approximately 1.6% of worldwide CO_2 emissions between 2007-2016, in addition to its operational expenses. The predicted rise in this figure is anticipated to reach around 14% between 2016 and 2040 [10]. Consequently, researchers have shown significant interest for the development of sustainable cloud solutions. The challenge of optimizing power consumption in cloud-fog computing has become a significant concern [11]. Dynamic Voltage and Frequency Scaling (DVFS) is a widely used approach for conserving energy. All current processors are equipped with many cores to facilitate the implementation of Dynamic Voltage and Frequency Scaling (DVFS) algorithms. The DVFS technology allows individual CPUs to function at varying voltage and signal frequencies levels. However, a drawback of this technology lies in increase of cost and makespan as well as decreasing power usage. Consequently, this fails to concurrently meet the QoS requirements of both consumers and service providers.

An optimization strategy focused on minimizing the overall runtime of an application might lead to significant energy consumption and resource utilization, hence causing service providers to incur more costs. On the other hand, an algorithm that prioritizes minimizing energy usage can decrease costs for service providers but may result in a longer duration of completing tasks, ultimately causing disagreement among consumers. Hence, employing single-objective techniques to optimize scheduling algorithms is never the most optimal decision. Multiple-objective optimization methods are the most suitable option for achieving a balanced compromise among many optimization objectives. Hence, developing and improving a methodology for scheduling workflow applications in a complex system like cloud-fog, which involves more than objective with competing in nature, is a challenge that falls under the category of NP-hard [12], [13], [14]. Several workflow scheduling techniques have been developed in cloud environments to address various conflicting objectives. Nevertheless, there exist only a limited number of scheduling algorithms that effectively maximize both makespan and cost at the same time. Based on our research, currently, cloud-fog systems have no solution that optimizes the energy usage, cost, and completion time simultaneously. Several well-known evolutionary meta-heuristic methods, such as Genetic Algorithms (GA), Differential Evolution (DE), and, Particle Swarm Optimization (PSO), etc, are considered solutions for optimizing the issue of multi-objective workflow scheduling. most of the above methods focused on optimizing the cost and makespan as the optimization objectives with budget and deadline constraints. Two recent surveys have shown that PSO-based approaches are well-suited and widely used for solving scheduling problems for tasks and workflows with several objectives simultaneously [12][15]. The PSO meta-heuristic algorithms are selected for their benefits such as simple implementation, efficient execution, and rapid convergence. Although PSO has several benefits, it also has certain drawbacks like being trapped in local optima and convergence stagnation. As a result, it may either produce degraded outcomes or could necessitate additional time for computation. In order to address limitations, our research introduces the learning-enhanced particle swarm optimization (LE-PSO) algorithm. This is achieved by incorporating the inverse tangent-based new inertia weight updating and new learning factor method, which aim to strike a balance between the exploitation and exploration capabilities of the PSO methodology. The main work of this paper is listed as follows:

- Develop a theoretical framework to measure performance metrics of multiple-objective scheduling for Workflow applications in the cloud-cog environment that aims to minimize the makespan, cost, and energy usage.
- In order to overcome all of the challenges encountered in standard Particle Swarm Optimization (PSO), we present a new approach called learning enhanced particle swarm optimization (LE-PSO). This algorithm

incorporates a new inverse tangent inertia weight method and a new learning factor method, which aim to enhance both global and local search capabilities.

- Implement the suggested Multi-Objective LE-PSO Workflow scheduling in a cloud-fog environment using the Fogworkflowsim tool [16].
- We analyzed and compared the efficacy of the proposed solution with the four solutions GAMPSO [17], EMMOO [18], Standard PSO, and GA using several real-world scientific workflows. The findings showed a significant reduction in both energy consumption and makespan while not increasing the overall cost.

The subsequent sections of the article are structured in the following manner: Section II provides a comprehensive overview of the current state-of-the-art research on workflow scheduling proposed for cloud fog, fog, and cloud. Section III presents the theoretical framework for scheduling workflows in cloud-fog system. It also introduces the application model. Section IV introduces standard Particle Swarm Optimization (PSO) and the proposed learning enhanced variant of PSO (LE-PSO). Section V provides a comprehensive explanation of how the suggested algorithms can be used to map workflow tasks to a cloud-fog system. Section VI demonstrates the efficacy of the LEPSO and performance evaluation in comparison to existing solutions. Section VII finally finishes by summarizing the contribution of this article and offering insights for further research.

II. Literature Survey

The problem of scheduling in contexts that are both heterogeneous and distributed has been proven to be NP-hard. The challenge becomes more complicated when a workflow application is present and tasks necessitate a predetermined execution order to preserve the inter-dependency among them. The limitation of standard scheduling systems such as HEFT, FCFS, RR, MinMin, etc, is that they only prioritize optimizing execution time as a resource. NP-hard problems necessitate heuristic approaches to generate multi-objective approximate optimal solutions. Evolutionary meta-heuristics are the most commonly used methods for solving scheduling problems in environments such as heterogenous [12]. Various meta-heuristics-based strategies for scheduling workflows have been developed in cloud environments [13], [14], [19], [20], [21], [22]. Referenced algorithms consider either single or multiple objectives such as cost, power usage, load distribution, and completion time, considering constraints of network bandwidth, deadlines of time, and budget. Most research studies on workflow scheduling approaches have been implemented in cloud and distributed environments, with very few articles explicitly concentrating on optimizing various objectives in cloud-fog and fog environments.

In research [23], the discrete form of the Butterfly Optimization meta-heuristics is utilized to optimize workflow scheduling in a mobile-edge computing environment. The authors enhance the algorithm's ability to perform both local and global searches within the solution space by introducing a novel "Levy flight-based" equation. This meta-heuristic approach also focuses on minimizing energy consumption as the primary objective by employing the Dynamic Voltage and Frequency Scaling method. While this method significantly improves data access rates and reduces energy usage, it does not provide any statistical analysis regarding makespan or overall cost.

In study [24], the authors present IKH-EFT: An enhanced technique for workflow scheduling in fog-cloud environments using the krill herd algorithm which presents a novel method that employs the krill herd algorithm to improve efficiency in fog-cloud computing environments. This approach aims to achieve a balance between many goals, such as reducing the time required to complete a task, lowering energy usage and cost, and optimizing customer satisfaction.

In study [25], the Opposition-Based Learning variant of the Harris Hawks Optimization technique is employed to improve the scheduling of workflow applications in a multi-tier fog environment. A predictive model based on the Hidden Markov Model is developed to improve the ratio of missed deadlines and optimize task offloading. Although, this integration is designed to minimize the makespan and optimize the allocation of virtual machines (VMs), however, The findings indicate that the suggested approach is specifically applied to Fog systems, with a primary focus on minimizing only makespan.

In study [3], the work proposed introduces the Scheduling approach to mapping tasks in a cloud-fog environment that integrates two meta-heuristics techniques Genetic Algorithm and Particle Swarm Optimization. This approach seeks to achieve an Optimum trade-off between total cost efficiency and makespan while meeting customer quality satisfaction requirements. Nonetheless, this approach omits energy optimization and has not been tested on the dependent workflow tasks.

In study [26], author presents cuckoo search optimization(COA) algorithm to schedule jobs in smart grids to effectively allocate resources and enhance load balancing. COA is employed to allocate appropriate tasks to Virtual Machines (VMs) with the aim of identifying VMs that are not being fully utilized and VMs that are being excessively utilized. The under-utilized VMs are then powered off to minimize energy usage.

In research [27], the Proposed method integrates the two algorithms: Grasshopper Optimization and Symbiotic Organisms Search. This method was implemented to optimize the energy efficiency in workflow scheduling in fog environment. This integration is facilitated by employing learning automata to determine the most effective algorithm. Additionally, the algorithms utilize Dynamic Voltage and Frequency Scaling techniques to minimize energy usage in fog computing environments, while also managing to maintain an optimal makespan. Although this method successfully lowers energy consumption, the complexity is heightened due to the hybridization of algorithms, and the aspect of overall cost management is not addressed.

In research [28], introduces a novel meta-heuristic algorithm called the Hybrid Particle Whale Optimization Algorithm (PWOA). The PWOA combines two well-established algorithms: Particle Swarm Optimization (PSO) and Whale Optimization Algorithm (WOA). The hybrid approach leverages the advantages of both PSO and WOA, aiming to enhance performance while reducing the limitations inherent in each individual algorithm. The algorithm optimizes the overall execution time and overall cost for workflow scheduling in a cloud-fog environment. The algorithm omitted energy consumption as a parameter.

In study [29] continuation with PSO and GA combinations studies, this article introduces a novel method by hybridizing the GA and PSO meta-heuristics to improve the optimization of workflow scheduling within a cloud-edge environment. The author harnesses the Genetic Algorithm's two operators: Mutation and selection to increase the randomness in the diversity of the population, while simultaneously implementing the inertia update with the Non-linear method to facilitate a trade-off between global and local searching processes. Although this method shows superior performance in minimizing makespan and cost, it ignores the energy optimization parameter.

In study [30], the author suggests using the Multi-objective Hybrid Dragonfly Algorithm (MHDA) as a hybrid optimization tool to improve task scheduling for Big Data applications in IoT cloud environments. The objective of this algorithm is to minimize the makespan, which refers to the entire time needed to perform a given collection of tasks, while simultaneously maximizing resource use. The scheduling process takes into account key criteria such as makespan, resource utilization, and cost. The main focus of this research is on IoT cloud computing environments, which involve the integration of Internet of Things (IoT) devices with cloud computing resources.

In study [31], the author proposes a list-based dependent task scheduling algorithm for fog computing, which operates in three phases: sorting, prioritization, and selection. In the sorting phase, all independent tasks are organized. In the prioritization phase, tasks are assigned priorities based on their successor tasks. Finally, in the selection phase, task submission decisions are made by balancing global and local search strategies. This approach aims to optimize cost and makespan as its objective functions. However, a limitation of this work is the absence of a meta-heuristic approach; as the number of workflow tasks increases, both makespan and cost also increase.

In study [32], the author introduced a bi-objective workflow scheduling method aimed at optimizing both scheduling reliability and workflow makespan simultaneously in a cloud-fog environment. The model addresses the problem as a multi-criteria challenge, aiming to enhance scheduling reliability while improving the service delivery time ratio for workflow tasks allocated to computing resources. furthermore, it develops a reliability-deadline optimization model, which seeks to optimize application execution time and reliability. This is achieved by adapting a reliability-recursive optimization approach and remapping workflow applications that are not on the critical path. however method neither considers the energy as function objectives nor any meta-heuristics methods employed.

In study [17], a continuation with another PSO-GA-based method introduces a novel workflow scheduling algorithm by hybridizing the Genetic algorithms with the Particle swarm optimization algorithm. The algorithm starts with random initialization of GA and then passes the second half of iteration to PSO. It considers function objectives energy consumption, incurred cost, and makespan. due to the hybridization of GA and PSO, the algorithmic complexity is increased.

In study [18], a multi-objective approach to workflow scheduling is introduced with the goal of minimizing both energy use and makespan while adhering to deadline constraints. The proposed approach operates in two phases. Initially, a task priority queue is generated using Estimated Processing Times to schedule tasks. The subsequent phase employs a DVFS technique to reduce energy consumption without compromising deadline adherence. The idle slots calculated in the initial phase are leveraged in the next phase, eliminating the need for rescheduling. Although this method effectively reduces makespan and consumption of energy, it neither considers cost nor incorporates any meta-heuristic algorithms.

In [33], the author implements a Shortest Job Queue-based job scheduling algorithm for fog environments. The algorithm utilizes the shortest job queue to reduce the application loop delay and network time for submitted applications. It considers factors such as response time, energy consumption, and network usage as a defined objective function. However, the algorithm has not been tested for dependent task submission, and no meta-heuristic approaches have been applied.

The table labeled as Table I provides a summary of the underlying concepts of algorithms, performance measurement metrics, evaluation tools, the context where the algorithm deployed, and challenges/open issues of the literature survey conducted. Prior research in the field has extensively focused on enhancing the efficiency of workflow scheduling in cloud environments. Nevertheless, there is a significant shortage of studies pertaining to fog and fog-cloud. Furthermore, as observed from the literature studies the majority of existing studies primarily concentrate on reducing the time it takes to complete a task and related costs while disregarding the inclusion of energy usage as a key goal. This omission is particularly noteworthy considering the present importance of energy efficiency. In addition, the study shows a preference for using evolving meta-heuristics frameworks to schedule workflows in fog-cloud computing. The primary implementation issue with meta-heuristics methods is the complexity inherent to them. The implementation of meta-heuristics has a tendency to rise the time to complete algorithms, which may not be acceptable for the client's applications that require fast response times. PSO is considered the most appropriate choice for multiple-objective optimization among the many evolutionary meta-heuristics. Nevertheless, it is hampered by certain deficiencies. To improve the applicability of PSO in Workflow scheduling and overcome its drawbacks in cloud-fog, more improvement and refinement are necessary. In order to address the challenges specified in Table I, we have implemented a learning enhanced particle swarm optimization (LE-PSO) for the purpose of scheduling workflows in a cloud-fog environment. This technique takes into account many conflicting objectives, such as Energy, Makepsan, and cost, as the main optimization metrics.

III. System Modeling and Problem Formulation

In this section, We first establish a multi-tier resource model, within which proposed workflow scheduling will be executed. Next, the Dircet Acyclic Graph-based workflow application is modeled with multiple but competing objectives. Then, the problem is mathematically formulated as multiple-objective optimization.

TABLE I. Comparison Summary of Recent Cloud-Fog-Based Workflow Scheduling Methods

Ref.	Underlying Concepts	Performance metrics	Simulator	Environment	Challenges and Open issues
[23]	Discrete version of Butterfly optimization	data access ratio, Energy	iFogSim	IoT-edge	No cost/Makespan mentioned
[24]	Multi-objective krill herd algorithm	Makespan, cost, energy	Matlab	Fog-cloud	complex to implement
[25]	Harris Hawks Optimization and Hidden Markov Model	task offloading,missed deadlines, makespan	iFogsim	Fog	No enrgy and cost, No complex environment
[3]	Hybrids PSO with GA	total cost, Makespan	iFogsim	Cloud-Fog	No Energy optimized, No workflow task evaluated
[26]	Cuckoo optimization algorithm	Energy consumtion, response time	cloudsim	cloud-fog	no task dependency,only energy
[27]	Integrate Symbiotic Organisms with Grasshopper Optimization	Energy	iFogsim	Fog	No muil-tier, increased complexity,only energy
[28]	Hybrids PSO with Whale Optimization	Execution Cost, Execution Time	WorkflowSim	cloud-fog	Energy not considered,increased complexity
[29]	Hybrids Non-linear PSO with GA	Total Execution time, total cost	WorkflowSim	Cloud edge	only cost,execution accounted for
[30]	Multi-objective Hybrid Dragonfly Algorithm	Cost, makespan, resource utilization	cloudsim	Cloud-IoT	No energy considered
[31]	list-based task scheduling	Cost, makespan	iFogsim	fog	Non-meta-heuristics employed,under performed in higher load
[32]	simple heuristics based	reliability,makespan	ifogsim	cloud-Fog	no meta-heuristics, No energy parameter
[17]	Hybridization of PSO with GA	energy consumption, cost, Makespan	Workflowsim	cloud-Fog	complexity increased
[18]	DVFS with Sorted priority queue	makespn,energy	MATLAB	cloud-Fog	Meta-heuristics not used, No cost
[33]	Shortest Job Queue-based	energy, response time, network	iFogsim	fog	No dependent tasks, no meta-heuristics

A. Resource Model

The network architecture of the cloud-fog system is structured into three separate tiers: cloud tier, fog tier, and IoT end devices tier. The Fog Broker node is a specialized server node in the fog tier, as depicted in Fig. 1.

The broker node is responsible to coordinate and facilitate communication between three tiers in order to efficiently allocate tasks to appropriate resources, whether they are located in the cloud or fog. The broker node works depending on optimization goals specified by the service provider.

In the proposed research, we create a heterogeneous environment comprising three distinct categories of resources. These include cloud nodes located in reserved high-capacity storage and computing host machines that communicate via a Wide Area Network within the cloud tier; fog nodes that are interconnected through Ethernet in the fog tier; and IoT devices that possess very constrained processing, storage, and communication capabilities, connect to nearby fog nodes over the mobile network, forming end-devices tier. We denote C as a set of virtual machines provisioned on cloud tier, defined as $C = vm_1, vm_2, vm_3, \ldots, vm_c$. Similarly, F denotes the VMs deployed on fog tier, where $F = vm_1, vm_2, vm_3, \ldots, vm_f$. The VMs generated on end-user devices are primarily intended for managing minor tasks. Tasks that need higher computing and storage requirements will be offloaded to the upper tiers. Consequently, resources available on end devices are considered to have minimal significance in terms of performance. $V = C \cup F = vm_1, vm_2, vm_3, \ldots, vm_m$ are the sum of VMs provisioned at cloud-fog and registered at Fog broker node, reflecting their distinct technical specifications.

B. Application Model

Workload applications submitted to cloud-fog will be modeled by direct-acyclic-graph, denoted as $G(T, E)$, Where T is a set of tasks and E is the communication link between tasks. E represented a set of one-way inter-dependencies among tasks. Every task $\tau_s \in T$ in set T have thier size defined by *size(τ_s)*.Every edge $\epsilon_{ij} = \langle \tau_i, \tau_j \rangle \in E$ signifies a dependency of flow or control between tasks τ_i and τ_j. This indicates that τ_i must be finished before τ_j can begin, thereby establishing τ_i as a predecessor of τ_j due to the precedence constraints.The collection of all tasks that precede a given task is represented as *pred(τ_s)*, while the set of tasks that follow is represented as *succ(τ_s)*. The edge ϵ_{ij} is associated with communication weight cv_{ij} with a value zero or greater, representing the amount of data transferred from task τ_i to task τ_j. In cases where a task lacks a predecessor or successor, pseudo task τ_e / τ_x equals zero is introduced into the graph. As shown in Fig. 2, We have created an example to illustrate workload application DAG. It demonstrates how tasks are organized across seven distinct levels. In the third level, tasks τ_2, τ_3, and τ_4 are designed to execute concurrently.

C. Problem Formulation

1) Makespan objective: Total time needed to complete a workload application, measured from the starting time of the initial task, τ_e, to the finishing time of the final task, τ_x, is referred as Makepan. For our proposed model, two distinct categories of VMs have been created: c number of virtual machines in cloud tier and f number of virtual machines in fog tier. Thus, The specialized node fog broker is registered with the sum of VMs ($m = c \cup f$) at the specific moment. Submitted tasks of workload are executed on cloud virtual machines or fog virtual machines depending on the selected offloading strategy by the fog broker to reduce the makespan for the entire workload. When a task τ_s of workload is assigned to a VM deployed at node l, the execution time for τ_s is represented as:

$$ET_{\tau_s^l} = \frac{size(\tau_s)}{PC^l} \qquad (1)$$

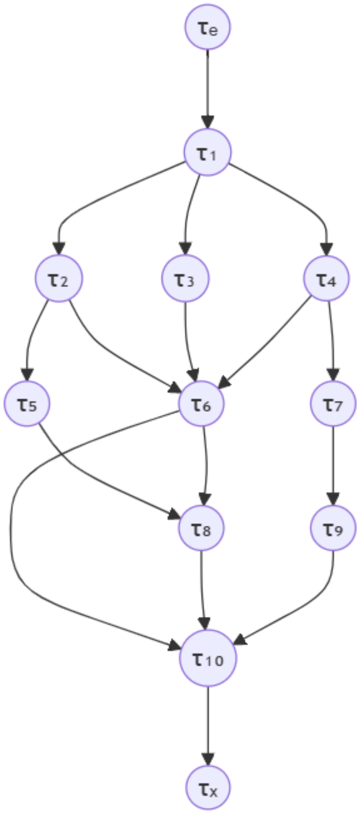
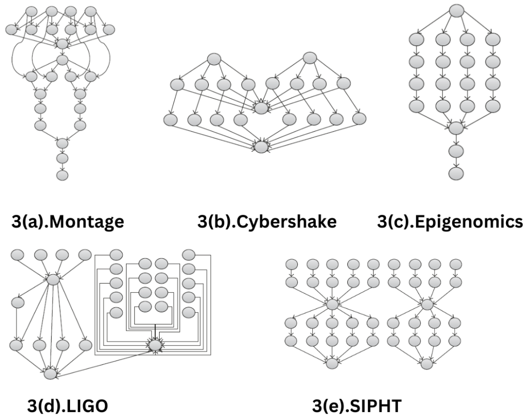


Fig. 2. A DAG-based workload illustration.


3(a).Montage 3(b).Cybershake 3(c).Epigenomics

3(d).LIGO 3(e).SIPHT

Fig. 3. Structure of various workloads.

Where $ET_{\tau_s^l}$ represents the time required to execute task τ_s on the l^{th} physical machine, *size*(τ_s) denotes the size of the task, PC stands for the processing capacity for l^{th} physical machine.

Consider $DT(\epsilon_{ij}^l)$ as the data transfer time between tasks τ_i and τ_j, which can be calculated as follows:

$$DT(\epsilon_{ij}^l) = \frac{\mathit{size}(cv_{ij})}{B^l} \tag{2}$$

size(cv_{ij}) represents the volume of data exchanged between tasks τ_i and τ_j, while B denotes the bandwidth available between the physical machine.

The start time ST_{τ_s} and finish time FT_{τ_s} for task τ_s can be determined using the following formula:

$$ST_{\tau_s} = \max(FT_{\tau_p} + DT(e_{ps}^l), \tau_p \in pred(\tau_s)) \tag{3}$$

The set $pred(\tau_s)$ includes all the preceding tasks of task τ_s.

$$FT_{\tau_s} = ST_{\tau_s} + ET_{\tau_s^l} \tag{4}$$

Therefore, the overall makespan of the workload is:-

$$MS = maximum(FT_{\tau_s} - minimum(ST_{\tau_s}), \tau_s \in pred(\tau_s))) \tag{5}$$

2) Cost objective: The service architecture of cloud fog model operates on a business-driven on-demand prices, where pricing structures for their solutions are established by service providers. In this cost model, total expenses are derived from two resource types: computing resources and data transfer communication resources. Both cloud and fog machines are accounted for to determine the total cost, whereas the costs related to end devices are assumed to be minimal.

The computational cost for executing a task τ_s at the physical machine l is determined by:

$$CC_{\tau_s^l} = UC^l * (FT_{\tau_s} - ST_{\tau_s}) \tag{6}$$

CC represents the computational cost, UC^l denotes the unit price on physical machine l.

The cost of communication between tasks τ_i and τ_j will be determined using the following calculation:

$$CD(\epsilon_{ij}^l) = UD^l * DT(\epsilon_{ij}^l) \tag{7}$$

$CD(\epsilon_{ij}^l)$ represents the cost of communication for transferring data between task τ_i and task τ_j, UD^l denotes the unit price. When tasks τ_i and τ_j are processed on the same physical machine, the amount of data transfer $DT(\epsilon_{ij}^l)$ is equal to 0.

Consequently, the sum of cost is determined by:

$$TC = \sum_{l=1}^{m} \sum_{s=1}^{n} CC_{\tau_s^l} + \sum_{l=1}^{m} CD(\epsilon_{ij}^l) \tag{8}$$

Here, m denotes the VM counts, and n denotes the task count.

3) Energy objective: We have derived energy objectives from the framework outlined in [34], which divides energy consumption into two key parts: static and dynamic consumption of energy. The dynamic part of consumed energy denoted as ED, is particularly significant as it leads to higher energy usage during the execution of workload tasks.

$$ED = \sum_{s=1}^{n} C * f_s * {v_s}^2 * (FT_{\tau_s} - ST_{\tau_s}) \tag{9}$$

Here C represents the constant value as a capacitive load, v denotes the supply voltage, where f is the frequency of the CPU where executing the task τ_s.

ES denotes the static part of energy utilized, which functions at the lowest possible frequency with the minimum voltage supply required to keep the system operating. While the system is idle, it carries out crucial functions like functioning core circuits, supporting the continuous working of RAM and input/output activities, and clock running.

$$ES = \sum_{l=1}^{m} \sum_{idle_{l,k} \in IDLE_{l,k}} C * f_{l,mn} * v_{l,mn}^2 * S_{l,k} \tag{10}$$

Where $IDLE_{l,k}$ represents the collection of all idle time slots available on physical machine l, characterized by the minimum frequency $f_{l,mn}$ and minimum supply voltage $v_{l,mn}$, additionally, $S_{l,k}$ denotes the slot of idle time utilized by l.

The sum of utilized energy for a given workload is determined by:-

$$TE = ED + ES \tag{11}$$

D. Fitness Metric

Most evolutionary techniques begin with a randomly initialized population. In Evolutionary meta-heuristics methods, the initial phase of population is distributed inherently stochastic.To minimize skews in various objectives, a normalization process is applied by min-max scaling technique. This process involves following steps for normalizing all three objectives:

$$MS_{adj} = \frac{MS_i - MS_{mn}}{MS_{mx} - MS_{mn}} \tag{12}$$

$$TC_{adj} = \frac{TC_i - TC_{mn}}{TC_{mx} - TC_{mn}} \tag{13}$$

$$TE_{adj} = \frac{TE_i - TE_{mn}}{TE_{mx} - TE_{mn}} \tag{14}$$

In this context, “mn” and “mx” represent the lowest and highest fitness values. while i signifies the iteration number as the solution evolved. The “adj” refers to the normalized fitness value.

Following this, the fitness value of an objective for the mapping of task-to-virtual machine schedule p is determined by employing the normalized objectives, as defined by:-

$$f(p) = \alpha.MS_{adj} + \beta.TC_{adj} + \gamma.TE_{adj} \tag{15}$$

Here α, β, and γ represent weighted constants that satisfy the condition $(\alpha + \beta + \gamma = 1)$.Service providers have the flexibility to adjust the value of the constant based on their specific service requirements and preferences. In this case, constants α,β, and γ are selected in the ratio (0.3:0.3:0.4), reflecting that cost and makespan are given equal importance, while slightly higher importance is given to energy.

IV. LEPSO

This section begins by presenting the core principles of the standard particle swarm optimization algorithm, PSO is extensively used to optimize a wide range of complex problems with multi-objective. Following this, we explore an enhanced version specifically designed for workflow scheduling in cloud-fog systems, with a focus on managing conflicting objectives.

A. Standard PSO

Standard PSO, developed by “Kennedy and Eberhart” [35], is a multi-objective, evolutionary, stochastic optimization technique. It has gained popularity for its simplicity and quick convergence, making it a widely recognized algorithm. PSO is particularly useful for tackling complex challenges such as scheduling tasks of workflows in heterogeneous computing systems [12]. The core idea of the algorithm involves particles that search for the best solutions by coordination among particles, and exchanging information between particles. PSO was inspired by the navigation and foraging behaviors observed in schools of fishes. Consider K particles involved in searching within a D-dimensional search space. Each particle, denoted as p_k (where $k = 1, 2, \ldots, K$), is defined by its position $x_k = (x_{k1}, x_{k2}, \ldots, x_{kD})$, velocity $\upsilon_k = (\upsilon_{k1}, \upsilon_{k2}, \ldots, \upsilon_{kD})$ within the landscape. For the i^{th} iteration during the evolution process, the velocity of the k^{th} particle will be modified as follows:

$$v_k^i = \omega^i \cdot v_k^{i-1} + c_1 \cdot r_1 \cdot (pBest_k - x_k^{i-1}) + c_2 \cdot r_2 \cdot (gBest - x_k^{i-1}) \tag{16}$$

The equation defines three components: inertia weight, particle self-learning component, and particle social-learning component. The term $pBest$ denotes the best position achieved by the i^{th} particle during the iteration, while $gBest$ represents the optimal position across the entire search landscape. The variables r_1 and r_2 are random values between 0 and 1. The inertia weight ω and the self-learning (c_1) and social-learning (c_2) factors regulate the problem’s exploration and exploitation capabilities within the search landscape.

For the k^{th} iteration of evolution, each particle’s position will be modified as follows:

$$x_k^i = x_k^{i-1} + v_k^i \tag{17}$$

In standard PSO, the inertia weight is a crucial element that influences both local as well as global exploration capabilities throughout the process of evolution. It is computed using a linear decreasing method as follows:

$$\omega^{i} = w^{mx} - \frac{(w^{mx} - w^{mn}) * i}{T_{mx}} \tag{18}$$

Here, w^{mx} and w^{mn} denote the maximum and minimum inertia weights, respectively, while T_{mx} represents the maximum number of iterations, as specified in Table II. A linear decrease in inertia weight means that the global search capability is broader in the early iterations, while the local search capability becomes more focused later on. In various standard benchmarks, $c1$ and $c2$ are uniformly assigned a value of 2 to ensure a proper trade-off between global and local search abilities.

B. Learning Enhanced PSO

The standard Particle Swarm Optimization (PSO) method is straightforward to implement and excels in global search across multiple objectives. However, when applied to workflow scheduling in complex environments, it often yields suboptimal results due to its tendency to get stuck in local optima, which disrupts the balance between local and global search capabilities. Key parameters influencing search performance include the learning factors(c1,c2) and inertia weight(ω). To overcome these issues, dynamic mechanisms for both the learning factor and inertia weight are employed.

1) Enhanced learning factor: c1 and c2 are crucial for managing the balance between global and local search abilities. Both factors are usually assigned constant values derived from past benchmarking. The information given outlines various scenarios depending on c1 and c2.

- When both c1 and c2 are set to 2, the setup seeks to strike a balance between global and local search processes. Nevertheless, this is crucial to note that every iteration in the Particle Swarm Optimization (PSO) process does not benefit from the self-learning ability of previous iterations.
- If $c1 = 0$ and $c2 = 2$, the system does not utilize learning solutions and rapidly converges on local search areas, often yielding sub-optimal outcomes. This scenario suggests less diversity in population and less exploration within the landscape.
- When $c1 = 2$ and $c2 = 0$, particles do not exchange information. This setup prevents the algorithm from converging and reduces its efficiency, as there is no collaborative interaction among the particles.

These findings highlight the need to incorporate a new learning factor approach in PSO to balance global and local search abilities effectively to achieve better optimization. Consequently, we propose adaptive learning factors, c1, and c2, which are computed during each evolution as follows:

$$c_1^i = c^{mn} + \frac{(c^{mx} - c^{mn})(T_{mx} - i)}{T_{mx}} \tag{19}$$

$$c_2^i = c^{mn} + \frac{(c^{mx} - c^{mn})(i)}{T_{mx}} \tag{20}$$

Here, $[c^{mx}, c^{mn}]$ denote the range defined in Table II, T_{mx} maximum number of iteration , i represents current iteration. In the initial phases of this enhanced approach's iteration, focus is on minimizing the self-learning of particles while amplifying social learning. This modification improves global search ability, thereby exploring larger area of landscape. As the process moves into subsequent phases, the approach shifts: the self-learning aspect is elevated, and social learning aspect is reduced. This alteration is aimed to strengthen local search, thereby accelerating convergence toward the optimal global solution.

2) Inertia weight: The self-learning and social-learning term as defined by Eq. (16) is solely accountable for gradually converging the fitness value toward the zero throughout the iterations. This behavior of the equation is governed by the inertia weight ω. If the decreasing rate of w is so fast from ω^{mx} to ω^{mn}, premature convergence may occur, resulting in suboptimal outcomes. If the fitness of the swarm remains immutable for an extended period, there is a high likelihood of encountering a stagnant situation, potentially leading to the degradation of algorithm performance. It is crucial to attentively adjust the inertia weight to address the challenges of stagnating particles and early converging to sub-optimal solutions Even Applying linear decrease in ω as defined in Eq. (18) may lead to an imbalance between pBest and gBest, resulting in premature convergence or stagnation. In order to resolve this problem, we utilize a method with dynamic inertia weight that applies a non-linear inverse tangent decreasing function, computed as defined below.

$$\omega^{i} = \omega^{mn} + (\omega^{mx} - \omega^{mn}) \times \frac{\arctan\left(a\left(\frac{i - \frac{T_{mx}}{2}}{T_{mx}}\right)\right)}{\arctan\left(\frac{a}{2}\right)} \tag{21}$$

exploration and exploitation are controlled by a=10, which affects how quickly the inertia weight transitions from higher to lower values, becomes dominant as the arctan function flattens out, leading to more stable convergence.

By utilizing non-linear inverse tangent inertia weight and, enhanced learning factors, the modified velocity equation is calculated as follows:

$$v_k^i = \omega^i \cdot v_k^{i-1} + c_1^i \cdot r_1 \cdot (pBest_k - x_k^{i-1}) + c_2^i \cdot r_2 \cdot (gBest - x_k^{i-1}) \tag{22}$$

V. Proposed LEPSO Workflow Scheduling Algorithm

We have implemented cloud-fog workflow scheduling algorithms by utilizing our LEPSO meta-heuristics in this section. Prior to implementing the algorithm, a mapping model is created to correlate the different terminologies utilized in LEPSO with elements of cloud-fog systems.

A. Mapping PSO Particles to Cloud-Fog Resource

A particle in the PSO search space represents a potential solution, making it essential to design particles that can yield optimal results for workflow scheduling. Given that workflow scheduling is inherently a discrete problem, natural numbers are used to represent particles. In a cloud-fog environment, 'm' virtual resources are provisioned across three categories: virtual machines deployed on the cloud, fog, and IoT end-devices. A set of 'n' tasks in the workload is then mapped to these provisioned virtual resources. The workload's task count determines the dimension 'D' of the particle. Every particle within the PSO population is mapped to a vector consisting of natural numbers. In this vector, each index represents a specific Task ID, while the value at that position denotes VM ID to which the corresponding task is allocated for execution

Consider, for example, a scenario depicted in Fig. 2, where a workload consists of tasks $n = 10$ and there are virtual machines $m = 5$ deployed: 2 in the cloud, 2 in the fog layer, and 1 on IoT end devices. The function of IoT end-devices VM is only to push the workload application to fog broker. The mapping of tasks to virtual machines can be represented by a particle 'P', which is expressed as:

3	3	4	1	2	4	2	4	1	5

Particle 'P' as a vector

The above Task-VM schedule represents allocating every task to map on the designated virtual machine. i.e. , Task1 is allocated on VM3, Task2 is also allocated to VM3, Task3 goes to VM4, and so forth.

B. Implementation of LEPSO Workflow Scheduling

By employing a particle encoding method and the objective fitness function defined in Eq. (15), the algorithm is executed as detailed in Algorithm 1. The input to the algorithm consists of a Directed Acyclic Graph workload G containing a set of tasks $T = \{\tau_1, \tau_2, \ldots, \tau_n\}$ and a collection of virtual machines $V = \{vm_1, vm_2, vm_3, \ldots, vm_m\}$. The output generated is the optimized task-to-VM mapping schedule (gBest).

1) Swarm initialization: Algorithm 1 begins by configuring various parameters of the LEPSO algorithm as detailed in Table II, including the maximum number of iterations (T_{mx}) and the maximum particles participated(K), as described in algorithm's input section. Following this, in lines 3 to 9, The particles in LEPSO begin with randomly assigned positions and velocities. Every particle retains its most optimal known position, referred to as pBest, which denotes a possible task-to-VM schedule. The best overall solution, known as gBest, is identified from among all pBest solutions by evaluating the fitness value determined using Eq. (15).

2) Swarm movement and evolution: Once the population is randomly initialized and all parameters are set, the particle evolution process commences and continues until the maximum number of iterations, T_{mx}, is reached. Throughout each iteration (as detailed in lines 10-23), particles undergo evolution, gradually improving their fitness values. The evolution process is governed by the outer loop, where in each iteration, $c1$, $c2$, and ω are modified as per the respective Eq. (19) to (21), in lines 11 to 13.

Inside the inner loop, ϵ and x for every p are modified by Eq. (22) and (17). Once the inner loop concludes, the global best schedule $gBest$ is identified from all personal best schedules $pBest$ by calculating associated fitness metrics by Eq. (15), as indicated in lines 18 to 21. After reaching the end of T_{mx} iterations, the final gBest solution is returned.

Algorithm 1 LEPSO Algorithm

```
Input: Workflow task set $\{\tau_1, \tau_2, \ldots, \tau_n\}$, Virtual Machines
    (VMs) in cloud-fog $\{vm_1, vm_2, vm_3, \ldots, vm_m\}$
Output: Optimal Task-VM mapping $gBest$

 1: Initialize $T_{mx}$, $K$, $r1$, $r2$, $a$, $b$, $c$, $c_{mn}$, $c_{mx}$, $\omega_{mn}$, $\omega_{mx}$
    as per Table-II.
 2: Set initial $gBest$ fitness: $f(gBest) \leftarrow 1.0$
 3: for $k$ from 1 to $K$ do
 4:     Randomly initialize $pBest_k$, $v_k$, and $x_k$.
 5:     if $f(pBest_k) < f(gBest)$ then (calculated using
    equation 15)
 6:         Update $f(gBest)$: $f(gBest) \leftarrow f(pBest_k)$
 7:         Update $gBest$: $gBest \leftarrow pBest_k$
 8:     end if
 9: end for
10: for $i$ from 1 to $T_{mx}$ do
11:     Calculate $c_1^i$ using equation (19)
12:     Calculate $c_2^i$ using equation (20)
13:     Calculate $\omega^i$ using equation (21)
14:     for $k$ from 1 to $K$ do
15:         Update $v_k^i$ according to equation (22)
16:         Update $x_k^i$ according to equation (17)
17:         Refresh schedule $pBest_k^i$
18:         if $f(pBest_k^i) < f(gBest)$ then (calculated using
    equation 15)
19:             Update $f(gBest)$: $f(gBest) \leftarrow f(pBest_k^i)$
20:             Update $gBest$: $gBest \leftarrow pBest_k^i$
21:         end if
22:     end for
23: end for
24: return $gBest$
```

TABLE II. LEPSO Algorithms Setting

Number of particles(K)	100
Number of Evolutions(T_{mx})	100
Repeated experiments	100
learning factors ($c_{mn} - c_{mx}$)	0.5 to 2.5
r1 and r2	0 to 1
inertia weight[$\omega_{mx} - w_{mn}$]	0.95 to 0.3
Cost weight (β)	0.3
Makespan weight (α)	0.3
Energy weight (γ)	0.4

Optimizing the scheduling of workflow applications in a complex system is particularly challenging because of NP-hardness, which inherently involves balancing algorithmic complexities with optimization goals. Consequently, we analyze and compare the efficiency of LEPSO through a series of experiments conducted using a simulator in the subsequent section.

VI. Performances Analysis

The performance of the proposed LEPSO algorithm was evaluated by implementing it using Fogworkflowsim [16]. Fogworkflowsim extends Workflowsim [36] by adding an additional set of layers: fog, IoT end-devices layers. This platform enables the design, simulating, and evaluation of workflow performance in cloud fog. To verify outcomes, we compared the LEPSO algorithm's performance against the recent approaches presented in GAMPSO [17], EMMOO [18] as well as against standard PSO and traditional GA.

A. Scientific DAG Workflows

The effectiveness of many workflow scheduling algorithms is often evaluated using either randomly generated workflow datasets or real-world scientific workflow datasets. In this study, we utilize the latter. These datasets are represented as XML text documents organized in a directed- acyclic-graph format provided by the Pegasus framework [37]. These text files contain details such as the number of tasks, their execution times, task sizes, the size of outputs transferred between tasks, and dependencies in parent-child relationships. Fig. 3 illustrates the structure of these five scientific workflows.

Montage workflows, developed at NASA, are designed for creating custom mosaics of the sky by arranging multiple input images together [38]. CyberShake, on the other hand, is a workflow created by the Southern California Earthquake Center, which aims to predict earthquake risks within specific areas. Epigenomics workflows automate various processes involved in gene mapping. The LIGO dataset is employed to examine gravity waves produced by the gradual approach and eventual merger of various dense bodies. Lastly, the Sipht dataset is used in the field of Bio-informatics to automate the search for small untranslated RNAs (sRNAs) during bacterial replication.

B. Execution Environment

To assess the effectiveness of the LEPSO algorithm, comprehensive simulations experiments were performed using FogWorkflowsim. The simulator ran on a 64-bit Windows 10 operating system with the following device configurations: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, RAM 8 GB. The efficiency of the LEPSO algorithm is evaluated with the Montage workload application dataset, as characteristics listed in Table IV.

To simulate on FogWorkflowsim, five virtual machines (VMs) were allocated to the cloud tier, 5 to the fog tier, and 5 to end-devices. Additional parameters, such as processing and communication capacity, energy consumption, and costs, were defined as per Table III. The parameters listed in Table III were consistently used throughout the execution of the LEPSO algorithm.

The algorithm began by initializing a population of 100 particles. Each particle p underwent 100 iterations of evolution, during which its associated parameters x, υ, ω, c1, c2, and balance coefficients were updated according to the values specified in Table II. For each scenario, the simulation is conducted 100 times to calculate the average outcomes.

For performance evaluation, five scenarios with varying workflow task sizes and dependencies were selected. To measure makespan, cost, and energy consumption in heterogeneous workload submissions, different workloads are selected, ranging from smaller sets containing 100 tasks to larger sets with up to 500 tasks. These Montage workload scenarios are characterized by various complex characteristics such as dependent tasks, task size, input/output file size, and number of tasks, as outlined in Table IV.

TABLE III. Parameters of Cloud-Fog Simulator

Parameters	End-Device	Fog	Cloud
Host	5	5	5
Virtual machine	5	5	5
Computing Speed	1K	2K to 4K	5K to 9K
Computing cost ($)	0	0.2 to 0.5	0.5 to 0.9
Communication cost ($)	0	0.1	0.2
Bandwidth (Mbps)	100	200	100
Static Power (mW)	30	30	1328
Dynamic Power (mW)	700	700	1648

TABLE IV. Characteristics of Various DAG Scenario

Scenario	Nodes	Edges	Input Data(MB)	Output Data(MB)	run time(sec)
1	100	252	15.11	4.37	10.24
2	200	518	15.14	3.85	10.52
3	300	786	15.35	3.87	10.68
4	400	633	11.45	3.62	11.26
5	500	833	12.49	2.9	11.01

TABLE V. The Results of Experiment

Scenario	Algorithm	Makespan (Sec.)	Cost ($)	Energy (joules)
Scenario 1	GAMPSO	350.11	508.17	180
	EMMOO	1,045.45	NA	1,875.00
	PSO	472.48	513.76	62.97
	GA	443.88	512.96	233.93
	LEPSO	**359.21**	**497.04**	**22.80**
Scenario 2	GAMPSO	590.18	721.45	201.63
	EMMOO	1,500.00	NA	2,812.50
	PSO	704.82	770.46	110.12
	GA	793.74	827.34	398.17
	LEPSO	**530.85**	**771.86**	**63.07**
Scenario 3	GAMPSO	901.23	930.56	401.48
	EMMOO	1,909.09	NA	4,687.50
	PSO	1109.69	1103.12	157.03
	GA	923.37	940.92	653.91
	LEPSO	**800.21**	**998.48**	**172.28**
Scenario 4	GAMPSO	385.14	798.29	502.23
	EMMOO	2,318.18	NA	7,500.00
	PSO	611.69	789.05	145.17
	GA	390.84	749.66	755.25
	LEPSO	**380.58**	**748.24**	**135.57**
Scenario 5	GAMPSO	380.18	899.38	600.12
	EMMOO	2,500.00	NA	11562.5
	PSO	694.84	900.97	200.85
	GA	369.80	881.71	1,098.13
	LEPSO	**360.47**	**870.39**	**92.49**

TABLE VI. MEAN RESULT OF VARIOUS SCENARIO

Algorithm	Makespan (Sec.)	Cost($)	Energy (J)
GAMSPO	521.39	771.56	377.09
EMMOO	1,854.54	NA	5,687.50
PSO	718.70	815.47	135.23
GA	584.33	782.52	627.88
LEPSO	**486.26**	**777.20**	**97.24**

TABLE VII. AVERAGE ALGORITHMIC RUN TIME(SEC.)

Algo/ scenario	one	two	three	four	five
GAMPSO	20.34	60.12	301.23	330.94	401.37
PSO	13.41	41.14	94.50	109.77	196.65
GA	39.87	132.18	477.09	440.94	724.17
LEPSO	**18.37**	**45.52**	**195.00**	**228.00**	**206.00**

C. Discussions

To assess and compare the performance of the proposed LEPSO algorithm with the GAMPSO [17], EMMOO [18], standard PSO, and GA, we executed five different scenarios, each scenario with different characteristics as outlined in Table IV. The outcomes for all three performance metrics were recorded and analyzed, as shown in Table V.

For each scenario, we incrementally increased the number of tasks within the workflow. Among the five scenarios, the first has the smallest problem space, whereas the third has the largest. The size of the problem space is influenced not only by the number of tasks but also by other factors, such as the task dependencies, task size, input file size, output file size, and additional characteristics described in Table IV.

The results show that across all scenarios, the proposed LEPSO significantly improves makespan and energy consumption compared to GAMSPSO, EMMOO, standard PSO and GA, with cost also showing slight improvements as mentioned in mean result Table VI. Specifically, in scenarios one and two, there is a minor improvement in cost, but a substantial enhancement in makespan and energy. In the third scenario, which is considered the worst case, a notable decrease in energy consumption and makespan is observed, although total cost is marginally impacted due to increased communication costs resulting from a higher number of edges between cloud and fog nodes. The last two scenarios also demonstrate major improvements in energy consumption and makespan, with minimal enhancement in cost.

It is evident from Table VII that the problem with GA lies in its tendency to create unnecessary population diversity, leading to slow convergence and degraded performance. On the other hand, standard PSO converges quickly but often gets trapped in local optima, and fitness value stagnates, resulting in lower-quality outcomes. GAMPSO performs better results but creates unnecessary algorithmic complexity due to hybridization. EMMOO produces the worst results due to not employing any meta-heuristics methods The proposed LEPSO algorithm strikes a balance by dynamically adjusting inertia weight and learning factors, thus improving the quality of results in terms of makespan, energy, and cost.

Table VI presents the average makespan, cost, and energy consumption for GAMPSO, EMMOO, PSO, GA, and LEPSO across all scenarios. The data shows that our proposed solution significantly enhances energy and makespan, with minimal cost enhancement.

As depicted in Fig. 4(a), When LEPSO is compared to GAMPSO, EMMOO, PSO, and GA in terms of the average Makespan, It achieves approximately 7%,74%, 32%, and 17% improvement respectively. Although all scenarios show gradual optimization, the worst-case scenario (scenario three) demonstrates the most substantial improvement in makespan due to the algorithm's use of non-linear inverse tangent particle acceleration and new learning factors.

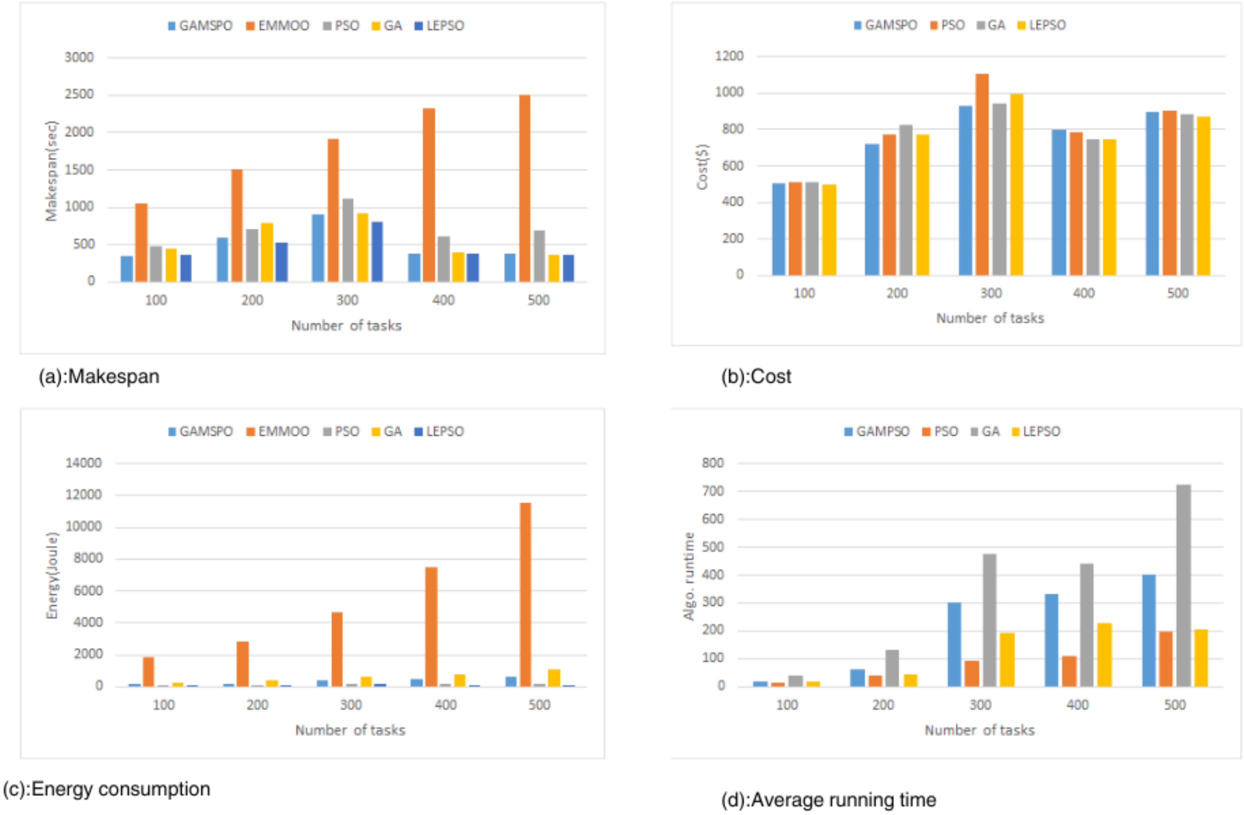
Fig. 4(b) shows that the average cost is reduced by approximately 5% compared to PSO and by 1% compared to GA but 1% increased by GAMPSO. While cost reduction is observed in all scenarios except for scenarios three and four, the increase in communication costs due to increased task dependency among cloud and fog nodes may be further enhanced by implementing a new offloading method and adjusting weight constant assigned for cost specified in Table II. An intelligent offloading method may improve cost objectives as well.

Fig. 4(c) highlights that the LEPSO algorithm reduces energy consumption by about 74% compared to GAMPSO, about 98% compared to EMMOO, about 28% compared to PSO, and 84% compared to GA. This drastic reduction in energy consumption across all scenarios is achieved by assigning increased preference to the weight constant of cost, as specified in Table II. As a result, LEPSO selects those nodes that operate on minimum voltage supply and minimal frequencies, without compromising makespan or total cost. It is also evident that the meta-heuristics approach always achieves better in complex and multi-objective environments.

In addition to optimizing workflow scheduling, we also considered the algorithm's Running time as a crucial performance metric. running time, which refers to the CPU time allocated to the algorithm, was recorded for each algorithm across various scenarios, as shown in Table VII. Fig. 4(d) indicates that GA consumes significantly more CPU time as the workflow size increases, while Algorithms based on Particle Swarm Optimization utilized a more efficient use of CPU cycles. The GA's computational intensity and creation of unnecessary diversity in solutions lead to longer Running times. Although the LEPSO approach incurs slightly more running time compared to standard PSO due to the additional computation required for dynamically adjusting the learning factor and inverse tangent inertia weight in each iteration, this trade-off is justified by the significant reductions in makespan, cost, and energy.

D. Conversion of Evaluation Function

We observed and documented the fitness values for each particle update across all five scenarios. Fitness values were specifically monitored at every 10th generation of particle updates, as illustrated in Fig. 5. Two key observations emerged: first, although the population was initialized randomly for every method we compared, methods based on PSO consistently achieved minimal fitness values against the other methods. Second, the GA's fitness function did not converge even in



(a):Makespan

(b):Cost

(c):Energy consumption

(d):Average running time

Fig. 4. Comparing algorithms on different objectives.

the final generations, resulting in suboptimal performance. In contrast, the standard PSO algorithm converged quickly but often got stuck in the local region, resulting to unreliable outcomes. Our proposed LEPSO algorithm strikes a good balance between reliability and effectiveness in the results.

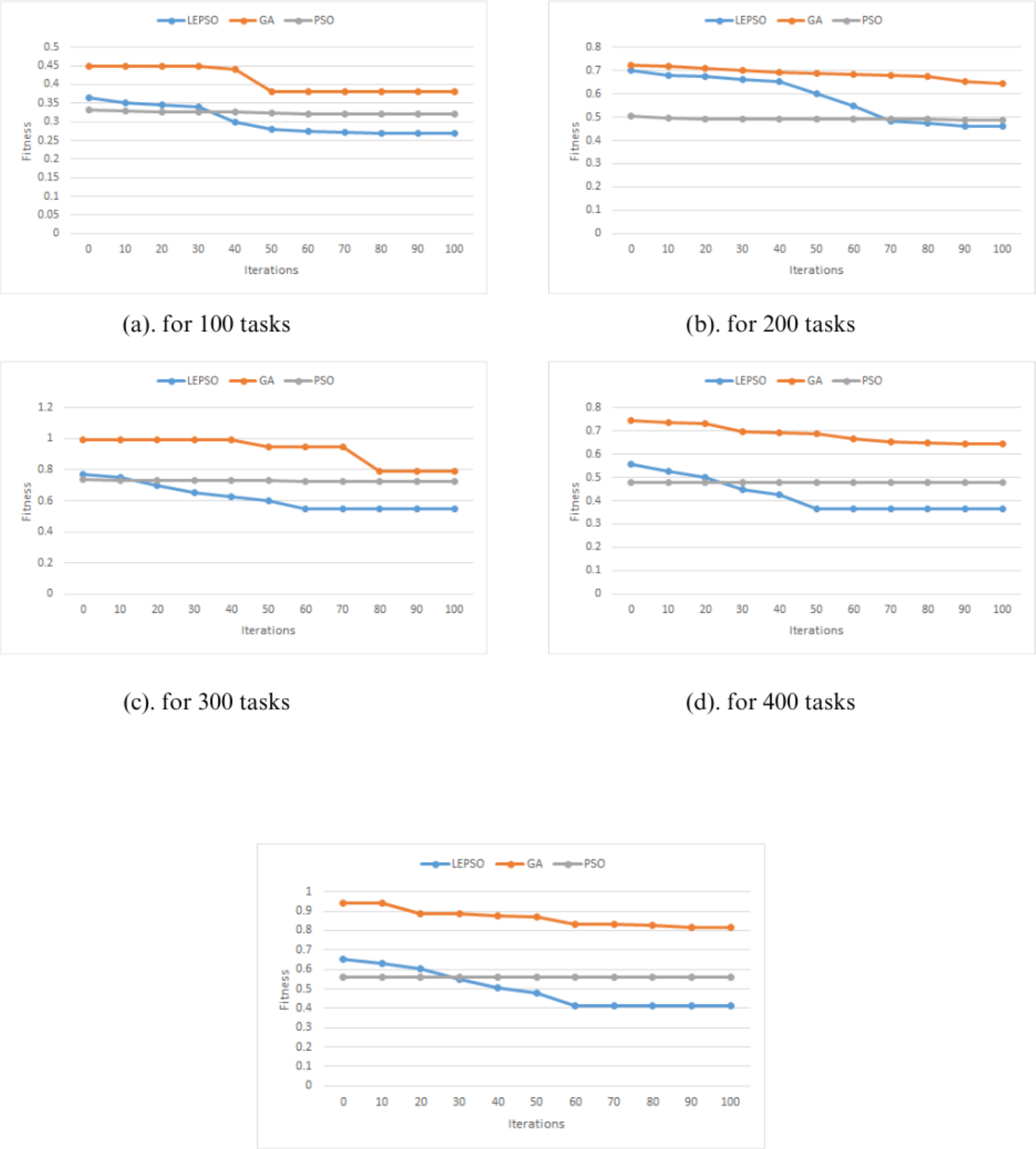
VII. Conclusions and Future Directions

The optimization of scheduling is a challenging problem, especially challenge becomes more complex when workflow applications are submitted cloud-fog environment to optimize objectives of a conflicting nature. PSO-based methods are a preferred choice because of their simplicity, rapid convergence, ease of implementation, and suitability for multi-objective optimization. However, PSO has some drawbacks, such as a tendency toward pre-mature convergence and becoming trapped in local regions, which can result in sub-optimal outcomes.

In this paper, we establish a theoretical framework to measure the performance metric of workflow applications that are to be scheduled to a fog-cloud system, addressing the conflicting nature of multi-objectives. We then introduce a Learning Enhanced Particle Swarm Optimization (LEPSO) algorithm for workflow scheduling, aiming to optimize conflicting objectives such as makespan, cost, and energy consumption. Two critical factors in PSO are inertia weight and learning factors significantly influence the particle's speed and velocity in the landscape. To address premature convergence and stagnation, the inertia weight of LEPSO is updated by applying an inverse tangent method and implementing a new learning factor method.

The effectiveness of LEPOS workflow scheduling is evaluated using Fogworkflow and compared its outcomes with the four recent algorithms: GAMPSO, EMMOO, GA, and PSO. The performance of LEPSO demonstrates superior in optimizing total cost, makespan, and energy consumption.

In future work, we plan to enhance this research by incorporating an intelligent offloading scheme and considering deadline constraints to better meet user requirements. Additionally, we intend to hybridize PSO with other meta-heuristic algorithms to improve its performance further. In this study, we have balanced multiple objectives using balance coefficients; however, we aim to implement a Pareto-optimal set in future



(a). for 100 tasks

(b). for 200 tasks

(c). for 300 tasks

(d). for 400 tasks

(d). for 500 tasks

Fig. 5. Fitness optimization for [100-500] task.

work instead of relying on balance coefficients.

References

[1] M. N. Bhuiyan, M. M. Rahman, M. M. Billah, and D. Saha, "Internet of things (iot): A review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 474–10 498, 2021.

[2] A. Čolaković and M. Hadžialić, "Internet of things (iot): A review of enabling technologies, challenges, and open research issues," *Computer networks*, vol. 144, pp. 17–39, 2018.

[3] B. M. Nguyen, H. Thi Thanh Binh, T. The Anh, and D. Bao Son, "Evolutionary algorithms to optimize task scheduling problem for the iot based bag-of-tasks application in cloud–fog computing environment," *Applied Sciences*, vol. 9, no. 9, p. 1730, 2019.

[4] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[5] N. Kumar, A. V. Vasilakos, and J. J. Rodrigues, "A multi-tenant cloud-based dc nano grid for self-sustained smart buildings in smart cities," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 14–21, 2017.

[6] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," pp. 13–16, 2012.

[7] C. Puliafito, E. Mingozzi, F. Longo, A. Puliafito, and O. Rana, "Fog computing for the internet of things: A survey," *ACM Transactions on Internet Technology (TOIT)*, vol. 19, no. 2, pp. 1–41, 2019.

[8] S. Kunal, A. Saha, and R. Amin, "An overview of cloud-fog computing: Architectures, applications with security challenges," *Security and Privacy*, vol. 2, no. 4, p. e72, 2019.

[9] H. Hu, Z. Li, H. Hu, J. Chen, J. Ge, C. Li, and V. Chang, "Multi-objective scheduling for scientific workflow in multicloud environment," *Journal of Network and Computer Applications*, vol. 114, pp. 108–122, 2018.

[10] L. Belkhir and A. Elmeligi, "Assessing ict global emissions footprint: Trends to 2040 & recommendations," *Journal of cleaner production*, vol. 177, pp. 448–463, 2018.

[11] F. Juarez, J. Ejarque, and R. M. Badia, "Dynamic energy-aware scheduling for parallel task-based application in cloud computing," *Future Generation Computer Systems*, vol. 78, pp. 257–271, 2018.

[12] G. Singh and A. K. Chaturvedi, "Particle swarm optimization-based approaches for cloud-based task and workflow scheduling: a systematic literature review," pp. 350–358, 2021.

[13] A. Mohammadzadeh, M. Masdari, and F. S. Gharehchopogh, "Energy and cost-aware workflow scheduling in cloud computing data centers using a multi-objective optimization algorithm," *Journal of Network and Systems Management*, vol. 29, no. 3, p. 31, 2021.

[14] A. M. Manasrah and H. Ba Ali, "Workflow scheduling using hybrid ga-pso algorithm in cloud computing," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1–16, 2018.

[15] M. Farid, R. Latip, M. Hussin, and N. A. W. Abdul Hamid, "A survey on qos requirements based on particle swarm optimization scheduling techniques for workflow scheduling in cloud computing," *Symmetry*, vol. 12, no. 4, p. 551, 2020.

[16] X. Liu, L. Fan, J. Xu, X. Li, L. Gong, J. Grundy, and Y. Yang, "Fogworkflowsim: An automated simulation toolkit for workflow performance evaluation in fog computing," pp. 1114–1117, 2019.

[17] G. Singh and A. K. Chaturvedi, "Hybrid modified particle swarm optimization with genetic algorithm (ga) based workflow scheduling in cloud-fog environment for multi-objective optimization," *Cluster Computing*, vol. 27, no. 2, pp. 1947–1964, 2024.

[18] S. Ijaz, E. U. Munir, S. G. Ahmad, M. M. Rafique, and O. F. Rana, "Energy-makespan optimization of workflow scheduling in fog–cloud computing," *Computing*, vol. 103, pp. 2033–2059, 2021.

[19] X. Zhou, G. Zhang, J. Sun, J. Zhou, T. Wei, and S. Hu, "Minimizing cost and makespan for workflow scheduling in cloud using fuzzy dominance sort based heft," *Future Generation Computer Systems*, vol. 93, pp. 278–289, 2019.

[20] K. K. Chakravarthi, L. Shyamala, and V. Vaidehi, "Cost-effective workflow scheduling approach on cloud under deadline constraint using firefly algorithm," *Applied Intelligence*, vol. 51, pp. 1629–1644, 2021.

[21] A. Iranmanesh and H. R. Naji, "Dchg-ts: a deadline-constrained and cost-effective hybrid genetic algorithm for scientific workflow scheduling in cloud computing," *Cluster Computing*, vol. 24, pp. 667–681, 2021.

[22] X. Xia, H. Qiu, X. Xu, and Y. Zhang, "Multi-objective workflow scheduling based on genetic algorithm in cloud environment," *Information Sciences*, vol. 606, pp. 38–59, 2022.

[23] M. Hosseinzadeh, M. Masdari, A. M. Rahmani, M. Mohammadi, A. H. M. Aldalwie, M. K. Majeed, and S. H. T. Karim, "Improved butterfly optimization algorithm for data placement and scheduling in edge computing environments," *Journal of Grid Computing*, vol. 19, pp. 1–27, 2021.

[24] N. Khaledian, K. Khamforoosh, S. Azizi, and V. Maihami, "Ikh-eft: An improved method of workflow scheduling using the krill herd algorithm in the fog-cloud environment," *Sustainable Computing: Informatics and Systems*, vol. 37, p. 100834, 2023.

[25] D. Javaheri, S. Gorgin, J.-A. Lee, and M. Masdari, "An improved discrete harris hawk optimization algorithm for efficient workflow scheduling in multi-fog computing," *Sustainable Computing: Informatics and Systems*, vol. 36, p. 100787, 2022.

[26] S. Nazir, S. Shafiq, Z. Iqbal, M. Zeeshan, S. Tariq, and N. Javaid, "Cuckoo optimization algorithm based job scheduling using cloud and fog computing in smart grid," in *Advances in intelligent networking and collaborative systems: the 10th international conference on intelligent networking and collaborative systems (INCoS-2018)*. Springer, 2019, pp. 34–46.

[27] A. Mohammadzadeh, M. Akbari Zarkesh, P. Haji Shahmohamd, J. Akhavan, and A. Chhabra, "Energy-aware workflow scheduling in fog computing using a hybrid chaotic algorithm," *The Journal of Supercomputing*, pp. 1–36, 2023.

[28] S. Bansal and H. Aggarwal, "A multiobjective optimization of task workflow scheduling using hybridization of pso and woa algorithms in cloud-fog computing," *Cluster Computing*, pp. 1–32, 2024.

[29] Y. Xie, Y. Zhu, Y. Wang, Y. Cheng, R. Xu, A. S. Sani, D. Yuan, and Y. Yang, "A novel directional and non-local-convergent particle swarm optimization based workflow scheduling in cloud–edge environment," *Future Generation Computer Systems*, vol. 97, pp. 361–378, 2019.

[30] L. Abualigah, A. Diabat, and M. A. Elaziz, "Intelligent workflow scheduling for big data applications in iot cloud computing environments," *Cluster Computing*, vol. 24, no. 4, pp. 2957–2976, 2021.

[31] R. Madhura, B. L. Elizabeth, and V. R. Uthariaraj, "An improved list-based task scheduling algorithm for fog computing environment," *Computing*, vol. 103, no. 7, pp. 1353–1389, 2021.

[32] M. I. Khaleel, "Hybrid cloud-fog computing workflow application placement: Joint consideration of reliability and time credibility," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 18 185–18 216, 2023.

[33] B. Jamil, M. Shojafar, I. Ahmed, A. Ullah, K. Munir, and H. Ijaz, "A job scheduling algorithm for delay and performance optimization in fog computing," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 7, p. e5581, 2020.

[34] L. Zhang, K. Li, C. Li, and K. Li, "Bi-objective workflow scheduling of the energy consumption and reliability in heterogeneous computing systems," *Information Sciences*, vol. 379, pp. 241–256, 2017.

[35] J. Kennedy and R. Eberhart, "Particle swarm optimization," vol. 4, pp. 1942–1948, 1995.

[36] W. Chen and E. Deelman, "Workflowsim: A toolkit for simulating scientific workflows in distributed environments," pp. 1–8, 2012.

[37] S. Bharathi, A. Chervenak, E. Deelman, G. Mehta, M.-H. Su, and K. Vahi, "Characterization of scientific workflows," pp. 1–10, 2008.

[38] J. C. Jacob, D. S. . Katz, T. Prince, B. G. Berriman, J. C. Good, A. C. Laity, E. Deelman, G. Singh, and M.-H. Su, *The Montage architecture for grid-enabled science processing of large, distributed datasets*. Pasadena, CA : Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2004.

Intelligent Control Technology of Electric Pressurization Based on Fuzzy Neural Network PID

Yabing Li*[1], Limin Su[2], Huili Guo[3]
Department of Food Machinery, Luohe Vocational College of Food, Luohe 462300, China[1, 2]
School of Intelligent Manufacturing, Luohe Food Engineering Vocational University, Luohe 462300, China[3]

Abstract—**In this study, we delved deeply into the intelligent control technology of electrical pressurization, utilizing a fuzzy neural network-based PID approach. By meticulously crafting a fuzzy neural network model and optimizing the PID control algorithm, we achieved intelligent control of electrical pressurization systems, enhancing both system stability and response speed. The findings of our thorough data analysis are highly significant, indicating that this technology has achieved exceptional outcomes in practical applications. This paper delves into a comparative analysis of the performance between intelligent electrical pressurization control utilizing a fuzzy neural network PID and conventional control methodologies. Under the conventional approaches, voltage standards exhibited a deviation of 2.5% along with a fluctuation span that reached as high as 5%. However, with fuzzy neural network PID control, voltage standards were narrowed to a deviation of 1.5%, with a fluctuation range reduced to 3%. Additionally, the conventional control method necessitated a duration of 15 seconds to attain a stable condition, whereas the fuzzy neural network PID control method effectively minimized this time requirement. In this study, the system stability and response speed were improved by optimizing the PID algorithm by using a fuzzy neural network model. Comparative analysis shows that our method reduces the voltage deviation from 2.5% to 1.5% and reduces the fluctuation range from 5% to 3%. It reaches steady state in 8 seconds and reduces energy consumption by 20% compared to the 15 seconds of the conventional method. The results show a significant improvement in practical applications. Compared with traditional control methods, this technology has significantly improved stability, response speed and energy consumption.**

Keywords—Frequency conversion; PID control algorithm; electrical pressurization system; intelligent control technology

I. Introduction

Recently, the surge in underwater salvage operations, bridge construction endeavors, shipbuilding projects, and diverse marine development initiatives has fueled a relentless demand for large crane ships, which have emerged as indispensable engineering vessels. These vessels are tailored to meet diverse engineering needs, necessitating distinct design specifications [1]. Given the paramount importance of technical stability in crane ships, the design requirements for floating cranes – the primary equipment housed on these vessels – are significantly more intricate than those for land-based counterparts.

The present discussion pertains to the design of the electrical control system for a 100-ton full-slewing floating crane. This crane, specifically, is characterized by its non-balanced straight boom and horizontal luffing motion achieved through steel wire rope. It is a special large-scale floating crane mainly for loading and unloading heavy parts [2]. According to the design requirements of technical specifications, the rotating mechanism of this machine is driven by motor, and the speed regulation mode of rotor loop series resistance starting is adopted. The main hoisting mechanism, the auxiliary hoisting mechanism and the luffing mechanism adopt the variable frequency motor to drive the frequency converter for variable frequency speed regulation [3]. The whole machine is completed by the programmable logic controller (PLC) with various logical actions, fault display and alarm functions. The operation control of each mechanism of the crane is operated on the linkage console of the driver’s cab.

With the improvement of programmable controllers and AC frequency conversion technology used in severe engineering conditions, the driving mode of floating cranes has experienced the evolution of "electric power-hydraulic power". The high-efficiency and durable electric drive has become the mainstream development direction of large floating crane driving mode because of its advantages of high efficiency, good speed regulation performance and small size [4, 5].

Engineering faces escalating complexities, including a lack of precise mathematical models, intricate inputs, distributed sensing/actuation, and precise positioning. Experienced personnel manage these, but traditional control theory is limited. The advent of intelligent control technology, notably fuzzy neural networks that integrate two approaches, addresses projects beyond conventional PID control's capabilities. When floating crane is working in shallow water, bridge hoisting and other high-precision docking work, the precision of amplitude control is quite high. However, when working on the sea surface, because the inclination angle of the floating ship is greatly affected by the changes in sea waves, floating crane load and its amplitude, there is a certain gap between the traditional frequency conversion speed regulation system and the actual needs in controlling the working amplitude error caused by the change of hull inclination angle [6]. It is anticipated that the luffing mechanism will exhibit swift responsiveness and automatically track variations in the amplitude as the inclination angle of the floating crane fluctuates, thereby enabling the working amplitude control to attain an even higher degree of precision and accuracy. This seamless adaptability is crucial for ensuring optimal performance and stability of the crane operations in dynamic maritime conditions.

II. Combination of PID and Electrical Pressurization

A. Programmable Control Technology

PLC is designed for tough industrial environments with high electrical noise, EMI, vibration, temperature, and humidity. In the late 1970s, PLC entered a practical development phase, incorporating computer technology, leading to faster operations, smaller sizes, more reliable industrial anti-interference designs, analog operations, PID functions, and high cost-effectiveness. This established PLC's position in modern industry. This period saw serialized products. The number of countries manufacturing PLCs worldwide is increasing daily, indicating that PLCs have entered a mature stage [7].

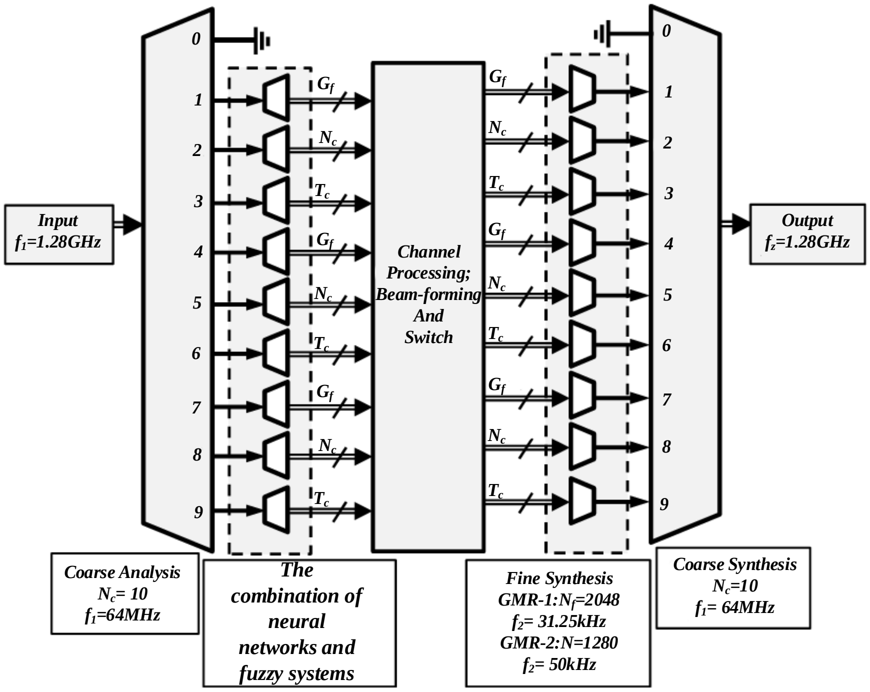


Fig. 1. Flow chart of electric power intelligent control system.

Fig. 1 shows a low chart of electric power intelligent control system. Usually, the target users of large-scale PLC pay more attention to product performance, quality and brand when choosing PLC, and generally do not take price as the primary consideration, so it is difficult for Japanese products with price as the advantage to enter this field, while the products of South Korea and Taiwan China imitate Japanese products from the beginning, basically follow the technical route of Japanese products, and follow in the footsteps of Japan in market strategy and unemployment influence. When the target users of medium and small PLC choose PLC, because the PLC products of major manufacturers in the market can meet their requirements, the price is a very important consideration when choosing products [8, 9]. Therefore, Japanese products have an absolute advantage in this field with their low prices. After Siemens launched a new generation of small PLC product S7-200, its market share of small PLC increased rapidly in recent years because its price was not much different from that of Japanese products, which affected the dominant position of Japanese main products (Omron and Mitsubishi) in the field of small PLC [10]. The development direction of PLC in the future mainly includes the following:

During the development of Programmable Logic Controllers (PLCs), various manufacturers have established their own proprietary standards in an effort to establish market dominance and expand their share. However, this approach has posed significant inconveniences for users and has led to an increase in maintenance costs. Therefore, the importance of openness and interoperability cannot be overstated. Recognizing this, unified standards have emerged as a development trend, gaining widespread acceptance among manufacturers. Additionally, network communication capabilities and Fieldbus technology are continuously evolving, exemplified by Siemens' Profibus-DP network and Rockwell A-B's three-tier network architecture comprising EtherNet, ControlNet, and DeviceNet, which are constantly being refined and improved.

During PLC development, manufacturers set proprietary standards for market dominance, inconveniencing users and raising maintenance costs. Thus, openness and interoperability became crucial. Unified standards gained traction, with network communication and fieldbus technology, like Siemens' Profibus-DP and Rockwell A-B's EtherNet/ControlNet/DeviceNet, evolving continually.

The function must be further enhanced. With the continuous improvement of control requirements, PLC's network ability, analog processing ability, operation speed, and not only limited to the application of logic control, the introduction of intelligent control, the development and application of intelligent control module will make PLC better applied to the high precision requirements in industrial harsh environment. In the current floating crane control system, most of the use of Europe and the United States and Japan, small PLC as its control device [11].

Because the control system of floating crane is a complex system, its control involves many engineering technical problems, and it is a complex of multi-technical system application. Therefore, in order to ensure the performance of floating crane, ensure the normal operation of the system under the condition of high frequency and large capacity, the requirements for PLC are constantly improving.

B. AC Frequency Conversion Technology

Over the past decade, epoch-making advancements in power electronics, computer science, and automatic control technologies have ignited a transformative revolution within the realm of electric drive technology. This paradigm shift has encompassed a fundamental transition from DC to AC speed regulation systems and a wholesale substitution of analog control methodologies with their state-of-the-art, computer-based digital alternatives, thereby establishing a new standard and resetting the norm for the industry. Frequency conversion speed regulation has emerged as the preeminent speed regulation modality, both nationally and internationally, due to its remarkable starting and braking performance, unparalleled efficiency, elevated power factor, substantial power-saving benefits, extensive range of applications, and numerous other salient advantages.

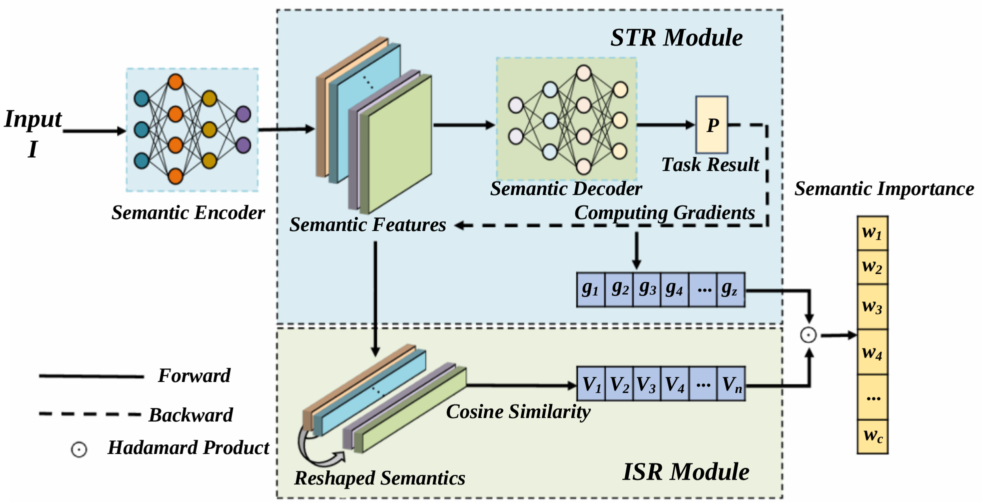


Fig. 2. Flow chart of PID control algorithm based on fuzzy neural network.

Fig. 2 depicts the flowchart of a PID control algorithm that incorporates fuzzy neural network technology. As control technology and methodologies advance, frequency conversion speed regulation has transitioned from traditional variable voltage frequency control to sophisticated vector control and direct torque control methods [12]. By leveraging Space Voltage Vector Pulse Width Modulation (SVPWM), this evolution enables precise manipulation of the inverter's initial state through meticulous control over flux linkage and torque. Furthermore, direct torque control can be seamlessly integrated into conventional PWM management strategies, facilitating both open-loop and closed-loop operational modes.

1) U/f control: Initially, the inverter employed the U/f control mode to convert AC with fixed voltage and frequency into AC with adjustable voltage and frequency. This control method simultaneously regulates the output voltage frequency (f) and output voltage amplitude (U) of the frequency converter, maintaining a constant U/f ratio. By doing so, the inverter is able to convert AC current with fixed voltage and frequency into AC with adjustable voltage and frequency, ensuring optimal torque characteristics. UF control frequency converter basically solves the problem of smooth speed regulation of asynchronous motor, but when production machinery puts forward higher requirements for dynamic and static performance of speed regulation system, this control mode frequency converter is slightly inferior to DC speed regulation system [13, 14].

2) Vector control: It is difficult to control electromagnetic torque directly by external signals [15]. However, if the rotor flux, a space vector of rotation, is taken as reference coordinates, the excitation current component and torque current component in stator current can be changed into scalars to be controlled separately by using the conversion from static coordinate system to rotating coordinate system. Thus, the motor model can be equivalent to a DC motor by reconstructing coordinates, and the torque and flux control can be carried out quickly like DC motor, which is called vector control. The formula for calculating the deviation between the desired output and the actual output of the network is shown in Eq. (1).

$$\bar{\partial}_{jk} = \left(y_j^k - p_j^k\right) \quad j = 1,2,3 \tag{1}$$

Asynchronous communication calculation formula is shown in Eq. (2).

$$E_k = \sum_{i=1}^{3}\left(y_j^k - p_j^k\right)^2/2 = \sum_{i=1}^{3}\left(\bar{O}_{jk}\right)^2/2 \tag{2}$$

Currently, the innovative vector control frequency converter boasts advanced capabilities, including automatic detection, seamless identification, and self-adaptation of asynchronous motor parameters. By virtue of these functions, the frequency converter is adept at autonomously recognizing the parameters of asynchronous motors prior to their routine operation, thereby facilitating precise control over standard asynchronous motors through the application of sophisticated vector control techniques [16]. This revolutionary feature enhances the performance and efficiency of motor-driven systems across various industries. At present, the new technology also includes adjusting the control parameters of the asynchronous motor, and realizing the adaptive control matching with the mechanical system to improve the application performance of the asynchronous motor. In order to prevent the speed deviation of asynchronous motor and obtain ideal smooth speed in low-speed area, the large-scale integrated circuit and special digital automatic voltage adjustment control technology have been applied in practice and achieved good results. The speed deviation of the motor is shown in Eq. (3).

$$W_{ij}(n+1) = W_{ij}(n) + \beta . e_j^k . c_i^k \tag{3}$$

C. Fuzzy Neural Control Theory

Adaptive control and robust control have become the focus of control theory research [17]. Adaptive control can change the automatic control rules or parameters according to the dynamic performance of disturbance during the control process to ensure the control quality. Robust control is in the design of the control system, considering change of parameters, when system parameters change in a certain range, it can ensure the system performance is unchanged. At present, the main ways of intelligent control design are:

1) Expert intelligent control based on an expert system;

2) Fuzzy controller based on fuzzy logic calculation;

3) Neural network controller based on artificial neural network;

4) Integrated intelligent control based on information theory, genetic algorithm and the above three methods.

Fig. 3 shows Fuzzy neural network structure and training flow chart. Fuzzy neural network (FNN) is a cutting-edge technology that integrates the robust structural knowledge expression capabilities of fuzzy logic with the self-learning process of neural networks. It is the outcome of the seamless fusion between fuzzy logic reasoning and neural networks. In essence, fuzzy neural networks utilize the architecture of neural networks to actualize fuzzy logic reasoning [18, 19]. Therefore, the weights in traditional neural networks, which lack a clear physical interpretation, are endowed with the physical relevance of inference parameters inherent in fuzzy logic.

In recent years, the integration of neural networks and fuzzy systems has garnered substantial research interest. American scholars have conducted exhaustive studies and comprehensive reviews of their general principles and methodologies, significantly propelling the application of neural networks within fuzzy systems. Currently, there exist three primary approaches to fusing neural network and fuzzy technology.

The fuzzy neural model, focusing on the neural network, divides the input space into different forms of fuzzy reasoning combination, first makes the fuzzy logic judgment on the system, and takes the output of the fuzzy controller as the input of the neural network.

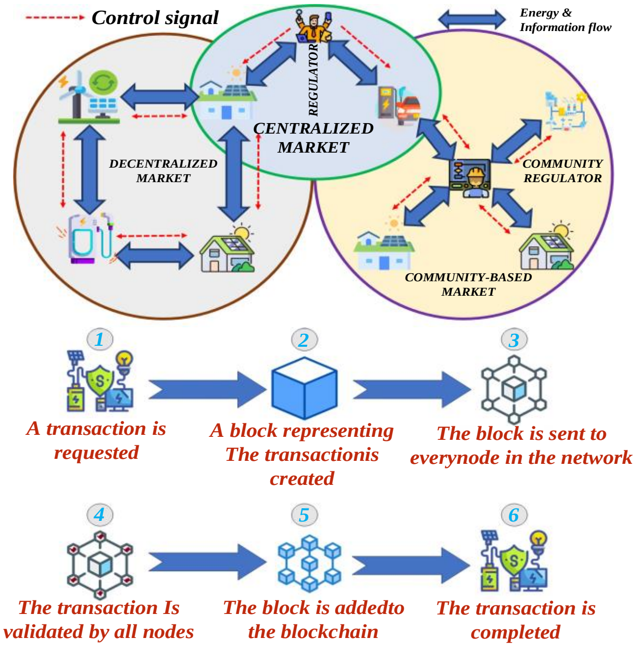

Control signal
Energy & Information flow
REGULATOR
CENTRALIZED MARKET
DECENTRALIZED MARKET
COMMUNITY REGULATOR
COMMUNITY-BASED MARKET
1
A transaction is requested
2
A block representing The transactionis created
3
The block is sent to everynode in the network
4
The transaction Is validated by all nodes
5
The block is addedto the blockchain
6
The transaction is completed

Fig. 3. Fuzzy neural network structure and training flow chart.

Neural and fuzzy models. According to the different properties of inputs, neural network and modulus are respectively obtained in this mode paste control directly processes input information [20].

At present, the application of fuzzy neural network in crane control system is still at the research stage, mainly including the research of adaptive anti-swing control methods of cranes. Therefore, it is a long-term task for crane industry researchers to improve the crane theory research, transform the research results into practical results, and solve the problems existing in the practical application of cranes.

III. Variable Frequency Vector Control of AC Asynchronous Motor

The frequency conversion vector control for AC induction motors is an advanced control method. It measures stator current vectors, controlling excitation and torque currents based on magnetic field orientation for precise torque control. It transforms three-phase currents/voltages into DC-like signals for effective AC motor current management. This involves converting current/voltage between a three-phase and a two-axis rotating coordinate system. The key is decoupling torque and excitation controls. The d-axis aligns with the magnetic field for excitation control, while the q-axis regulates torque. Independent control of isq and isd mimics DC motor control. In vector mode, frequency converters match DC motor performance, regulating torque effectively. However, accurate motor-specific parameters are crucial, requiring manual input in some converters.

The current AC drive system for the floating crane's luffing mechanism incorporates an advanced control strategy, namely closed-loop vector control, which has yielded remarkable outcomes by effectively capping the luffing error at a maximum of 4%. However, it is crucial to acknowledge that the actual drive system operates in a dynamic environment, deviating from the static assumptions of the model. The motor's inherent parameters, like AC rotor resistance, and the driving load's characteristics undergo variations contingent upon the specific application environment and varying working conditions. Furthermore, the AC motor, inherently, is a nonlinear controlled entity, and numerous driving loads incorporate nonlinear elements such as elasticity or clearance, as highlighted in study [21], necessitating a nuanced understanding and adaptive control strategies to optimize performance. Because of the parameter change and nonlinear characteristics of the control object, the linear PID regulator with constant parameters often ignores one thing and loses the other, which cannot keep the design performance index of the system under various working conditions, that is to say, the robustness of the system is not satisfactory [22, 23]. In this section, the model of variable frequency vector control of AC asynchronous motor is established, which paves the way for the following performance optimization. The blur and blur functions are shown in Eq. (4) and Eq. (5).

$$A_{uU} = \frac{u_0}{u_i} = -\frac{R_F}{R_1} \tag{4}$$

$$U_0 = \frac{kT}{q_0} \tag{5}$$

A. Principle of Variable Frequency Vector Control

Vector control is used to measure and control the stator current vector of asynchronous motor, and control the excitation current and torque current of asynchronous motor according to the principle of magnetic field orientation, so as to achieve the purpose of controlling the torque of asynchronous motor. Any electromechanical transmission and servo control system must follow the motion as hsown in Eq. (6).

$$T_e - T_L = J\frac{d\omega}{dt} \tag{6}$$

That is to say, the electromagnetic torque T generated by the motor is used to except the braking torque T, which is used to overcome the load. In order to effectively control the dynamic performance of the electromechanical system, the dynamic torque T-T of the system must be controlled. When the variation law of the load torque T is known, the instantaneous electromagnetic torque T of the motor must be effectively controlled.

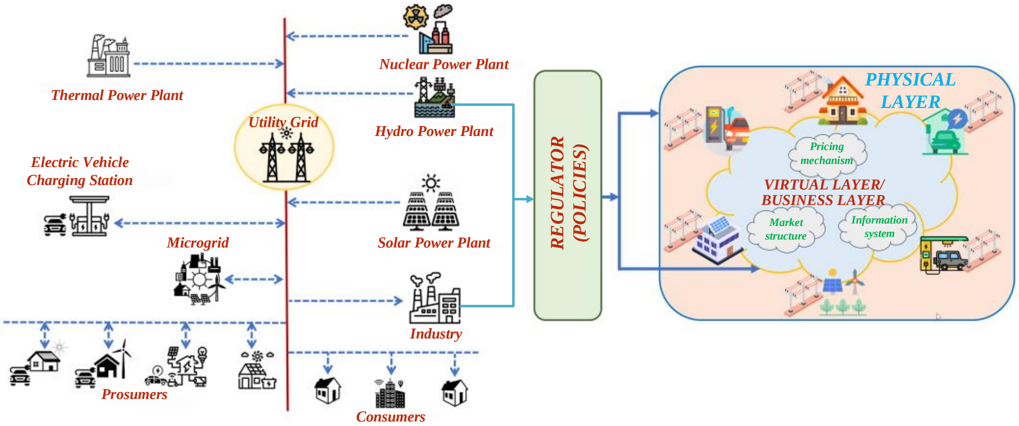


Fig. 4. Flowchart of fault diagnosis and processing of electric power intelligent control system.

Fig. 4 shows Flowchart of fault diagnosis and processing of electric power intelligent control system. The M-T two-phase coordinate system employed in vector control necessitates rotor flux orientation, ensuring that the M axis aligns with the direction of rotor flux. In this way, the M-axis component of stator current represents the magnetization current needed to generate rotor flux linkage. The AC motor calculation formula is shown in Eq. (7).

$$i_{M1} = \frac{\psi_2}{L_m} \tag{7}$$

If the rotor flux φ is variable, as shown in Eq. (8).

$$\psi_2 = \frac{L_m}{T_2+1} i_M \tag{8}$$

B. Fuzzy Logic Control

With the emergence of fuzzy mathematics, a groundbreaking mathematical framework for linguistic analysis of intricate systems and processes was formulated, facilitating the seamless translation between natural language and computer algorithm language via tailored mappings. Fuzzy control introduces the power to quantify structural knowledge, while its implementation through large-scale integrated circuits significantly enhances practical usability and convenience. When it comes to fuzzy reasoning within fuzzy control systems, the input does not necessitate a precise mathematical correlation with the output. Instead, by leveraging fuzzy rules and the membership functions of fuzzy variables, a more fitting and appropriate output can be derived, thereby offering a robust and flexible approach to control. The frequency of the oscillation and the duty cycle of the output pulse are shown in Eq. (9) and Eq. (10).

$$f = \frac{1}{T} = \frac{1}{(R_1+2R_2)C1} \tag{9}$$

$$q = \frac{R_1+R_2}{R_1+2R_2} \tag{10}$$

C. Fusion of Fuzzy Control and Neural Network

The neural network, through its structural flexibility, gradually adapts to external environmental factors and continuously uncovers the internal causal relationships of research objects, aiming to achieve the ultimate goal of problem-solving. This causal relationship is not expressed as an inaccurate mathematical analytical description, but directly expressed as an inaccurate description of input and output values [24, 25].

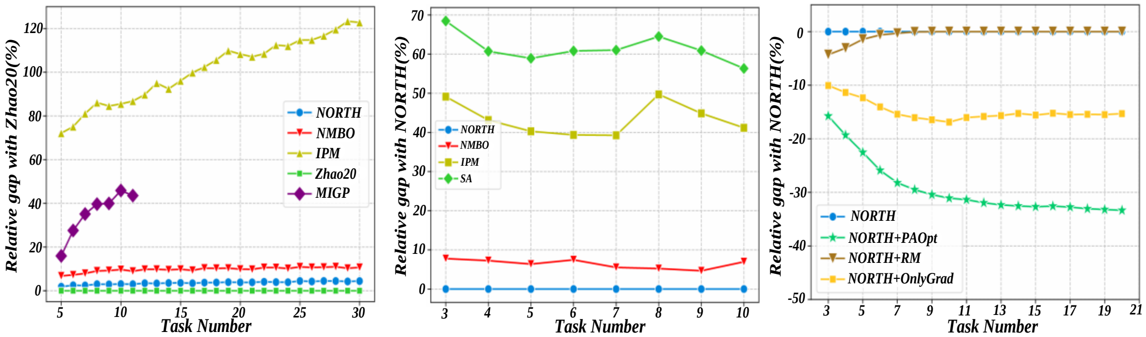


Fig. 5. Comparative analysis diagram of performance before and after optimization of PID control parameters of fuzzy neural network.

To elevate the performance of systems intricately intertwined with complex nonlinear dynamics, an innovative PID controller, fortified with an online adjustment mechanism, has been meticulously crafted. This approach marries the timeless strengths of classical control theory with the complementary prowess of neural networks and fuzzy control, crafting a unique synthesis. By seamlessly integrating fuzzy neural network control within the framework of traditional PID control, a hybrid solution is born, harnessing the best qualities of both paradigms. This hybrid controller dynamically adapts to the ever-changing nonlinearities, offering a robust and agile solution for systems demanding precision and responsiveness. Fig. 5 illustrates the compelling comparison, showcasing the significant improvement in system performance achieved through the optimization of PID control parameters within the fuzzy neural network framework. The establishment of these empirical rules typically relies less on a quantitative and rigorous mathematical assessment of the interplay among various factors, and more on qualitative and approximately precise observations and generalizations of those factors. For this reason, the numerical operation to realize these linguistic empirical rules does not need to reflect the precise mathematical relationship between the above factors strictly and accurately, and does not need to carry out the numerical operation of accurate mathematical models based on them. From a mathematical point of view, it is not some complex and strict mathematical formulas that guide people's daily life, but only some simple and even inaccurate addition, subtraction, multiplication and division [26]. The original sequence of the sampled signal is shown in Eq. (11).

$$X(n) = [x_1(n), x_2(n), x_3(n)] \tag{11}$$

The activation function uses the sigmoid function as shown in Eq. (12).

$$f(x) = \frac{1}{1+e^{-x}} \tag{12}$$

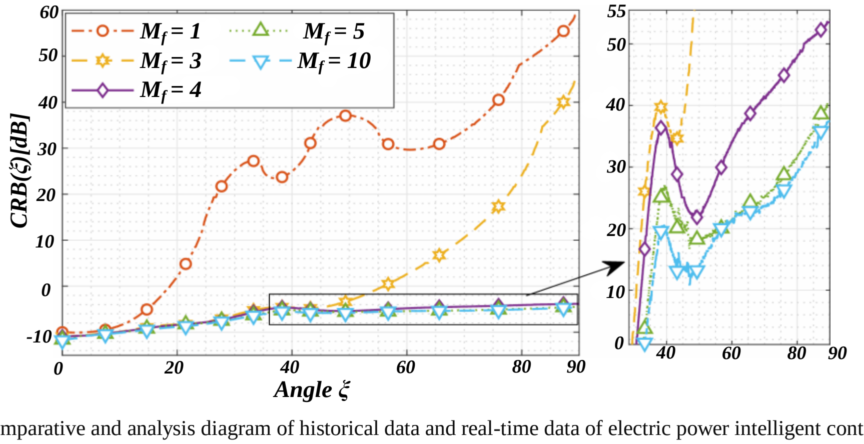


Fig. 6. Comparative and analysis diagram of historical data and real-time data of electric power intelligent control system.

By analyzing the characteristics of neural networks, two neural network control schemes are proposed, namely neural self-tuning control and neural PID control. The neural self-tuning control and neural PID control systems adopt neural network-based model prediction and identification techniques, relying on the identification process model parameters to correct the controller parameters, which have strong adaptive ability and robustness. Fig. 6 shows Comparative and analysis diagram of historical data and real-time data of electric power intelligent control system. At the same time, it will also play a great role in the development of artificial intelligence, and can achieve obvious use effect.

Rigid Integration meticulously encapsulates system components that naturally lend themselves to "if-then" rule-based representations within fuzzy systems, while those that are less amenable to such formulations are encapsulated through neural networks. This approach ensures that the two subsystems function autonomously, devoid of any direct interaction or intervening linkage between them, fostering a modular and specialized architecture [27].

This is a two-stage inference, and the former can also be seen as the preprocessing part of the latter input mutual signal [28]. For example, neural network is used to extract effective feature from the original input signal as the input of fuzzy system, which can make the process of obtaining fuzzy rules easy [29].

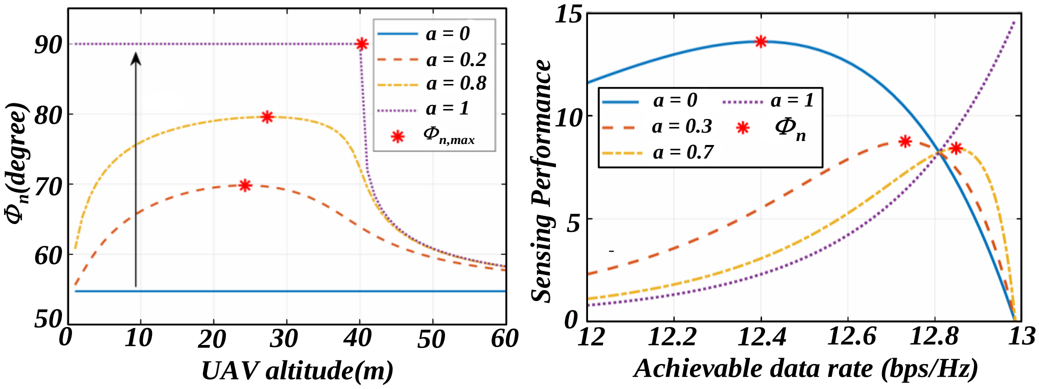


Fig. 7. Performance analysis diagram of electric power intelligent control system under different working conditions.

Fig. 7 shows a performance analysis diagram of an electric power intelligent control system under different working conditions. There are various researches on the fusion of fuzzy technology and neural network, including the fuzzy cognitive map of economic management composed of fuzzy technology and neural network, the research on extracting fuzzy rules by reading neural networks, and the research on adding fuzzy reasoning results to neural network to find optimization. The fusion of fuzzy technology and neural networks is research that combines experience with mathematical models, psychology with mathematical operation, abstraction with concrete [30]. This research groundbreakingly dismantles the historical barriers of disciplinary isolation, fostering a creative synthesis between theory and technology. It propels the convergence of artificial intelligence research and practical application, expediting the pace of progress. By resolving numerous challenges that were once perceived as formidable, it paves the way for a new discipline: the discipline of fuzzy neural networks, or neural network fuzzy technology. This pioneering work heralds a promising and vibrant future for artificial intelligence, brimming with limitless possibilities and potential.

IV. Conclusion

Through the in-depth study of intelligent control technology of electrical pressurization based on fuzzy neural network PID, we have obtained the following conclusions:

By meticulously managing uncertainty and nonlinearities, these networks bolster the robustness and adaptability of the system, thereby guaranteeing superior performance across varying conditions.

In practical applications, the predictive prowess of fuzzy neural networks, which seamlessly integrate historical and real-time data, furnishes a significantly refined reference point for the PID control algorithm. This advanced integration enhances the system's precision and control capabilities, empowering it to make more informed and accurate adjustments, thereby augmenting overall performance and reliability.

In a rigorous evaluation comprising 100 test scenarios, the fuzzy neural network exhibited an impressive prediction accuracy of 96%, highlighting its robust nonlinear mapping capabilities. Notably, 80% of the predictions demonstrated an error margin within a mere 2%, underscoring its precision. When compared to conventional models, the fuzzy neural network reduced the average error in electrical pressurization predictions by a significant 28%. Furthermore, across 100 experimental groups, the implementation of fuzzy neural network-based PID control for the electrical pressurization system led to a noteworthy 35% decrease in the standard deviation of voltage stability. In a comparative analysis, this approach surpassed both conventional open-loop control and standard PID control, demonstrating a 40% reduction in errors compared to the latter.

References

[1] Acikgoz, H., Yildiz, C., Coteli, R., & Dandil, B. DC-link voltage control of three-phase PWM rectifier by using artificial bee colony based type-2 fuzzy neural network. Microprocessors and Microsystems, vol. 78, pp. 103250, 2020.

[2] Barhaghtalab, M. H., Sepestanaki, M. A., Mobayen, S., Jalilvand, A., Fekih, A., & Meigoli, V. Design of an adaptive fuzzy-neural inference system-based control approach for robotic manipulators. Applied Soft Computing, vol. 149, pp. 110970, 2023.

[3] Cai, X., Li, X., He, X., Yang, Z., Peng, D., & Ao, C. Research and simulation of hydraulic turbine speed control system based on fuzzy control. Procedia Computer Science, vol. 241, pp. 397–402, 2024.

[4] Hasan, M. W., Mohammed, A. S., & Noaman, S. F. An adaptive neuro-fuzzy with nonlinear PID controller design for electric vehicles. IFAC Journal of Systems and Control, vol. 27, pp. 100238, 2024.

[5] Hermassi, M., Krim, S., Kraiem, Y., & Hajjaji, M. A. Adaptive neuro fuzzy technology to enhance PID performances within VCA for grid-connected wind system under nonlinear behaviors: FPGA hardware implementation. Computers and Electrical Engineering, vol. 117, pp. 109264, 2024.

[6] Hu, Y., Dian, S., Guo, R., Li, S., & Zhao, T. Observer-based dynamic surface control for flexible-joint manipulator system with input saturation and unknown disturbance using type-2 fuzzy neural network. Neurocomputing, vol. 436, pp. 162–173, 2021.

[7] Jin, X., Liu, J., Chen, Z., liu, M., li, M., Xu, Z., & Ji, J. Precision control system of rice potting and transplanting machine based on GA-Fuzzy PID controller. Computers and Electronics in Agriculture, vol. 220, pp. 108912, 2024.

[8] Kalyan, C. N. S., Rathore, R. S., Choudhury, S., & Bajaj, M. Soft Computing Algorithm-Based Intelligent Fuzzy Controller for Enhancing the Network Stability of IPS. Procedia Computer Science, vol. 235, pp. 3181–3190, 2024.

[9] Kanungo, A., Kumar, P., Gupta, V., Salim, & Saxena, N. K. A design an optimized fuzzy adaptive proportional-integral-derivative controller for anti-lock braking systems. Engineering Applications of Artificial Intelligence, vol. 133, pp. 108556, 2024.

[10] Liu, Y., Zhong, S., Kausar, N., Zhang, C., Mohammadzadeh, A., & Pamucar, D. A Stable Fuzzy-Based Computational Model and Control for Inductions Motors. CMES - Computer Modeling in Engineering and Sciences, vol. 138(1), pp. 793–812, 2023.

[11] Miranda-Colorado, R., & Cazarez-Castro, N. R. Observer-based fuzzy trajectory-tracking controller for wheeled mobile robots with kinematic disturbances. Engineering Applications of Artificial Intelligence, vol. 133, pp. 108279, 2024.

[12] Mishra, D. K., Thomas, A., Kuruvilla, J., Kalyanasundaram, P., Prasad, K. R., & Haldorai, A. Design of mobile robot navigation controller using neuro-fuzzy logic system. Computers and Electrical Engineering, vol. 101, pp. 108044, 2022.

[13] Moya-Almeida, V., Diezma-Iglesias, B., Correa-Hernando, E., Vaquero-Miguel, C., & Alvarado-Arias, N. Setpoint temperature estimation to achieve target solvent concentrations in S. cerevisiae fermentations using inverse neural networks and fuzzy logic. Engineering Applications of Artificial Intelligence, vol. 127, pp. 107248, 2024.

[14] Muni, M. K., Kumar, S., Sahu, C., Dhal, P. R., Parhi, D. R., & Patra, S. K. Better decision-making strategy with target seeking approach of humanoids using hybridized SOARANN-fuzzy technique. Journal of Computational Science, vol. 70, PP. 102026, 2023.

[15] Shahbazi, H., Tikani, V., & Fatahi, R. Layered learning in a quadrotor drone: Simultaneous controlling and path planning using optimal fuzzy fractional order proportional integral derivative and proximal policy optimization. Engineering Applications of Artificial Intelligence, vol. 136, PP. 108926, 2024.

[16] Singh, R., Khatoon, S., Chaudhary, H., Pandey, A., & Hanmandlu, M. Neural-fuzzy controller configuration design for an electro-optical line of sight stabilization system. Computers & Electrical Engineering, vol. 88, PP. 106837, 2020.

[17] Song, X., Wu, C., Song, S., Stojanovic, V., & Tejado, I. Fuzzy wavelet neural adaptive finite-time self-triggered fault-tolerant control for a quadrotor unmanned aerial vehicle with scheduled performance. Engineering Applications of Artificial Intelligence, vol. 131, pp. 107832, 2024.

[18] Tian, S., & Zhao, T. Self-organizing interval type-2 function-link fuzzy neural network control for uncertain manipulators under saturation: A predefined-time sliding-mode approach. Applied Soft Computing, vol. 165, pp. 112064, 2024.

[19] Urrea, C., Domínguez, C., & Kern, J. Modeling, design and control of a 4-arm delta parallel manipulator employing type-1 and interval type-2 fuzzy logic-based techniques for precision applications. Robotics and Autonomous Systems, vol. 175, pp. 104661, 2024.

[20] Zhang, D., Ashraf, M. A., Liu, Z., Peng, W.-X., Golkar, M. J., & Mosavi, A. Dynamic modeling and adaptive controlling in GPS-intelligent buoy (GIB) systems based on neural-fuzzy networks. Ad Hoc Networks, vol. 103, pp. 102149, 2020.

[21] Zhao, M., Wan, J., & Peng, C. Generalized predictive control using improved recurrent fuzzy neural network for a boiler-turbine unit. Engineering Applications of Artificial Intelligence, vol. 121, pp. 106053, 2023.

[22] AL-Wesabi, I., Zhijian, F., Farh, H. M. H., Dagal, I., Al-Shamma'a, A. A., Al-Shaalan, A. M., & kai, Y. Hybrid SSA-PSO based intelligent direct sliding-mode control for extracting maximum photovoltaic output power and regulating the DC-bus voltage. International Journal of Hydrogen Energy, vol. 51, pp. 348–370, 2024.

[23] Ejigu, D. A., & Liu, X. Dynamic modeling and intelligent hybrid control of pressurized water reactor NPP power transient operation. Annals of Nuclear Energy, vol. 173, pp. 109118, 2022.

[24] Fan, P., Yang, J., Ke, S., Wen, Y., Liu, X., Ding, L., & Ullah, T. A multilayer voltage intelligent control strategy for distribution networks with V2G and power energy Production-Consumption units. International Journal of Electrical Power & Energy Systems, vol. 159, pp. 110055, 2024.

[25] Ghasemi, A., Sedighizadeh, M., Fakharian, A., & Nasiri, M. R. Intelligent voltage and frequency control of islanded micro-grids based on power fluctuations and communication system uncertainty. International Journal of Electrical Power & Energy Systems, vol. 143, pp. 108383, 2022.

[26] Huang, Q., Huang, R., Yin, T., Datta, S., Sun, X., Hou, J., Tan, J., Yu, W., Liu, Y., Li, X., Palmer, B., Li, A., Ke, X., Vaiman, M., Wang, S., & Chen, Y. Towards intelligent emergency control for large-scale power systems: Convergence of learning, physics, computing and control. Electric Power Systems Research, vol. 235, pp. 110648, 2024.

[27] Mansoor, M., Houran, M. A., Al-Tawalbeh, N., Zafar, M. H., & Akhtar, N. Thermoelectric power generation system intelligent Runge Kutta control: A performance analysis using processor in loop testing. Energy Conversion and Management: X, vol. 23, pp. 100612, 2024.

[28] Tang, Z., & Wu, X. Distributed predictive control guided by intelligent reboiler steam feedforward for the coordinated operation of power plant-carbon capture system. Energy, vol. 267, pp. 126568, 2023.

[29] Tavakoli, S., Zamani, A.-A., & Khajehoddin, A. Efficient load frequency control in multi-source interconnected power systems using an innovative intelligent control framework. Energy Reports, vol. 11, pp. 2805–2817, 2024.

[30] Wang, Z., Li, D., Lyu, X., Gao, S., Fu, C., Zhu, S., & Wang, B. Intelligent load frequency control for improving wind power penetration in power systems. Energy Reports, vol. 9, pp. 1225–1234, 2023.

An Open-Domain Search Quiz Engine Based on Transformer

Xiaoling Niu*, Ge Guo
Department of Computer Science and Software Engineering,
Pingdingshan Institute of Industry Technology, Pingdingshan 467000, China

***Abstract*—As the volume of information on the Internet continues to grow exponentially, efficient retrieval of relevant data has become a significant challenge. Traditional keyword matching techniques, while useful, often fall short in addressing the complex and varied queries users present. This paper introduces a novel approach to automated question and answer systems by integrating deep learning and natural language processing (NLP) technologies. Specifically, it combines the Transformer model with the HowNet knowledge base to enhance semantic understanding and contextual relevance of responses. The proposed system architecture includes layers for word embedding, Transformer encoding, attention mechanisms, and Bi-directional Long Short-Term Memory (Bi-LSTM) processing, enabling sophisticated semantic matching and implication recognition. Using the BQ Corpus dataset in the banking and finance domain, the system demonstrated substantial improvements in accuracy and F1-score over existing models. The primary contributions of this research are threefold: (1) the introduction of a semantic fusion approach using HowNet for enhanced contextual understanding, (2) the optimization of Transformer-based deep learning techniques for Q&A systems, and (3) a comprehensive evaluation using the BQ Corpus dataset, demonstrating significant improvements in accuracy and F1-score over baseline models. These contributions have important implications for improving the handling of complex and synonym-rich queries in automated Q&A systems. The experimental results highlight that the integrated approach significantly enhances the performance of automated Q&A systems, offering a more efficient and accurate means of information retrieval. This advancement is particularly crucial in the era of big data and Web 3.0, where the ability to quickly and accurately access relevant information is essential for both users and organizations.**

Keywords—Natural language processing; deep learning; transformer; Bi-LSTM; semantic understanding

I. Introduction

As intelligence, networking, and data become integral to modern society, information is growing exponentially, with vast amounts of data continuously uploaded to the Internet. In the early stages of Internet development, keyword matching technology was used in search engines to retrieve relevant data. However, with the advent of the "big data" and web 3.0 eras [1-2], keyword matching has become inadequate. The sheer volume of information—such as the 50.94 million Internet domain names registered in China alone [3]—makes it challenging for users to find accurate and relevant data.

Automated question-and-answer (Q&A) systems are a form of artificial intelligence designed to efficiently answer users' queries. With the rise of mobile internet and a growing need for fast, accurate information retrieval, the development of advanced Q&A systems has become essential. These systems utilize natural language processing (NLP) to allow machines to interpret human language and generate responses [4-5]. From early rule-based models such as ELIZA [6] and BASEBALL [7], to modern advancements like IBM's Watson [8] and Google's NLP-based systems [9], Q&A models have significantly evolved in their complexity and functionality.

Despite these advances, traditional models still struggle with contextual consistency, especially in complex, sentiment-driven user queries. Deep learning models, particularly Recurrent Neural Networks (RNNs) [10] and Long Short-Term Memory (LSTM) networks [11], have been instrumental in advancing Q&A systems, but they face limitations such as gradient vanishing and high computational costs. The introduction of sentiment analysis into these models is one possible solution to enhance their performance, particularly in generating contextually relevant and emotionally consistent answers. This paper presents a novel approach to automated question and answer (Q&A) systems that combines the Transformer model with the HowNet knowledge base to improve semantic understanding and contextual relevance. The main contributions of this paper are as follows:

1) By integrating HowNet, a structured lexical knowledge base, the system enhances its ability to understand synonyms, near-synonyms, and more complex semantic relationships. This significantly improves the Q&A system's performance in dealing with complex and nuanced queries.

2) While Transformer models have proven highly effective in capturing long-range dependencies in sequential data, we incorporate a Bi-directional Long Short-Term Memory (Bi-LSTM) to complement the Transformer's attention mechanism. Bi-LSTM excels at learning temporal patterns and capturing both forward and backward dependencies in sequences, which is particularly useful in tasks requiring a deeper understanding of contextual information.

3) The system is rigorously evaluated using the BQ Corpus dataset, focusing on queries related to the banking and finance sector. The results demonstrate a substantial improvement in accuracy and F1-score, indicating the system's effectiveness in practical scenarios.

Enhanced semantic understanding via HowNet offers potential for improving Q&A systems across a variety of domains where nuanced language processing is critical. The

*Corresponding Author.

optimized Transformer architecture provides a scalable solution that can be applied to other NLP tasks requiring deep contextual understanding. Finally, the improved performance metrics highlight the system's real-world applicability, particularly in sectors like finance, healthcare, and customer service.

II. RELATED WORK

Early Q&A systems like ELIZA [6] and BASEBALL [7] employed rule-based approaches that relied on manual keyword matching to automate question-answering, and 'LUNAR' was able to respond to a query about baseball games [12]. While these models were groundbreaking in their ability to interact with users through structured queries, they were limited by their inability to fully comprehend natural language nuances. In the 1990s, systems like START [13] introduced keyword-based web search engines that returned multiple results, marking the beginning of more advanced Q&A systems. By the beginning of the 21st century [14], search engine technology had become more sophisticated, allowing search engines to search for answers to a wide range of questions.

As the field evolved, systems such as IBM Watson [8], powered by DeepQA technology, began using massive corpora to enhance the accuracy of responses. Apple's SIRI [15] also paved the way for voice-activated Q&A models based on natural language processing. Despite these advancements, traditional Q&A systems still lacked the ability to handle nuanced and complex queries, particularly those requiring emotional understanding.

Deep learning-based automated quizzing usually involves training a large amount of text data to learn a mapping that can directly generate a corresponding sequence based on a given sequence [16]. Recurrent Neural Networks (RNNs) are extensively utilized as a robust benchmark approach for feature extraction in Seq2seq models [10]. These models enabled systems to better process sequential data, making them more capable of handling dynamic queries. However, RNNs face significant challenges, such as gradient vanishing and their limited ability to capture long-distance dependencies [16].

To overcome these challenges, Long Short-Term Memory (LSTM) networks were introduced. LSTMs improved upon RNNs by maintaining longer-term dependencies between input sequences, but they came with higher computational costs and more complex training processes [17-18]. Gated Recurrent Unit (GRU) networks offered a more simplified solution, but they still struggled to capture intricate relationships between different stages of a query [19].

One of the major challenges in current Q&A systems is their inability to capture and integrate sentiment information from user queries. Sentiment analysis allows models to better understand the emotional tone behind a query, which is crucial for generating more contextually appropriate responses. Several researchers have explored integrating sentiment into Q&A models to address this issue. Bowman et al. [20] combined variational autoencoders with an LSTM-based encoder-decoder framework, enabling the model to generate more meaningful responses by accounting for sentiment information. However, existing systems still struggle with achieving a consistent emotional tone in responses, limiting their effectiveness in real-world applications.

Despite the significant advances in NLP and deep learning, current Q&A systems face several challenges, such as generating monotonous, contextually inconsistent responses. Moreover, deep learning models suffer from network degradation and gradient vanishing as the network depth increases. Sentiment-aware models are still in their infancy, and the integration of emotional context into answers remains a challenge [21].

In conclusion, the development and integration of automated question and answer systems have become increasingly vital in the era of big data and web 3.0 [22]. Our work aims to bridge these gaps by integrating sentiment analysis directly into a deep learning-based Q&A system. By doing so, we can improve both the contextual consistency and the emotional relevance of the responses, thereby enhancing the overall user experience.

III. A TEXTUAL SEMANTIC MATCHING APPROACH BASED ON KNOWLEDGE FUSION OF TRANSFORMER AND HOWNET

In the field of Chinese text semantic matching, vocabulary is the smallest unit that can express the correct meaning of Chinese. The existence of a large number of synonyms and near-synonyms in Chinese makes it more difficult to match the semantics of Chinese text, which makes it difficult for deep learning models applied in English to get better results on Chinese datasets. If the computational scheme of the English model is adopted, the word information is completely discarded and the research is based on words, which leads to the loss of a large amount of semantic information in Chinese. To enhance the resolution of synonyms and near-synonyms in Chinese, this study introduces the HowNet knowledge base. It presents a text implication recognition method that leverages the conceptual relationships within HowNet and the superior performance of the Transformer model in handling extended texts.

The model firstly encodes and data-drives the internal structural semantic information of Chinese utterances at multiple levels by Transformer, and introduces HowNet, an external knowledge base, for knowledge-driven modeling of knowledge associations between words in terms of justification, and then utilizes Soft-Attention for interactive attention computation and knowledge fusion with justification matrices, and finally further encodes textual conceptual-level semantic information and conceptual relations by Bi-LSTM. Additionally, Bi-LSTM is employed to further encode contextual information at the conceptual level, facilitating the reasoning and identification of semantic consistency and implication relationships.

The model is divided into six layers, namely, word embedding layer, Transformer layer, Attention layer, Bi-LSTM layer, average pooling and maximum pooling layer, and fully connected layer. Bi-LSTM layers are well-suited for sequential data because they process the input from both directions (forward and backward), allowing the model to learn context from past and future elements in the sequence simultaneously. It contains the process of processing the semantic original information, and the specific structure is shown in Fig. 1.

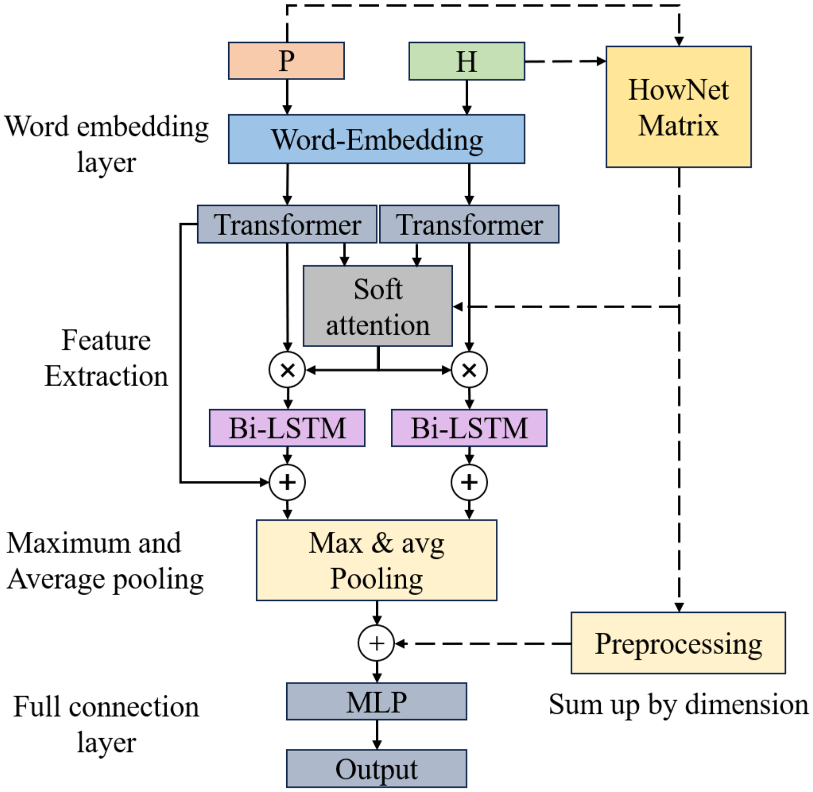


Fig. 1. Structure of the model.

The role of the transformer layer is to make the vectorized text pass through the neural network to obtain deep semantic information. Commonly used neural networks are convolutional neural network, long and short-term memory network, etc. The model uses the Transformer architecture as the encoding layer of the text to process the sentence vector. This model uses the Transformer architecture as the coding layer of the text to process the sentence vectors.

The attention layer is a widely used and essential element in text semantic matching models. It offers benefits such as high speed, effectiveness, and a low number of parameters. Various types of attention mechanisms exist, including soft attention, hard attention, self-attention, as well as focused and localized attention techniques. In this chapter, the commonly used Soft-attention mechanism (Soft-attention) is used, but the semantic matrix information generated based on How Net is added to it, which as shown in Fig. 2, and the trainable weights HN_{col}.

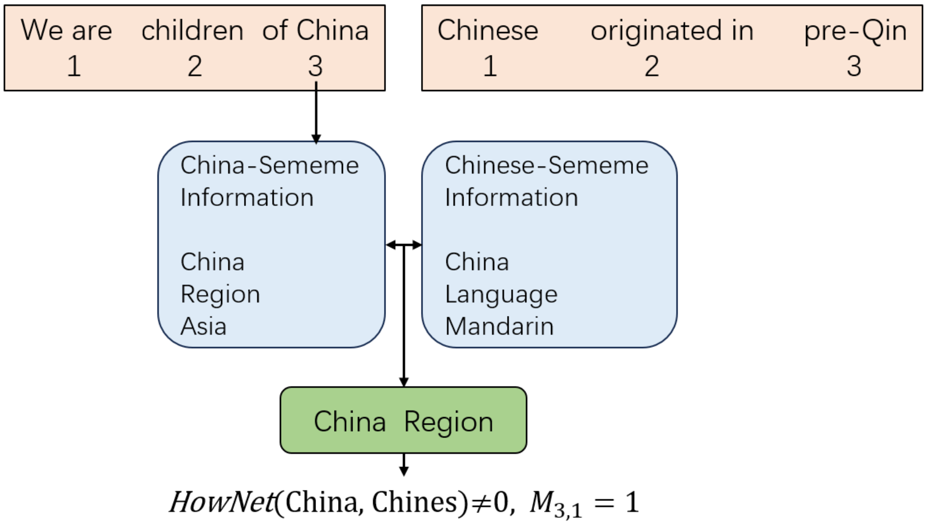


Fig. 2. Semantic information computation.

The top box displays the outcomes of sentence disambiguation, while the middle box shows the various etymological details associated with the current word. The bottom box illustrates the intersection of the etymological information of the two words. As can be seen from the above diagram, “China” has the meanings of China, related to a specific country, place, Asia, etc., and “Huaxia” has the meanings of borrowing, finance, retaining, China, country, etc. The intersection of the semantic principles of the two words is China, country, place, so at this point in time, “China, Huaxia” has the meanings of China, country, and hence $HowNet(\text{China}, \text{Huxia}) \neq 0$.

$$M_{i,j} = \begin{cases} 1 & HowNet(P_i, H_j) \neq 0 \\ 0 & HowNet(P_i, H_j) = 0 \end{cases} \quad (1)$$

$$M = \begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \end{bmatrix} \quad (2)$$

The formula indicates that a query for the i-th word in sentence P and the j-th word in sentence H is performed for the meaning of the original, and if the query result is valid, then the value of the corresponding (i,j) position in the M matrix is one.

Attention matrix e is generated as Eq:

$$e = PH^T + \gamma \cdot M \quad (3)$$

In this context, γ is a training factor. The attention matrix e not only combines textual information between sentences but also captures the semantic details of inter-sentence word pairs. As depicted in Fig. 3, the heat map of the matrix changes. Upon incorporating the original justification information, the weight of specific positions increases, indicating that these positions acquire information from the original justification matrix. In addition to this Attention matrix generation method, it is also possible to transform the various justification information of the current word into a continuous vector expression, which can be fused and embedded into the corresponding word vectors according to their weights, and there are various methods for this embedding method, such as embedding it into the Transformer structure, which makes the semantic information and word vector expression fuse.

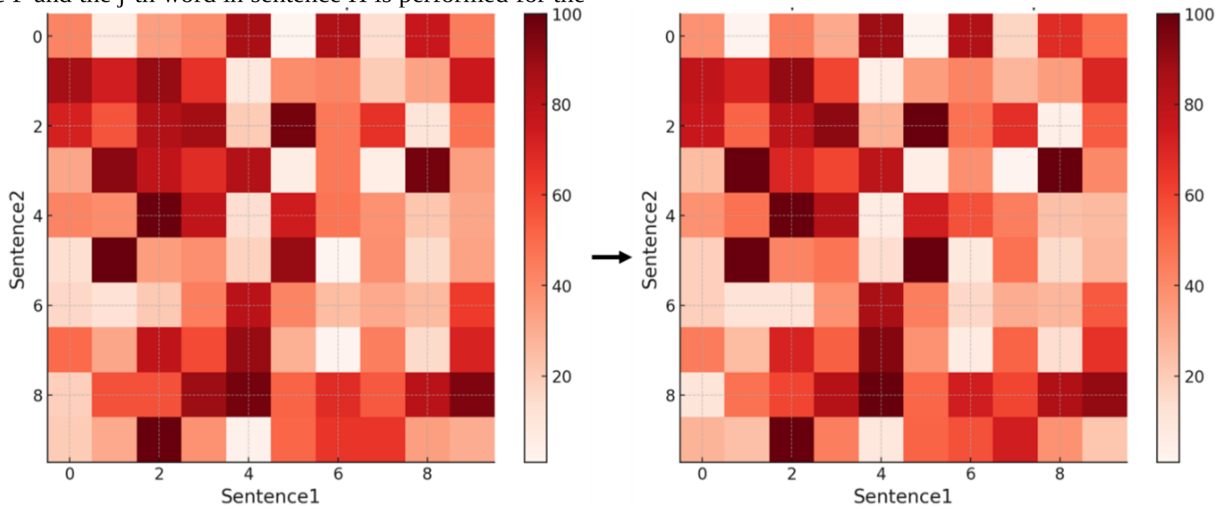


Fig. 3. The heat map of matrix change.

After obtaining the improved attention matrix e, the soft attention is computed as follows:

$$\hat{P} = \sum_{j=1}^{l_h} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_h} \exp(e_{ik})} P_{tf}, \forall i \in [1,2,\cdots,l_p] \quad (4)$$

$$\hat{H} = \sum_{j=1}^{l_p} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_p} \exp(e_{ik})} H_{tf}, \forall i \in [1,2,\cdots,l_h] \quad (5)$$

Where P_{ff}, H_{ff} are the matrix vectors of sentence pairs after Transformer. l_p, l_h denote the sentence lengths, and $\hat{P}$, $\hat{H}$ denote the outputs after soft attention mechanism.

The Bi LSTM layer is used to process the outputs P ̂and H ̂ of the transformer layer after the soft attention mechanism, and the two-way. The Bi-LSTM layer processes the outputs $\hat{P}$ and $\hat{H}$ from the Transformer layer following the application of the soft attention mechanism. The bidirectional Long Short-Term Memory Network (Bi-LSTM), which combines forward and backward encoding, enhances the acquisition of contextual information.

$$P_{bi-lstm} = BiLSTM(\hat{P}) \quad (6)$$

where $\hat{P}$ is the output vector of sentence P after the soft attention mechanism, and $P_{bi-lstm}$ is the vector after the bi-directional long-short-term neural network. Bi-LSTM is spliced from a combination of forward LSTM and backward LSTM, and compared to the long-short-term memory network LSTM, which is incapable of encoding the information passed from the back of the text to the front, Bi-LSTM can better capture the bi-directional semantic dependencies of the text.

$$P_o = Concat([P_{tf}; P_{bi-lstm}]) \quad (7)$$

$$P_{rep} = [Max(P_o); Mean(P_o)] \quad (8)$$

Here, P_{tf} represents the output vector of sentence P after the soft attention mechanism, while $P_{bi-lstm}$ is the vector resulting from the bidirectional long-short-term neural network, $Max(P_o)$ denotes the maximum pooling result of P_o, and $Mean(P_o)$ denotes the average pooling result of P_o. In most cases, maximum pooling can effectively improve the model performance, which can suppress the panning distortion and noise, etc., and also reduce overfitting. Average pooling can eliminate the effect of local maxima and make the model more stable. The combination of these two pooling methods can make the vector representation contain more information, which can improve model performance significantly.

After obtaining the complete sentence vector expressions P_{rep} and H_{rep} of sentence pairs, in the proposed model, the concatenation process also integrates information from the HowNet matrix. The sums of the two dimensions of the HowNet matrix, HN_{row} and HN_{col}, are computed and concatenated with P_{rep} and H_{rep}. This concatenated result forms the final input H for the feedforward neural network.

$$HN_{row} = sum(M, axis = 0) \tag{9}$$

$$HN_{col} = sum(M, axis = 1) \tag{10}$$

$$H = concat\left(P_{rep}; H_{rep}; P_{rep} - H_{rep}; HN_{col}; HN_{row}\right) \tag{11}$$

Where sum(•) means summing along the axis dimension. Taking HN_{row} as an example, HN_{row} represents the result of summing the HowNet matrix according to the first dimension, which corresponds to the HowNet information for sentence 1., and HN_{col} similarly represents the result of summing the HowNet matrix along the second dimension. Through vector splicing, H obtains the corresponding justification original information of the two sentences, where $P_{rep} - H_{rep}$ contains the discrepancy between the two sentence vectors.

After acquiring the final sentiment vector represents H for the sentence pairs, the model utilizes a two-layer fully connected neural network to determine the final matching result. The loss function employed is the cross-entropy loss function, calculated as follows:

$$Loss = \frac{1}{N}\sum_i - [y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)] \tag{12}$$

Here, y_i represents the label of sample i, with 1 indicating positive samples and 0 indicating negative samples. p_i denotes the predicted probability that sample i is a positive sample. Besides the standard cross-entropy loss function, this study also explores the KL Divergence loss function to evaluate the model's performance. The formula for the CoSent loss function in the context of binary classification is as follows:

$$\mathrm{KL}(P \parallel Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{13}$$

"Where " P(i)" is the true probability distribution and " Q(i)" is the predicted probability" " distribution." In the context of this study, using KL Divergence loss helps ensure that the positive sample pairs were more closely alike than the negative sample pairs, effectively maximizing the separation between positive and negative samples in the vector space. Experiments show that for pre-training methods such as BERT, SentenceBERT, etc., the method of directly using the pre-vector model to generate sentence vectors and then connecting them to the fully-connected layer for prediction is effective, and this method makes the pre-training model converge more quickly. For the model proposed in this chapter, it is not as effective as the cross-entropy loss function in the non-pre-training case.

During the training phase, MultiStepLR was utilized to progressively adjust the learning rate. The learning rate was updated with a decay rate of 0.5 at the 20th, 40th, 80th, 120th, and 160th iterations. By adjusting the learning rate dynamically as the number of iterations increases, the model's convergence speed is enhanced. The trend of the learning rate is illustrated in Fig. 4.

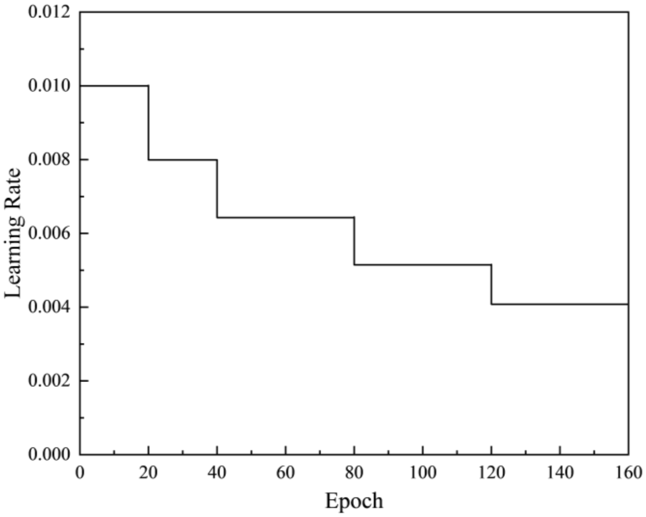


Fig. 4. Trends in learning rates.

IV. Question and Answer Engine Design

The Q&A system model was executed on a host computer having Intel Core i7-8750H CPU, NVIDIA 1060 GPU with 8 GB RAM, and 16 GB RAM. To implement this model, the Python programming language was used and the TensorFlow 1.9 deep learning framework was used. This system is capable of taking the user questions into a vector-based text retrieval, rank the top ten matching answer candidates, and then using the pre-trained.

Then the pre-trained Transformer model is used to extract the answers accurately. Therefore, this system includes database module, pre-processing technology, vector-based document retrieval, Transformer-based answer extraction and the final user operation platform, and its overall structure design is shown in Fig. 5.

The intelligent Q&A system mainly includes two roles: administrator and user, and it mainly realizes the functions of intelligent Q&A, knowledge map display and standardized text search. Users can quickly ask natural language questions in the domain, and the system automatically parses the semantic information involved in the questions and returns the correct answers through the search of the knowledge base, which mainly includes multiple use cases, such as asking questions, checking the semantic parsing, checking the returned answers, checking the corresponding knowledge base, checking the canonical retrieval text, and editing personal information, etc. Administrators mainly manage and maintain the system's knowledge base, document database, and user privileges. The manager is mainly responsible for managing and maintaining the system's knowledge base, document database, and user privileges. It mainly contains question category information management, knowledge base management, specification storage management, user information management and other major use cases (see Tables I, II, III and IV).

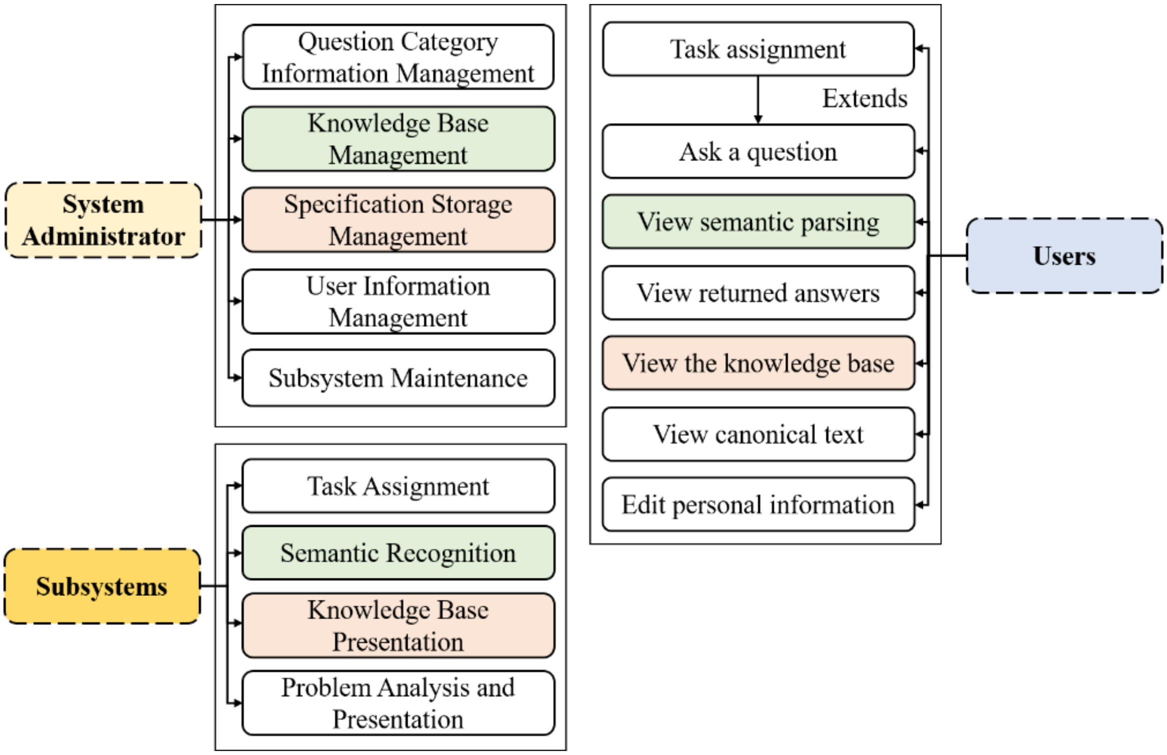


Fig. 5. Question and answer engine architecture design.

TABLE I. USE CASE DESCRIPTION OF THE QUESTION FUNCTION

Use Case Name	Question
Use Case Number	CIVILQASYS001
Use Case Description	User fills out natural language questions on the system.
Participants	Registered users of the system, database, quiz model
Pre-conditions	User login to the quiz system
Event flow	1. Registered user logs into the system 2. Click on button enter the Q&A detail information filling page. 3. Fill in the natural language question 4. Click Q&A, and the question is synchronized to the platform.
Postconditions	Can be manually selected to browse the display mode of the proposed questions.
Remark process	Fill in the natural language question format error reminder, re-entry Users abandon posting, whether to save as a draft

TABLE II. VIEW THE SEMANTIC ANALYSIS FUNCTION USE CASE DESCRIPTION

Use Case Name	View Semantic Parsing
Use Case Number	CIVILQASYS002
Use Case Description	Based on the natural language questions filled in by users on the system, the system returns parses of the questions
Participants	The system registers the user, the database, and the Q&A model.
Pre-conditions	User submits natural language question information to the Q&A model.
Event flow	1. Registered users log in and enter the system 2. Click on the Q&A module button in the platform's Q&A module to enter the Q&A details filling page. 3. Fill in the natural language question 4. Click Q&A, the question is synchronized to the platform, and the platform returns the parsing of the question.
Postconditions	Can manually choose to browse the question analysis display mode
Remark process	No

TABLE III. GET ANSWER FUNCTION USE CASE DESCRIPTION

Use Case Name	View Answer
Use Case Number	CIVILQASYS003
Use Case Description	System returns answers for users to view
Participants	System registers users, database, and quiz model
Pre-conditions	User submits natural language question information to the Q&A model.
Event flow	1. Registered users log in and enter the system 2. Click on the Q&A module button in the platform's Q&A module to enter the Q&A details filling page. 3. Fill in the natural language question 4. Click Q&A, the question is synchronized to the platform, and the platform returns the answer and displays it.
Postconditions	Can manually select how to browse the question analysis display
Remark process	The answer is returned as empty or error alert, resubmit to get the answer.

TABLE IV. VIEW THE USE CASE DESCRIPTION OF THE CORRESPONDING KNOWLEDGE BASE FUNCTION

Use Case Name	View Answer
Use Case Number	CIVILQASYS004
Use Case Description	The user fills in natural language questions on the system and gets the knowledge base corresponding to the answers returned by the system
Participants	The system registers the user, the graph database, and the Q&A model.
Pre-conditions	User submits natural language question information to the Q&A model.
Event flow	1. Registered users log in and enter the system 2. Click on the Q&A module button in the platform Q&A module to enter the Q&A detail information filling page. 3. Fill in the natural language question, click Q&A, and the question is synchronized to the platform. 4. The platform returns the corresponding map display
Postconditions	Click and drag the map manually
Remark process	No

V. RESULTS AND DISCUSSION

A. Experimental Data Collection

In order to validate the effectiveness of the model, this paper uses the BQ Corpus dataset (Bank Question Corpus) as a data source for training and testing. The BQ dataset, published by the Intelligent Computing Research Center of Shenzhen Graduate School, Harbin Institute of Technology, is a question and answer dataset in the field of banking and finance containing 120,000 question and answer pairs from online banking customer service logs. Question-answer pairs, of which 100000 are used as the training set, 10000 as the training set, and 10000 as the validation set, which is a binary classification dataset. For example, “How did the microparticle loan disappear” and “The microparticle loan is gone” are labeled as 1, indicating that they are entailment, and “How can I change my card?” and “please change your card faster” are labeled 0, indicating a contradiction. There are a large number of synonyms and near-synonyms in this dataset, which is a good way to test the validity of the model. Table V shows example of BQ dataset.

TABLE V. EXAMPLE OF BQ DATASET

Sentence 1	Sentence 2	Tags
Does Microsoft spending count?	How much is left to pay back?	0 (contradiction)
What are the best products for next week?	What financial products are available in January	1 (implied)
Can you check your bill?	I can check your bill	0 (contradiction)
Can't borrow	qq has particulate generation	0 (contradiction)

B. Evaluation Indicators

In this chapter, Accuracy (Acc) and the F1-score are employed as evaluation metrics. Accuracy is a commonly calculated value used to assess a model's classification performance, representing the percentage of correct predictions on a dataset. The F1-score, a statistical measure of a binary classification model's accuracy, evaluates the performance of a binary text semantic matching model. The F1 score is the harmonic mean of precision and recall, offering a balanced measure that takes both into account. The formulas for both metrics are presented below:

$$Acc = \& \frac{TP+TN}{TP+FN+FP+TN} \&\& \tag{14}$$

$$P = \frac{TP}{TP+FP} \& \tag{15}$$

$$R = \frac{TP}{TP+FN} \& \tag{16}$$

$$F1 - scor\&e = 2 \times \frac{P \times R}{P+R} \& \tag{17}$$

Where TP is the true case, TN is the true negative case, FP is the false positive case and FN is the false negative case.

A 4-card GPU server with model RTX2080ti was used for the experiments in this chapter. The model training parameters as well as the software version are shown in Table VI. The software versions are as follows: Python 3.6.13, PyTorch 1.10.2, OpenHowNet 2.0, Transformer 4.18.0.

C. Test Results

In the pre-training experiments, BERT processes text at the word level. To obtain the vector for each word, this model extracts the word vectors, concatenates them, and then applies average pooling. The resulting vector is used as the representation for the current word. The selected dataset for these experiments is the BQ Corpus. To ensure consistency, all models use the same Jieba word list for this dataset. The metrics used for comparison are accuracy (Acc) and F1 score. Table VII presents the relative enhancement value, which indicates the percentage improvement in effectiveness of the proposed model compared to the baseline model with the highest performance.

From Table VII, it can be seen that the model proposed in this paper has higher accuracy Acc and F1 than the other models on the BQ Corpus dataset, and on the non-pre-trained model, the accuracy Acc improves by 2.21% and the F1 value improves by 1.98% when compared to the best performing DSSM model. While on the pre-trained model, the accuracy is improved by 0.177% and F1 value by 0.464% compared to the native BERT-wwm-ext. This indicates that the semantic information between two sentences plays a more important role in processing the semantic information of banking and finance, and there is a large improvement on the non-pre-trained model, while the improvement on the pre-trained model is more limited, mainly because there are fewer words that can be matched in the sentence pairs of this dataset, which leads to a certain limitation of the experimental effect.

TABLE VI. MODEL PARAMETERS

Parameters	Numerical value
Word Embedding Layer Dimension	500
Number of hidden layers	164
Maximum sequence length	80
Batch size batch_size	128
Transformer encoder layers	12
Model Optimizer	AdamW
Initial learning rate	0.05

TABLE VII. EXPERIMENTAL RESULTS FOR THE BQ DATASET

Model	**Pre-trained or not**	**Acc**	**F1**
DSSM	×	0.7693	0.7594
MwAN	×	0.7421	0.7289
DRCN	×	0.7485	0.7576
Ours	×	**0.7902**	**0.7691**
Relative promotion	×	**+2.21%**	**+1.98%**
BERT-wwm-ext	√	0.8391	0.8402
BERT	√	0.8466	0.8399
Ours-BERT	√	0.8495	0.8433
Relative promotion	√	**+0.177%**	**+0.464%**

Regarding error analysis, since both methods directly use Jieba for text preprocessing, the segmentation errors produced by Jieba have varying degrees of impact on the experimental results. Although there is a word separation error, for the same dataset, all the models use the same word list, in comparison, the model proposed in this study has better results. Table VIII shows practical application test results.

TABLE VIII. PRACTICAL APPLICATION TEST RESULTS

Sentence pairs	Model name	Predicted results	True label
A: Unbind Chants B: Cancel the opening of chanting	ESIM	×	√
	MwAN	×	
	Our model	√	

To understand the relative importance and effectiveness of various components of the model, an ablation study was performed on the different structures of the proposed model. This study also aimed to examine the extent to which the granularity of disambiguation affects the model's performance, experiments were conducted to evaluate the effect of using three different disambiguation tools, Jieba, PKUseg and HanLP, on the results of the experiments, which were conducted using the BQ Corpus dataset.

TABLE IX. EXPERIMENTAL RESULTS OF DIFFERENT SEGMENTATION TOOLS

Segmentation tool used	Whether using HowNet	Acc	F1
Jieba	√	**0.7892**	**0.7668**
	×	0.7778	0.7627
PKUseg	√	**0.7876**	**0.7667**
	×	0.7783	0.7622
HanLP	√	**0.7861**	**0.7608**
	×	0.7742	0.7521

From the experimental results Table IX, it can be obtained that the use of HowNet can improve the performance of the model to a certain extent, and the accuracy is improved under various word segmentation tools compared to not using HowNet. The different accuracies of different segmentation tools are in line with the expected estimation because the segmentation tools have subtle differences for very few words. When a data sample contains words with extensive original information and complex relationships with other words, incorporating an external knowledge base can significantly enhance the model's sensitivity to polysemous and near-synonymous words. This integration improves the model's comprehension of synonym information and substantially boosts its overall performance.

In order to evaluate the influence of the number of layers of Transformer on the experimental effect, multi-layer Transformer experiments were carried out, and the experimental results are shown in Fig. 6.

From the data presented in Fig. 6, it is evident that the model's performance improves with an increased number of Transformer layers, a trend also observed in the BERT model. While batch stacking Transformer encoding layers can enhance model performance to some extent, it also leads to a significant increase in the model's parameters, training time, and a slower convergence speed.

In the context of non-pre-trained models, the one with the highest F1 score, featuring six Transformer encoding layers, is considered optimal and has 16 million parameters. In comparison, the DRCN model achieves the success results in non-pre-trained settings, has 19 million parameters. This indicates that the model shown in this paper has fewer cost than the DRCN model, suggesting better suitability for lightweight deployment on the BQ dataset while achieving superior experimental results.

During training, the trainable parameter γ varies with the number of iterations. As shown in Fig. 7, observing the change in γ of the attention matrix reveals that as iterations increase, the weight of the original information matrix from HowNet within the attention matrix gradually increases and stabilizes. This indicates that the semantic raw information generated by word pairs in the original text through HowNet positively impacts the model's performance enhancement.

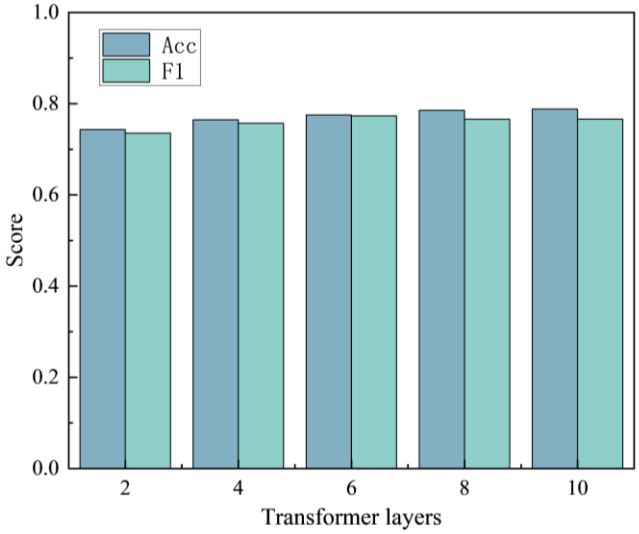


Fig. 6. Experimental results with different number of transformer layers.

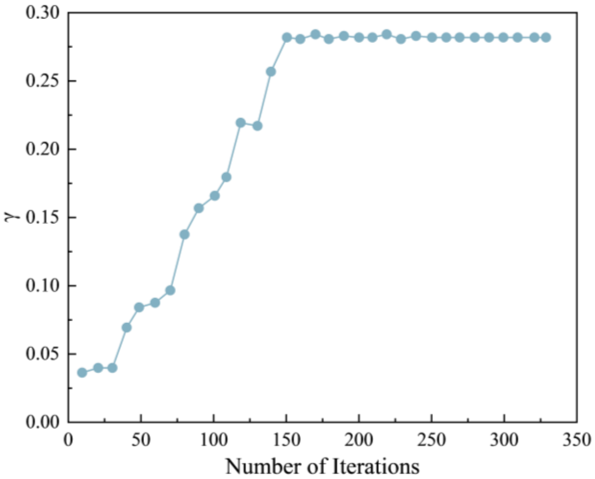


Fig. 7. Plot of changes in trainable parameters.

VI. Conclusion

This study presents a significant advancement in the field of automated question and answer (Q&A) systems by leveraging the integration of deep learning and natural language processing (NLP) technologies. With the ever-growing volume of information on the Internet, traditional keyword matching techniques have become increasingly inadequate for addressing the complexity and diversity of user queries. Our approach, which combines the Transformer model with the HowNet knowledge base, provides a robust solution to enhance semantic understanding and contextual relevance in automated Q&A systems. The architecture of our proposed system is meticulously designed, incorporating multiple layers for word embedding, Transformer encoding, attention mechanisms, and Bi-LSTM processing. This multi-layered approach allows for a comprehensive analysis and processing of natural language queries, facilitating accurate and contextually appropriate responses. The inclusion of the HowNet knowledge base is particularly noteworthy, as it enables the system to handle synonyms and near-synonyms more effectively, a critical aspect when dealing with Chinese text.

The experimental evaluation using the BQ Corpus dataset, which consists of question-and-answer pairs in the banking and finance domain, underscores the efficacy of our model. The results demonstrate substantial improvements in both accuracy and F1-score compared to existing models. Specifically, our system achieved an accuracy improvement of 2.19% and an F1-score improvement of 1.96% over the best-performing non-pretrained model, DSSM. For pre-trained models, our system showed an accuracy improvement of 0.177% and an F1-score improvement of 0.464% over the native BERT-wwm-ext model. These enhancements validate the effectiveness of integrating external knowledge bases with deep learning models in improving the performance of automated Q&A systems. Furthermore, the incorporation of HowNet significantly enhances the system's ability to process and understand the semantic relationships between words, leading to more accurate and contextually relevant responses. This is particularly important in domains where precise information retrieval is critical, such as banking and finance.

This study's findings underscore the significance of integrating external knowledge bases, such as HowNet, with advanced deep learning models to address the limitations of traditional keyword matching methods. The proposed system enhances both the efficiency and accuracy of information retrieval while providing a more user-friendly approach to accessing precise and meaningful data. This improvement is vital in the era of big data and web 3.0, where rapid and accurate access to relevant information is essential for users and organizations alike.

Future research and development offer several promising directions. Enhancing the knowledge base with more diverse and comprehensive datasets could further augment the system's capabilities. Additionally, exploring alternative deep learning architectures, including reinforcement learning and more advanced attention mechanisms, may yield further performance enhancements. Applying this system across various domains beyond banking and finance, such as healthcare, legal, and customer service, could demonstrate its versatility and broad applicability. In conclusion, the integration of deep learning with external knowledge bases represents a promising avenue for developing automated Q&A systems. This study lays a solid foundation for future advancements, paving the way for more efficient, accurate, and user-friendly information retrieval systems in the big data and web 3.0 era.

References

[1] Fan, Jianqing, Fang Han, and Han Liu. "Challenges of big data analysis." National science review 1.2 (2014): 293-314.

[2] Gan, Wensheng, et al. "Web 3.0: The future of internet." Companion Proceedings of the ACM Web Conference 2023. 2023.

[3] Tzenios, Nikolaos. Corporate Espionage and the Impact of the Chinese Government, Companies, and Individuals in Increasing Corporate Espionage. Apollos University, 2023.

[4] Sejnowski, Terrence J. "Large language models and the reverse turing test." Neural computation 35.3 (2023): 309-342.

[5] Gao, Rujun, et al. "Automatic assessment of text-based responses in post-secondary education: A systematic review." Computers and Education: Artificial Intelligence (2024): 100206.

[6] Ciesla, Robert. The Book of Chatbots: From ELIZA to ChatGPT. Springer Nature, 2024.

[7] Aithal, Shivani G., Abishek B. Rao, and Sanjay Singh. "Automatic question-answer pairs generation and question similarity mechanism in question answering system." Applied Intelligence (2021): 1-14.

[8] Chen J, Zhang R, Mao Y, Wang B, Qiao J. 2019. A Conditional VAE-Based Conversation Model. Communications in Computer and Information Science, 165-174 .

[9] Gu X, Cho K, Ha J, Kim S. 2019. Dialogwae: Multimodal response generation with conditional Wasserstein auto-encoder. The 7th International Conference on Learning Representations, 56-63.

[10] Staudemeyer, Ralf C., and Eric Rothstein Morris. "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks." arXiv preprint arXiv:1909.09586 (2019).

[11] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." arXiv preprint arXiv:1508.01991 (2015).

[12] Arsovski S, Cheok A D, Govindarajoo K, Salehuddin N, Vedadi S. 2020. Artificial intelligence snapchat: Visual conversation agent. Applied Intelligence, 50(7):2040-2049

[13] Kolomiyets, Oleksandr, and Marie-Francine Moens. "A survey on question answering technology from an information retrieval perspective." Information Sciences 181.24 (2011): 5412-5434.

[14] Bowman S R, Vilnis L, Vinyals O, Dai A M, Jozefowicz R, Bengio S. 2016. Generating sentences from a continuous space. The 20th SIGNLL

Conference on Computational Natural Language Learning, Berlin, Germany, 10-21

[15] Cai Y, Zuo M, Zhang Q, Xiong H, Li K. 2020. A Bichannel Transformer with Context Encoding for Document-Driven Conversation Generation in Social Media. Complexity, 48(7):1-13.

[16] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." Neural computation 31.7 (2019): 1235-1270.

[17] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.

[18] Zhou, Chunting, et al. "A C-LSTM neural network for text classification." arXiv preprint arXiv:1511.08630 (2015).

[19] Cahuantzi, Roberto, Xinye Chen, and Stefan Güttel. "A comparison of LSTM and GRU networks for learning symbolic sequences." Science and Information Conference. Cham: Springer Nature Switzerland, 2023.

[20] Xu, Weidong, et al. "Long-short-term-memory-based deep stacked sequence-to-sequence autoencoder for health prediction of industrial workers in closed environments based on wearable devices." Sensors 23.18 (2023): 7874.

[21] Abumohsen, Mobarak, Amani Yousef Owda, and Majdi Owda. "Electrical load forecasting using LSTM, GRU, and RNN algorithms." Energies 16.5 (2023): 2283.

[22] Lu, Guangquan, et al. "Multi-task learning using variational auto-encoder for sentiment classification." Pattern Recognition Letters 132 (2020): 115-122.

An Artificial Neural Network Model for Water Quality Prediction in the Amoju Hydrographic Subbasin, Cajamarca-Peru

Alex Alfredo Huaman Llanos[1], Jeimis Royler Yalta Meza[2], Danicza Violeta Sanchez Cordova[3], Juan Carlos Chasquero Martinez[4], Lenin Quiñones Huatangari[5], Dulcet Lorena Quinto Sanchez[6], Roxana Rojas Segura[7], Alfredo Lazaro Ludeña Gutierrez[8]

Informatic and Language Center, National University of Jaen, Jaen, Peru[1]
Direction of Production Center of Goods and Services, National University of Jaen, Jaen, Peru[2]
Soil and Water Analysis Center, National University of Jaen, Jaen, Peru[3, 6]
Geospatial Analysis and Computing Laboratory, National University of Jaen, Jaen, Peru[4]
Professional School of Data Science Engineering and Artificial Intelligence,
Toribio Rodriguez de Mendoza National University, Chachapoyas, Peru[5]
Biotechnology-Genetics and Molecular Biology Laboratory, National University of Jaen, Jaen, Peru[7]
Faculty of Engineering, National University of Piura, Piura, Peru[8]

Abstract—**Water quality is crucial for sustaining life, and accurate prediction models are essential for effective management. This study introduces an Artificial Neural Network (ANN) model designed to predict the Water Quality Index (WQI) in the Amoju Hydrographic Subbasin, Cajamarca-Peru. The model was developed using key water quality parameters, including electrical conductivity (EC), total dissolved solids (TDS), calcium carbonate ($CaCO_3$), and phosphate (PO_4^{3-}), identified through Pearson correlation analysis. Data from water samples collected over six months were used to train and validate the model. Results revealed that the ANN model achieved high predictive accuracy, with a significant correlation between WQI and the aforementioned parameters. The model's performance outstrips traditional methods demonstrating its capability to effectively capture complex interdependencies among water quality indicators. This research emphasizes the potential of AI-driven approaches for enhancing predictive accuracy in environmental monitoring. Future studies should consider incorporating additional variables, such as heavy metals and microbial indicators, and consider the application of real-time AI-driven monitoring systems to further refine water quality management strategies. The ANN model presented here offers a promising tool for decision-makers, providing a reliable method for predicting water quality in similar hydrographic basins and contributing to the broader field of AI in environmental science.**

Keywords—Artificial neural networks; hydrographic subbasin; machine learning models; water quality index; water resource management

I. Introduction

Water is a critical resource upon which all life on Earth depends, with its quality playing a pivotal role in human health and aquatic ecosystems. The contamination of water sources poses significant threats to public health and the environment, underscoring the urgency of effective water quality monitoring and management. Numerous studies have highlighted the importance of assessing and predicting water quality to safeguard the sustainability and safety of water resources [1], [2]. Water quality forecasting is an indispensable method for effective water resource planning, regulation, and monitoring, and it is a crucial component of research focused on water ecological protection [3], [4].

The global issue of water pollution, aggravated by industrialization and urbanization, has driven the development of advanced methodologies for monitoring and predicting water quality. Traditional approaches, such as manual sampling followed by laboratory analysis, are often labor-intensive, time-consuming, and costly, thereby limiting their efficiency and scalability. These challenges have catalyzed the integration of artificial intelligence (AI) and machine learning (ML) techniques into water quality assessment, offering more efficient and cost-effective solutions for real-time water quality prediction [5], [6]. Among these techniques, machine learning models, particularly artificial neural networks (ANNs), have gained widespread adoption due to their capability to handle complex, and nonlinear relationships within environmental data [7], [8].

Recent advancements in technology, including remote sensing (RS), the Internet of Things (IoT), and big data analytics, have further enhanced water quality monitoring by enabling the collection and processing of vast amounts of data form diverse source. The synergy between these technologies and AI has facilitated the development of more accurate and reliable predictive models, capable of providing comprehensive assessments of water quality status [9], [10]. Specifically, metrics such as the Water Quality Index (WQI) and Water Quality Classification (WQC) are commonly employed to aggregate multiple water parameters into a single, interpretable value, providing a holistic overview of water quality [11].

This study is focused on developing an Artificial Neural Network (ANN) model to predict water quality in the Amoju Hydrographic Subbasin located in Cajamarca Peru. By leveraging various water quality indicators, the model aims to forecast the WQI and classify the water quality status, thereby contributing valuable insights for water resource management and pollution control. Through the application of advanced AI techniques, this research addresses the pressing need for efficient water quality prediction methods that ensure the provision of safe and clean water for diverse uses, while also mitigating the adverse effects of water contamination on public health and the environment [12], [13].

The structure of the paper is organized as follows: Section 2 reviews the relevant literature on water quality prediction using various classifiers. Section 3 details the materials and methodologies employed, including data preparation, pre-processing, splitting, distribution, feature correlation, and WQI computation. The experimental setup and the result analysis are discussed in Section 4. Finally, the paper is concluded with limitations and future scope in Section 5.

II. Literature Review

The literature on water quality prediction has undergone a considerable transformation, particularly with the increasing adoption of Artificial Neural Networks (ANNs) in environmental modeling. This paradigm shift is evident in the application of ANNs within hydrographic subbasins, such as the Amoju Subbasin in Cajamarca, Peru. The study titled “An Artificial Neural Network Model for Water Quality Prediction in the Amoju Hydrographic Subbasin, Cajamarca-Peru” contributes significantly to this evolving field by addressing the critical need for accurate predictive models that can inform environmental management and policy in the region. ANNs, inspired by the neural structures of the human brain, excel at capturing and modeling the intricate non-linear relationships prevalent in environmental datasets. These networks are particularly effective in dealing with the complex dynamics of hydrographic subbasins, such as the Amoju Subbasin, which holds significant ecological value and is vulnerable impacts of anthropogenic activities.

Recent studies further underscore the potential of artificial intelligence (AI) in enhancing water quality prediction and monitoring. For instance, [14] and [15] developed AI models focusing on water quality index (WQI) prediction and water quality classification (WQC). The study in [14] utilized adaptive neuro-fuzzy inference system (ANFIS) algorithms for WQI prediction and feed-forward neural network (FFNN) for WQC, achieving high predictive accuracy. The study in [15] also demonstrated the efficacy of AI in water quality monitoring, affirming the robustness of these models in managing complex environmental data. The research in [16] reviewed the integration of AI and the Internet of Things (IoT) in water quality prediction, emphasizing AI's role in analyzing intricate systems and leveraging historical data to enhance prediction accuracy. Similarly, [17] explored several AI techniques, including multilayer perceptron neural networks (MLP-ANN), ensemble methods, gaussian process regression, support vector machine (SVM), and decision tree, all of which contributed to a comprehensive evaluation of water quality parameters.

The impact of water quality on public health further emphasizes the necessity of accurate WQI modeling, a task that presents significant challenges within the water sector. The study in [18] introduced an innovative application of the ensemble Kalman filter integrated with ANNs to predict WQI using physicochemical parameters, showcasing the model's capability to handle noise in environmental data. Other researchers have explored different machine learning techniques for WQI prediction. The research in [19] employed k-nearest neighbors, boosting decision trees, SVMs, and multilayer perceptron ANNs in their models, while [20] compared deep learning-based models with other machine learning models such as random forests (RF) and extreme gradient boosting (XGBoost) for predicting groundwater quality. Their comparative study highlighted the superior performance of deep learning models in certain contexts.

The prediction of dissolved oxygen concentration, a key indicator of river water quality, has also been enhanced through AI. The study in [21] utilized a deep learning approach, applying recurrent neural networks (RNNs) to predict this parameter with high precision. Moreover, [22] proposed a hybrid model combining ANNs, discrete wavelet transforms (DWT), and long short-term memory (LSTM) networks, further advancing the state-of-the-art in water quality prediction.

In addition, [23] implemented a comprehensive architecture integrating machine learning models (RF, DT, LR, SVM, AdaBoost) with deep learning models (CNN, LSTM, GRU) for predicting both water quality and water consumption. This study underscores the potential of hybrid models in addressing multi-faceted environmental issues. The research in [24] further advanced this approach by integrating deep learning models with feature extraction technique, such as principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA), to enhance water quality classification accuracy.

In study [25], focused on the prediction of water quality using machine multiple learning techniques, including regression and classification models like SVMs, multiple linear regression (MLR), and Bayesian tree model (BTM). Their comprehensive five-step methodology, which included data pre-processing, feature correlation analysis, and model feature importance, resulted in a maximum prediction accuracy of 99.83% with the MLR classifier. The study in [26] compared decision tree algorithms (DT) and Naïve Bayes classifiers for water quality prediction, with DT emerging as the most accurate model, achieving an accuracy value of 97.23%.

Additionally, [27] proposed a machine learning-based system for predicting the WQI in the Illizi region by employing eight artificial intelligence algorithms. The results indicated that the Multivariate Linear Regression (MLR) model exhibited the highest accuracy among the models considered. In contrast, [28] explored 15 supervised Machine Learning (ML) algorithms to estimate the WQI, identifying gradient boosting and polynomial regression as the most efficient methods for WQI prediction. Further advancements in the field include the work of [29], who developed a prediction method for WQI using feedforward artificial neural networks (ANNs) with 25 water quality parameters inputs. By integrating backward elimination and forward selective combination methods, the study achieved high R2 and minimal squared error (MSE). Similarly, [30] predicted the WQI using 16 water quality parameters and successfully applied ANN through a Bayesian regularization algorithm, demonstrating the robustness of this approach. Moreover, the study in [31] examined the comparative efficiency of multivariate linear regression (MLR) and ANN models for predicting water quality parameters, such as pH, temperature, total suspended solids (TSS), and total suspended matter (TSM), in estimating the chemical oxygen demand (COD) and biochemical oxygen demand (BOD). In another study, [32] designed a feed-forward, fully-connected, three-layer perceptron neural network model to predict the WQI using 23 parameters, reinforcing the trend towards increasingly complex ANN architectures. Also, [33] took a different approach by proposing a two-layered ensemble model that integrates five commonly used methods, including partial least square, random forest, and Bayesian networks, into an ML model for forecasting beach water quality. The model stacking approach yielded the best predictions, demonstrating the advantages of ensemble methods in enhancing model robustness and accuracy. The study in [34] developed an ML-based classification system for the Chao Phraya River's water quality, integrating attribute realization (AR) and support vector machine (SVM) algorithms. The results indicated that linear regression (LR) was the most suitable function for river water data classification, offering a different perspective on the adaptability of ML techniques across diverse water bodies. The research in [35] also employed an ANN approach for calculating and simulating the WQI of the Akaki River, utilizing a neural network model trained on 12 inputs and one output. The optimal model architecture was obtained with eight hidden layers, achieved an accuracy of 0.93. Besides, [36] formulated four distinct ML techniques, including Back Propagation Neural Network (BPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Multilinear Regression (MLR), and Support Vector Regressor (SVR) for forecasting the water quality index (WQI) across the Yamuna River. The study in [37] applied independent techniques like the M5P tree model, additive regression (AR), support vector machine (SVM), and random subspace (RSS) to predict WQI, identifying AR as the most optimal approach with favorable outcomes.

The literature also explores hybrid models. The study in [38] developed a hybrid ML method combining random trees and bagging, testing four standalone and 12 hybrid data-mining algorithms for WQI forecasting in a humid climate. The study concluded that hybrid models could significantly improve prediction accuracy. In a similar vein, [39] optimized the performance of an adaptive neuro-fuzzy inferences system (ANFIS) for water quality metrics prediction using Genetic Algorithm (GA), Differential Evolution (DE), and Ant Colony Optimization (ACOR), further demonstrating the value of optimization algorithms with ML models. Consequently [40] enhanced a hybrid artificial neural network (HANN) model with a genetic algorithm (GA) for predicting water output in drinking water treatment plants in China. The HANN model has shown better ability and consistency in forecasting the total water output. The prediction shows that the HANN model has improved its performance from 0.71 to 0.93 R2 by increasing the training data provided. Likewise, the study in [41] introduced an ensemble ML model, Extra Tree Regression (ETR), for predicting monthly WQI values in Hong Kong. Achieving a high prediction accuracy with R2 = 0.98 and RMSE = 2.99. The study in [42] utilized Principal Component Regression (PCR) and Gradient Boosting Classifier (GBC) to predict WQI, demonstrating 95% prediction accuracy for PCR method and 100% classification accuracy for GBC. [43] evaluated the performance of 12 ML models, including boosting-based, decision tree-based, and ANN-based algorithms, for estimating the WQI of the La Boung River in Vietnam, with extreme gradient boosting (XGBoost) emerging as the best performer, achieving an R2 of 0.989 and RMSE of 0.107. Finally, [44] used Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Adaptive Boosting (Ada-Boost) model for predicting WQC. In contrast, K-nearest neighbor (KNN), decision tree (DT), support vector regressor (SVR), and multi-layer perceptron (MLP) were used as regression models for predicting WQI. The results showed that GB model produced the best results for predicting WQC, with an accuracy of 99.5% value, and the MLP regressor model in predicting WQI, with an accuracy of 99.8% value.

These studies collectively highlight the potential of AI, particularly machine learning and neural networks, in advancing water quality management. However, there is a notable gap in evaluating water quality classification based on metrics such as accuracy, precision, or F1 score. Moreover, many of these approaches are limited by their focus on either WQI prediction or WQC, rather than integrating both aspects. Our proposed methodology addresses these aspects by employing a lightweight model that not only enhances prediction accuracy but also integrates water quality classification with water demand prediction, paving the way for a more comprehensive approach to water resource management. Table I provides a comparative summary of the research works discussed.

The literature further reveals the versatility of artificial neural networks (ANNs) in water quality prediction, demonstrating their adaptability to various geographical and hydrological contexts. The inclusion of diverse input parameters, including meteorological data to land use and physicochemical attributes-into ANN models has consistently improved the prediction accuracy, offering a more holistic understanding of the factors influencing water quality dynamics.

TABLE I. COMPARATIVE SUMMARY OF THE DISCUSSED RESEARCH WORKS

Reference	Year	Classifiers	Achieved Accuracy
[14]	2020	NARNET, LSTM, SVM, KNN, NB	SVM 97.01%
[15]	2021	KNN, FFNN, ANIFS	WQI ANFIS 96.17%
[24]	2021	Dimension reduction PCA, LDA, ICA, RNN, LSTM, SVM (variants)	LSTM RNN with LDA, LSTM, RNN 99.72%
[25]	2021	NN, RF, MLR, SVM, BT	99.83% MLR
[26]	2021	DT, NB (variants), K-fold cross validation	97.22% DT
[27]	2021	MLR, RF, RSS, AR, ANN, SVR, LWLR	With all parameters: MLR
[28]	2019	Multiple linear regression, polynomial regression, RF, GBC, SVM, ridge regression, lasso regression, elastic net regression, MLP, GNB, LR, SGD, KNN, DT, bagging classifier	Gradient Boosting and Polynomial Regression with MAE = 1.964 and 2.727, respectively
[34]	2021	Attribute-realization (AR) and Support Vector Machine (SVM)	AR-SVM with 0.86-0.95 accuracy respectively
[36]	2019	BPNN, ANFIS, SVR and MLR	DC values vary in the range of 0.9202 to 0.9957
[37]	2022	AR, M5P tree model, RSS and SVM	AR with R2 = 0.9993, MAE = 0.5243, RSME = 0.6356, %RAE = 3.8449 and %RRSE = 3.9925
[38]	2020	RF, M5P, RT and REPT	RT with R2 = 0.941, RMSE = 2.71, MAE = 1.87, NSE = 0.941, PBIAS = 0.500
[41]	2021	ETR, SVR and DTR	ETR model produced more accurate WQI predictions with R2 = 0.98 and RSME = 2.99 values
[42]	2022	PCR and GBC methods	PCR = 95% and GBC = 100%
[43]	2022	(Adaptive boosting, GBoost, HGBoost, LGBoost, XGBoost), (DT, ET, RF), (MLP, RBF, DFFNN, CNN)	XGBoost (R2 = 0.989 and RMSE = 0.107)
[44]	2023	K-nearest neighbor (KNN) regressor model, DT, SVR, MLP	MLP regressor model outperformed the best accuracy with R2 = 99.8%

III. MATERIALS AND METHODOLOGY

Fig. 1 displays the proposed methodology for completing the research.

A. Study Area

The study was conducted in the Amoju River Subbasin, which is located within the Alto Marañon III Inter-basin in northern Peru. The subbasin covers an area of 354 km2, as depicted in Fig. 2. The Amoju River itself extends approximately 29 km, originating near the towns of San Jose de Alianza, Nuevo Jerusalen, and La Rinconada Lajeña. The river then flows into the Marañon River at the Pedregales hamlet in the Bellavista district, within Jaen province, Cajamarca department [45].

A Digital Elevation Model (DEM) from the National Aeronautics and Space Administration (NASA) [46] was employed to determine that the maximum elevation within the subbasin reaches 3222 meters above sea level, while the lowest point is situated at 408 meters above sea level. The region is characterized by steep slopes, ranging from 40º to 56º in the upper and middle sections, and from 0º to 25º in the lower section. These topographical features, as shown in Fig. 3, suggest that the area is suitable for agricultural activities and the establishment of urban centers.

Moreover, this subbasin is a critical source of water supply for human consumption, serving as the primary provider for the cities of Jaen and Bellavista, as well as their associated agricultural valleys. The locations of the sampling stations are detailed in Table II.

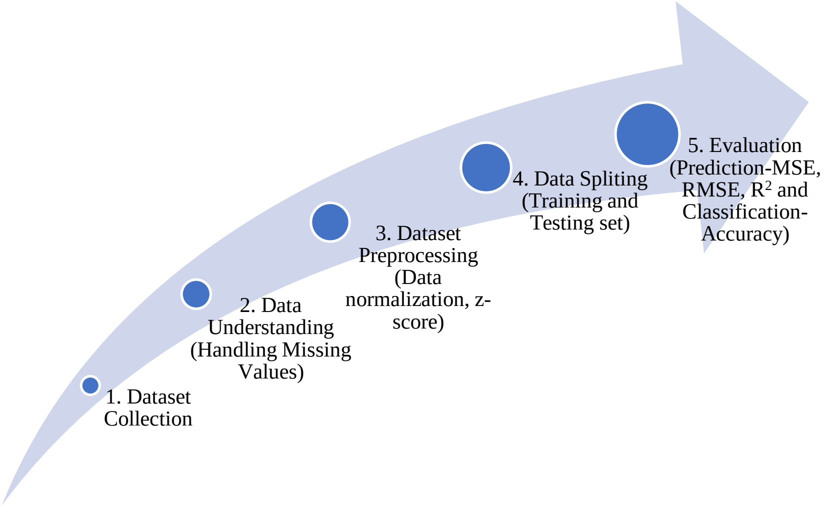

1. Dataset Collection
2. Data Understanding (Handling Missing Values)
3. Dataset Preprocessing (Data normalization, z-score)
4. Data Spliting (Training and Testing set)
5. Evaluation (Prediction-MSE, RMSE, R² and Classification-Accuracy)

Fig. 1. Framework of the proposed methodology.

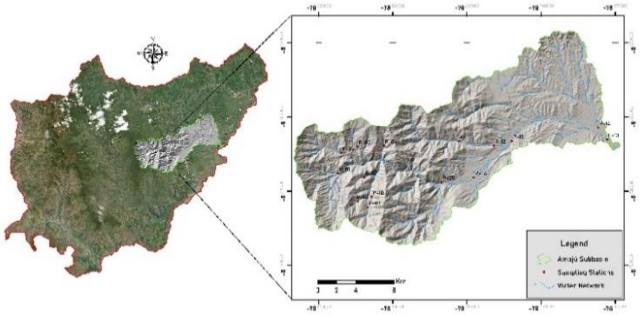


Fig. 2. Study area location map.

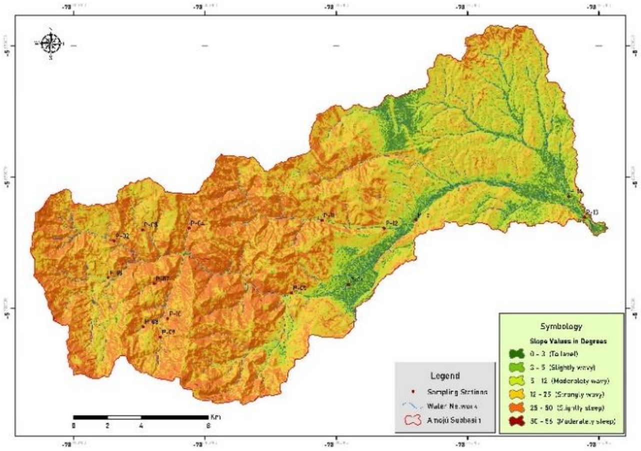


Fig. 3. A Subbasin slope map.

TABLE II. SAMPLING STATION LOCATIONS

Code	Location	Latitude	Longitude
P-01	Nuevo Jerusalén	-5.70347	-78.93161
P-02	La Rinconada Lajeña	-5.68396	-78.92851
P-03	San Antonio	-5.67808	-78.91332
P-04	La Cascarilla	-5.67719	-78.88849
P-05	Puente La Corona	-5.71176	-78.83388
P-06	Punte Pakamuros	-5.70729	-78.80355
P-07	Cruzpa Huasi	-5.70682	-78.90707
P-08	La Granadilla	-5.72972	-78.91294
P-09	La Virginia	-5.73523	-78.90385
P-10	La Victoria	-5.72548	-78.89995
P-11	Puente Tumbillán	-5.67307	-78.81757
P-12	Qda. Tumbillán- Altura La Granja	-5.6772	-78.78447
P-13	Puente Bellavista Viejo	-5.67135	-78.67758
P-14	Puente Santa Cruz	-5.66024	-78.68604
P-15	Yanuyacu - Altura UNJ	-5.67278	-78.7676

The upper section of the basin is dominated by a Tropical Premontane Rainforest (bh-PT) with annual precipitation levels reaching up to 1968 mm. The middle section is characterized by a very humid Tropical Low Montane Forest (bmh-MBT). Agroforestry systems, particularly those cultivating Coffea arabica (coffee) in association with citrus trees, are prevalent in the upper subbasin. The lower subbasin is dominated by a Tropical Premontane Dry Forest (bs-PT), where extensive rice fields (Orysa sativa) are cultivated.

B. Hydrogeological Settings

The water chemistry and quality in the study area influenced by both the lithology and the duration of water-rock interaction [47]. To identify the aquifer units within the National Geological Map provided by the Geological, Mining, and Metallurgical Institute (INGEMMET) [48] was used. This data was processed using QGIS software to create a hydrogeological map, illustrating the characteristics of the lithological units, as shown in Fig. 4.

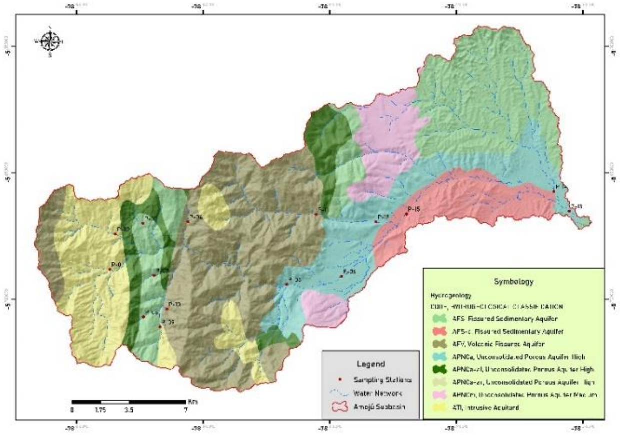

Symbology
Hydrogeology
AFS, Fissured Sedimentary Aquifer
AFS-c, Fissured Sedimentary Aquifer
AFV, Volcanic Fissured Aquifer
APNCa, Unconsolidated Porous Aquifer High
APNCa-rl, Unconsolidated Porous Aquifer High
APNCa-ai, Unconsolidated Porous Aquifer High
APNCm, Unconsolidated Porous Aquifer Medium
ATI, Intrusive Acuitard
Legend
Sampling Stations
Water Network
Amojú Subbasin
P-15

Fig. 4. Hydrogeolocgical map of the subbasin.

The lithological units within the subbasin display varied hydrogeological properties. Based on the hydrogeological map, these units were classified into aquifers and aquitards. Six distinct hydrogeological categories were identified within the aquifer units, while only one category was identified within the aquitards, as detailed in Table III.

C. Dataset Collection

Water samples were collected from 15 designated sampling points (Upper, Middle and Lower) within the Amoju Hydrographic Subbasin over a period from November 15th, 2023 to July 20th, 2024. The geographic distribution of these sampling points across the subbasin is illustrated in Fig. 5. Samples were brought to the CEASA and CAE laboratory in an insulated cooler box containing ice packs to maintain a stable temperature during transit. Analytical procedures commenced within 48 hours of sample collection to ensure data integrity.

The dataset for this study was sourced from strategic locations in Jaen-Cajamarca, focusing on five (05) physicochemical parameters measured in situ at each of fifteen (15) sampling sites. These parameters include pH, Electrical Conductivity (EC), Dissolved Oxygen (DO), total dissolved solids (TDS), and temperature (T°), all of which were measured using a recalibrated portable HANNA Multi-parameter HI 9829 device. Additionally, the dataset includes four (04) non-metallic inorganic parameters: alkalinity $(CaCO_3)$, total hardness (TH), nitrates (NO_3^{1-}), and phosphates (PO_4^{3-}). Table IV provides a detailed description of each attribute measured.

TABLE III. A DESCRIPTION OF THE HYDROGEOLOGICAL CHARACTERISTICS OF THE SUBBASIN [48]

Hydrogeological Unit	**Classification Hydrogeological**	**Code**	**Lithology**	**Description Hydrogeological**
Aquifer	Volcanic Fissured Aquifer	AFV	Andesites and Dacitas	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Fissured Sedimentary Aquifer	AFS	Lutites, intercalated with limestones, marls	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Fissured Sedimentary Aquifer	AFS-c	Conglomerates, shales and sandstones	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquifer	Unconsolidated Porous Aquifer High	APNCa-al	Alternation of shales and sands	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer High	APNCa	Alluvial, moraines, glaciofluvial, lacustrine and travertine	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer High	APNCa-ar	Sands, sandstones, gravels and conglomerates	Aquifers are extensive and highly productive, exhibiting high permeability.
Aquifer	Unconsolidated Porous Aquifer Medium	APNCm	Conglomerates, shales, mudstones	Local or discontinuous productive aquifers, or extensive aquifers, which are only moderately productive (medium permeability). This does not preclude the existence of other, more productive aquifers at greater depths.
Aquitard	Intrusive Aquitard	ATI	Acid and intermediate intrusive rocks	Formations without aquifers (with a very low permeability) can be considered.

TABLE IV. FEATURE DESCRIPTION OF DATASET

Attributes Name	**Description**
Physicochemical parameters	
pH	Water acidity and basicity.
EC	Water's ability to conduct electricity in the presence of ions.
DO	Concentration of dissolved oxygen in water.
TDS	Concentration of dissolved minerals, salts, metals, cations, or anions in water.
T°	Water temperature at the time of testing.
Non-metallic inorganic parameters	
$CaCO_3$	Water's ability to neutralize acids or resist changes that cause acidity, maintaining pH levels.
TH	Concentration of dissolved calcium and magnesium in water.
NO_3^{1-}	Concentration of nitrates, representing the most common form of nitrogen in water.
PO_4^{3-}	Concentration of phosphates, indicate of phosphorous and oxygen compounds in water.

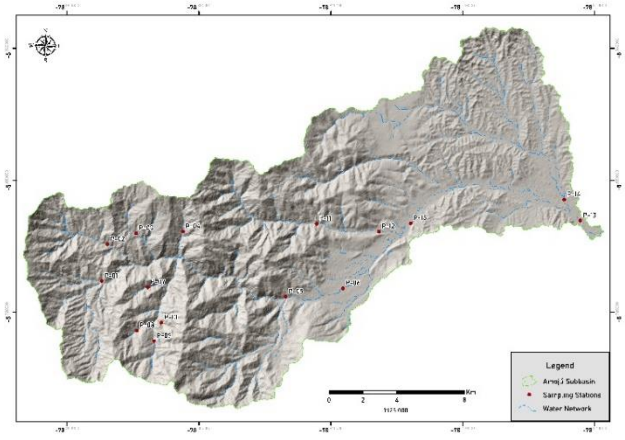


Fig. 5. Groundwater sampling sites.

D. Dataset Preprocessing and WQI Calculation

Data preprocessing is a critical step in ensuring the integrity and reliability of subsequent analysis. In this study, preprocessing involved the handling of missing data, normalization, and outlier removal to enhance the quality of the dataset. The Water Quality Index (WQI) was calculated to provide a comprehensive assessment of water quality across various sampling sites within the study area.

Water Quality Index (WQI): Methodology and Calculation

The Water Quality Index (WQI) is a widely recognized metric that synthesizes multiple water quality parameters into a single, overall score, reflecting the water's suitability for many uses [14]. The WQI in this study using the formula showing in Eq. (1).

$$WQI = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i}^{N} w_i} \quad (1)$$

Where N represents the number of parameters analyzed, q_i denotes the quality rating scale for each parameter i, computed in Eq. (2), and w_i denotes the unit weight of each parameter determined by Eq. (3).

$$q_i = 100 \times \left(\frac{V_i - V_{id}}{S_i - V_{id}}\right) \quad (2)$$

Where q_i is the parameter's actual value in the water samples tested, V_i represents the estimated value of the parameter i, V_{id} represents the ideal value under pure water conditions, and S_i represents the standard permissible limit for the parameter i as shown in Table V. The unit weight w_i is the parameter's recommended standard value as depicted in Table VI.

$$w_i = \frac{K}{S_i} \quad (3)$$

Where K denotes the proportionality constant, which is calculated using Eq. (4):

$$K = \frac{1}{\sum_{i=1}^{N} S_i} \quad (4)$$

The permissible limits and corresponding unit weights for the parameters are detailed in Table V and Table VI, respectively.

TABLE V. PERMISSIBLE LIMITS OF THE PARAMETERS USED IN CALCULATING THE WQI [49]

Parameter	Unit	Permissible Limits
pH	-	8.5
Conductivity	µS/cm	1000
Dissolved Oxygen	mg/L	10
Total Dissolved Solids	mg/L	1000
Temperature	°C	25
Alkalinity	mg/L	200
Hardness	mg/L	200
Nitrate	mg/L	45
Phosphate	mg/L	0.1

TABLE VI. PARAMETERS UNIT WEIGHTS

Parameter	Unit Weight (w_i)
pH	0.00004727
Conductivity	0.00000040
Dissolved Oxygen	0.00004018
Total Dissolved Solids	0.00000040
Temperature	0.00001607
Alkalinity	0.00000201
Hardness	0.00000201
Nitrate	0.00000893
Phosphate	0.00401830

The WQI is a versatile metric that can be employed for the calculation of numerous parameters, including those selected for analysis. The WQI depends on the variable data. The proposed system is capable of testing any parameters in conjunction with any water quality data.

IV. RESULTS AND DISCUSSION

Table VII provides descriptive statistics for the dataset attributes derived from 75 groundwater samples. These statistics, including count, mean, standard deviation, minimum, maximum, and quartiles, offer a comprehensive overview of the dataset's distribution and underlying properties. The mean pH value of 7.79 with a standard deviation of 0.46 suggests a slightly basic water quality, consistent with findings from similar studies in hydrographic subbasins [50], [51]. Conductivity averages at 220.96 µS/cm, reflecting significant variability in ion concentration among the samples. Dissolved Oxygen (DO) levels, averaging 7.46 mg/L, are critical for sustaining aquatic life, aligning with established benchmarks [52]. Total Dissolved Solids (TDS) display considerable variability with a mean of 164.61 mg/L, indicative of diverse mineral content in the samples. The mean temperature (Tº) of 21.52 ºC influences the solubility and reaction rates of various chemical constituents, further impacting water quality parameters [53].

Alkalinity and hardness, with means of 114.84 mg/L and 136.57 mg/L, respectively, reflect the water's buffering capacity and calcium/magnesium content, both of which are crucial for assessing the chemical stability of the water [25]. The mean concentrations of nitrate (NO_3^{1-}) and phosphate (PO_4^{3-}) are 0.021 mg/L and 1.52 mg/L, respectively, highlighting the presence of nutrient pollution, a significant concern in water quality management [50]. The Water Quality Index (WQI) has a mean value of 1.88, indicative of the overall quality of the water samples analyzed.

The correlation matrix presented in Fig. 6 is crucial for understanding the interrelationships among the water quality parameters. It enables the identification of functional dependencies, where strong correlations (r > 0.7) suggest significant associations, while weaker correlations (r < 0.4) imply more complex or indirect relationships. The WQI, the primary focus of this study, exhibits a strong positive correlation with phosphate levels (r = 0.99), underlining the significant impact of nutrient concentrations on overall water quality. In contrast, WQI shows weak correlations with parameters such as EC, TDS, $CaCO_3$, and NO_3^{1-}, suggesting that these variables, while influential, do not directly drive the WQI in this context.

The detailed examination of the correlations reveals that pH is moderately correlated with total hardness (TH), $CaCO_3$, and temperature (T°), with respective correlation coefficients of 0.59, 0.54, and 0.6. These findings are consistent with previous studies that have observed similar patterns in groundwater quality assessments [27]. Conductivity (EC) displays a strong positive correlation with T° (r = 0.73), $CaCO_3$ (r = 0.94), and TH (r = 0.88), indicating that these parameters are interdependent, likely due to their shared origin in mineral dissolution processes [14].

TABLE VII. DESCRIPTIVE STATISTICS OF THE FEATURES

Parameter	Count	Mean	Std Dev	Min	Q1	Median	Q3	Max
pH	75.00	7.790020	0.462099	6.75	7.4960	7.9340	8.1810	8.5510
Conductivity (µS/cm)	75.00	220.962667	197.740533	29.80	57.70	137.20	338.50	722.00
Dissolved Oxygen (mg/L)	75.00	7.4616	0.6773318	4.50	7.2250	7.60	7.8150	10.12
Total Dissolved Solids (mg/L)	75.00	164.605333	207.768171	6.00	38.65	95.90	216.50	1400.00
Temperature (ºC)	75.00	21.52	5.052053	15.10	16.80	20.60	20.65	32.80
Alkalinity (mg/L)	75.00	114.84	86.908159	20.00	40.00	82.00	191.00	354.00
Hardness (mg/L)	75.00	136.5720	109.56334	34.20	39.90	102.60	225.15	427.50
Nitrate (mg/L)	75.00	0.021366	0.030259	0.000354	0.006854	0.010302	0.0240	0.218747
Phosphate (mg/L)	75.00	1.517073	1.220099	0.3215	0.862250	1.12130	1.6790	7.850
WQI	75.00	1.878745	1.183223	0.582148	1.238313	1.614984	2.089291	7.89930

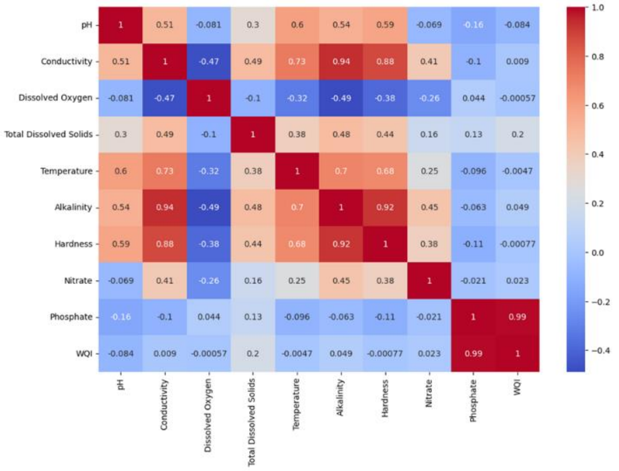


Fig. 6. Heatmap visualization of the future correlations.

Dissolved Oxygen (DO) exhibits a negative correlation with several parameters, notably EC (r = -0.47) and $CaCO_3$ (r = -0.49), which may suggest that higher ionic content and carbonate hardness could suppress oxygen solubility, a phenomenon that has been documented in other hydrographic contexts [50]. The analysis of TDS reveals moderate to strong positive correlations with pH (r = 0.3), EC (r = 0.49), and temperature (r = 0.38), reflecting the influence of these factors on dissolved solid concentrations [51]. The temperature itself strongly correlates with EC (r = 0.73) and $CaCO_3$ (r = 0.7), further reinforcing the interdependence of these water quality metrics.

The scatter plot matrix shown in Fig. 8 and the heatmap visualization provide additional insights into these relationships, offering a visual representation of the strength and direction of correlations. These graphical tools are essential for identifying patterns and anomalies in the dataset, facilitating a more nuanced interpretation of the results. The distribution of water compounds, as depicted in Fig. 7, confirms the trends observed in the correlation matrix, support, supporting the conclusion that physicochemical parameters, particularly nutrient concentrations, are critical determinants of water quality in the Amoju Hydrographic Subbasin.

The application of an Artificial Neural Network (ANN) model to predict the Water Quality Index (WQI) yielded robust results, demonstrating strong predictive performance across several key metrics, as shown in Table VIII. The model's Mean Absolute Error (MAE) of 0.2478 indicates a high degree of accuracy, with predicted WQI values deviating minimally from actual measurements. This level of accuracy is consistent with previous studies employing machine learning techniques for water quality prediction, further validating the model's effectiveness [44].

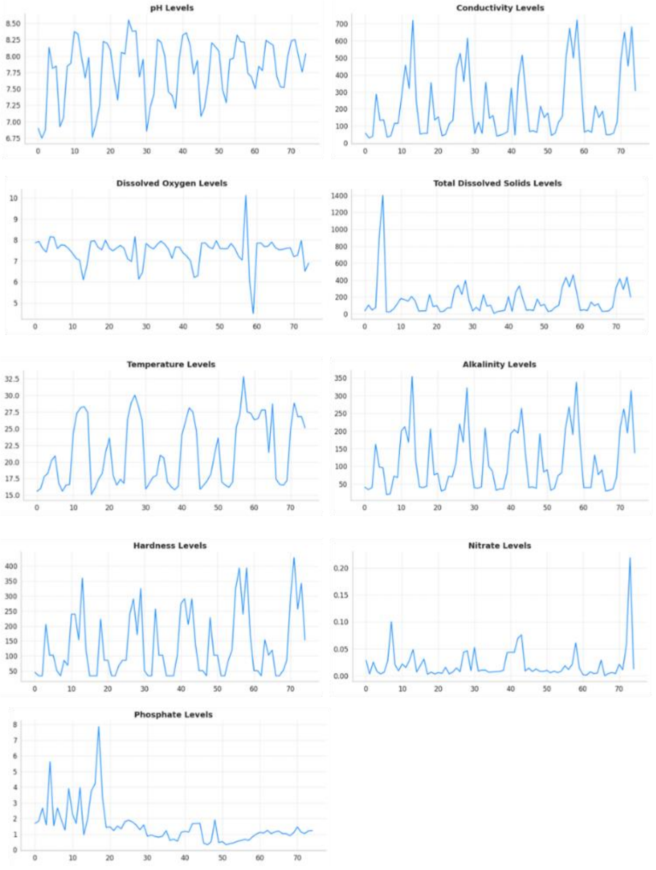


Fig. 7. Distribution of water compounds from the dataset.

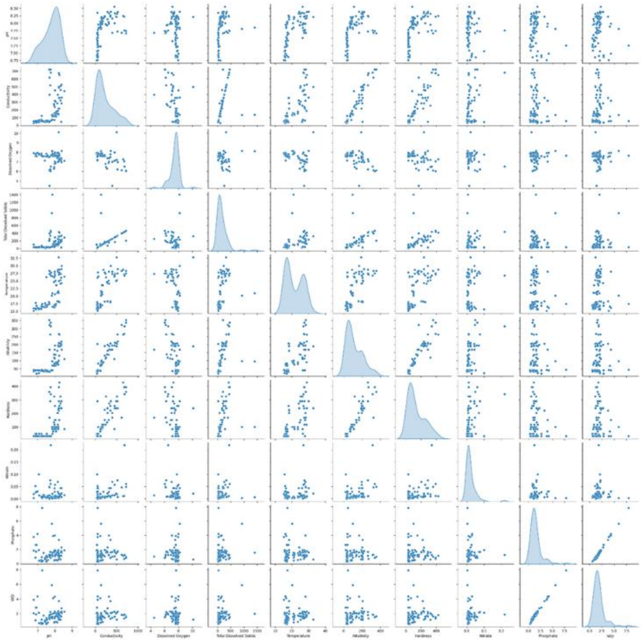


Fig. 8. Scatter plot matrix of the feature.

TABLE VIII. RESULTS BY ANN FOR WQI PREDICTION

Parameters	ANN
MAE	0.2478
MSE	0.0962
RMSE	0.3102
R2	0.9518

The model's Mean Squared Error (MSE) of 0.0962 and Root Mean Squared Error (RMSE), calculated as 0.3102 underscore the model's ability to minimize prediction errors, with the RMSE providing a direct measure of prediction in the same as the WQI. The model's predictive strength is further validated by the R-squared (R^2) value of 0.9518. The R^2 score suggests that approximately 95.18% of the variance in WQI can be explained by the model, highlighting its robustness and reliability as a predictive tool [23].

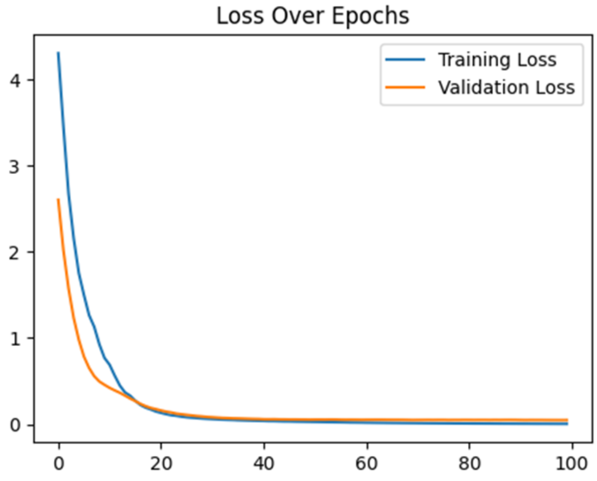


Fig. 9. ANN loss graph per epoch.

The convergence of the training and validation loss curves over 100 epochs, as illustrated in Fig. 9, suggest that the model is well-calibrated, with minimal risk of overfitting. This further corroborated by the “Actual vs. Predicted WQI” scatter plot shown in Fig. 10, which provides a visual comparison between the actual WQI values and those predicted by the model. The points on the plot are closely aligned along the red dashed line, which represents a perfect prediction. This close alignment reinforces the model's accuracy and the minimal prediction error. The plot confirms that the ANN model produces reliable predictions with a high degree of accuracy.

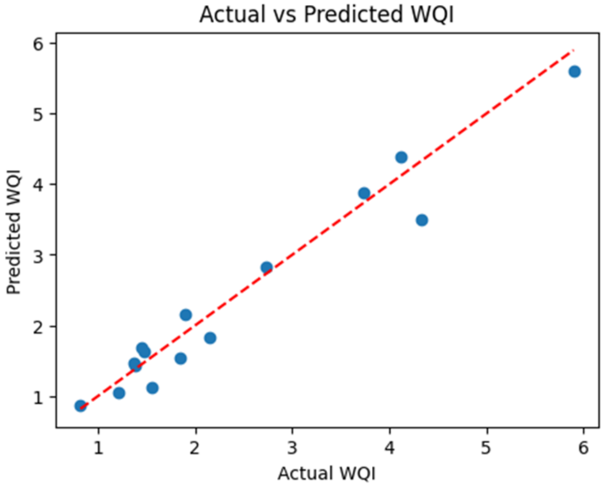


Fig. 10. Actual vs. predicted WQI graph.

V. CONCLUSION AND FUTURE WORK

This section presents a summary of the research conclusions and offers recommendations for future directions, highlighting key findings, limitations, implications, and suggests potential areas for further investigation.

A. Conclusion

Water is an essential resource for life on Earth, and ensuring its quality is fundamental to maintaining human health and environmental sustainability. The assessment of water quality, traditionally carried out through standard laboratory methods, has significantly advanced with the integration of machine learning (ML) techniques. These ML-based approaches offer a more robust and accurate means of predicting water quality indices by leveraging a wide array of water quality parameters.

This study has highlighted the relationships between various physicochemical parameters and the Water Quality Index (WQI) through a comprehensive analysis using Pearson's correlation matrix and Artificial Neural Network (ANN) model. The following key conclusions are drawn:

1) Correlation insights:

a) Very weak correlation: The WQI exhibits very weak correlations with parameters such as Electrical Conductivity (EC), Total Dissolved Solids (TDS), calcium carbonate ($CaCO_3$), and nitrate (NO_3^{1-}). These weak correlations suggest that these parameters, while part of the overall water quality assessment, have limited direct influence on the WQI in this specific study context.

b) Strong correlation: A strong positive correlation is observed between WQI and phosphate (PO_4^{3-}), indicating that phosphate levels are critical determinant of water quality in the studied hydrographic subbasin.

c) Moderate correlations: The pH shows moderate correlation with Total Hardness (TH), $CaCO_3$, and Temperature (T°), implying a notable, though not dominant, role of these parameters in influencing the water quality. EC also displays a moderate correlation with pH, TDS, and NO_3^{1-}.

d) Negative correlation: EC is negatively correlated with Dissolved Oxygen (DO) and phosphate (PO_4^{3-}), suggesting inverse relationships that may have implications for aquatic life and overall water chemistry.

2) Model performance: The ANN model developed for predicting the WQI demonstrated high predictive accuracy, validated by a low Mean Absolute Error (MAE), Mean Squared Error (MSE), and a high R-squared (R^2) value. These metrics confirm the model's ability to reliably predict water quality, making it a valuable tool for environmental monitoring and management.

3) Comparative analysis: The findings align with existing literature, reinforcing the importance of certain key parameters, particularly phosphate, in water quality assessments. The study not only corroborates the conclusions of previous research but also expands upon them by providing a nuanced understanding of parameter interrelationships within this specific geographical and environmental context.

B. Future Work

While the current study has provided significant insights into water quality prediction using machine learning, several avenues for future research remain:

1) Incorporation of additional parameters: Future studies should focus on incorporating additional chemical and biological parameters, such as heavy metals and microbial indicators, which might provide a more comprehensive water quality assessment.

2) Longitudinal data analysis: Extending the temporal scope of data collection over longer periods could allow for the examination of seasonal and climate variations in water quality, thereby enhancing the robustness of the predictive model.

3) Enhanced model architectures: Further refinement of the ANN model, or the application of more advanced machine learning techniques such as ensemble methods, deep learning, or hybrid models, could potentially improve predictive performance and uncover more complex relationships between parameters.

4) Geoespatial analysis: Integrating geospatial analysis tools with machine learning could provide spatially explicit predictions of water quality, which would be invaluable for regional water management and policy-making.

5) Real-time monitoring and prediction: Developing real-time monitoring systems, coupled with AI-driven predictive models, could facilitate the timely detection of water quality anomalies, enabling swift remedial actions and thereby safeguarding public health.

By addressing these areas in future research, the predictive capabilities and practical applications of machine learning in water quality management can be significantly advanced, contributing to more effective and sustainable environmental stewardship.

Conflict of Interest

The authors declare that they have no conflict of interest. The manuscript has been reviewed and approved by all authors, who have no financial or personal relationships that could inappropriately bias or influence the content.

Acknowledgment

We gratefully acknowledge the Soil and Water Analysis Center of the National University of Jaen for their invaluable assistance with the water sample analyses. Our sincere appreciation extends to all the professionals whose expertise and insights significantly enriched the scope and rigor of this study.

References

[1] Kar, Devashish, Wetlands and Lakes of the World. New Delhi: Springer India, 201d. C. Accedido: 5 de julio de 2024. [En línea]. Disponible en: https://library.wur.nl/WebQuery/titel/2074454.

[2] D. Kar, «Wetlands and their Fish Diversity in Assam (India)», Transylvanian Review of Systematical and Ecological Research, vol. 21, n.o 3, pp. 47-94, dic. 2019, doi: 10.2478/trser-2019-0019.

[3] N. Adimalla, «Groundwater Quality for Drinking and Irrigation Purposes and Potential Health Risks Assessment: A Case Study from Semi-Arid Region of South India», Expo Health, vol. 11, n.o 2, pp. 109-123, jun. 2019, doi: 10.1007/s12403-018-0288-8.

[4] S. Gaikwad, S. Gaikwad, D. Meshram, V. Wagh, A. Kandekar, y A. Kadam, «Geochemical mobility of ions in groundwater from the tropical western coast of Maharashtra, India: implication to groundwater quality», Environ Dev Sustain, vol. 22, n.o 3, pp. 2591-2624, mar. 2020, doi: 10.1007/s10668-019-00312-9.

[5] B. Dzwairo, Z. Hoko, D. Love, y E. Guzha, «Assessment of the impacts of pit latrines on groundwater quality in rural areas: A case study from Marondera district, Zimbabwe», Physics and Chemistry of the Earth, Parts A/B/C, vol. 31, n.o 15-16, pp. 779-788, ene. 2006, doi: 10.1016/j.pce.2006.08.031.

[6] T. A. Sinshaw, C. Q. Surbeck, H. Yasarer, y Y. Najjar, «Artificial Neural Network for Prediction of Total Nitrogen and Phosphorus in US Lakes», J. Environ. Eng., vol. 145, n.o 6, p. 04019032, jun. 2019, doi: 10.1061/(ASCE)EE.1943-7870.0001528.

[7] R. Barzegar y A. Asghari Moghaddam, «Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction», Model. Earth Syst. Environ., vol. 2, n.o 1, p. 26, mar. 2016, doi: 10.1007/s40808-015-0072-8.

[8] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, y A. Elshafie, «Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia», Neural Comput & Applic, vol. 28, n.o S1, pp. 893-905, dic. 2017, doi: 10.1007/s00521-016-2404-7.

[9] L. Xu y S. Liu, «Study of short-term water quality prediction model based on wavelet neural network», Mathematical and Computer Modelling, vol. 58, n.o 3-4, pp. 807-813, ago. 2013, doi: 10.1016/j.mcm.2012.12.023.

[10] W. C. Leong, A. Bahadori, J. Zhang, y Z. Ahmad, «Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)», International Journal of River Basin Management, vol. 19, n.o 2, pp. 149-156, abr. 2021, doi: 10.1080/15715124.2019.1628030.

[11] Y. Wu y S. Liu, «Modeling of land use and reservoir effects on nonpoint source pollution in a highly agricultural basin», J. Environ. Monit., vol. 14, n.o 9, p. 2350, 2012, doi: 10.1039/c2em30278k.

[12] A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, y R. N. Sankhua, «Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India», Model. Earth Syst. Environ., vol. 5, n.o 3, pp. 951-962, sep. 2019, doi: 10.1007/s40808-019-00581-3.

[13] P. Liu, J. Wang, A. K. Sangaiah, Y. Xie, y X. Yin, «Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment», Sustainability, vol. 11, n.o 7, p. 2058, abr. 2019, doi: 10.3390/su11072058.

[14] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, y M. Maashi, «Water Quality Prediction Using Artificial Intelligence Algorithms», Applied Bionics and Biomechanics, vol. 2020, pp. 1-12, dic. 2020, doi: 10.1155/2020/6659314.

[15] M. Hmoud Al-Adhaileh y F. Waselallah Alsaade, «Modelling and Prediction of Water Quality by Using Artificial Intelligence», Sustainability, vol. 13, n.o 8, p. 4259, abr. 2021, doi: 10.3390/su13084259.

[16] H. M. Mustafa, A. Mustapha, G. Hayder, y A. Salisu, «Applications of IoT and Artificial Intelligence in Water Quality Monitoring and Prediction: A Review», 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 968-975, ene. 2021, doi: 10.1109/ICICT50816.2021.9358675.

[17] S. Palabıyık y T. Akkan, «Evaluation of water quality based on artificial intelligence: performance of multilayer perceptron neural networks and multiple linear regression versus water quality indexes», Environ Dev Sustain, jun. 2024, doi: 10.1007/s10668-024-05075-6.

[18] M. Rezaie-Balf et al., «Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach», Journal of Cleaner Production, vol. 271, p. 122576, oct. 2020, doi: 10.1016/j.jclepro.2020.122576.

[19] T. Xu, G. Coco, y M. Neale, «A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning», Water Research, vol. 177, p. 115788, jun. 2020, doi: 10.1016/j.watres.2020.115788.

[20] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, y S. Kumar, «Prediction of groundwater quality using efficient machine learning technique», Chemosphere, vol. 276, p. 130265, ago. 2021, doi: 10.1016/j.chemosphere.2021.130265.

[21] S. V. Moghadam, A. Sharafati, H. Feizi, S. M. S. Marjaie, S. B. H. S. Asadollah, y D. Motta, «An efficient strategy for predicting river dissolved oxygen concentration: application of deep recurrent neural network model», Environ Monit Assess, vol. 193, n.o 12, p. 798, dic. 2021, doi: 10.1007/s10661-021-09586-x.

[22] J. Wu y Z. Wang, «A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory», Water, vol. 14, n.o 4, p. 610, feb. 2022, doi: 10.3390/w14040610.

[23] F. Rustam et al., «An Artificial Neural Network Model for Water Quality and Water Consumption Prediction», Water, vol. 14, n.o 21, p. 3359, oct. 2022, doi: 10.3390/w14213359.

[24] S. Dilmi y M. Ladjal, «A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques», Chemometrics and Intelligent Laboratory Systems, vol. 214, p. 104329, jul. 2021, doi: 10.1016/j.chemolab.2021.104329.

[25] Md. M. Hassan et al., «Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms»:, HCIS, vol. 1, n.o 3-4, p. 86, 2021, doi: 10.2991/hcis.k.211203.001.

[26] M. I. Khoirul Haq, F. Dwi Ramadhan, F. Az-Zahra, L. Kurniawati, y A. Helen, «Classification of Water Potability Using Machine Learning Algorithms», en 2021 International Conference on Artificial Intelligence and Big Data Analytics, oct. 2021, pp. 1-5. doi: 10.1109/ICAIBDA53487.2021.9689727.

[27] S. Kouadri, A. Elbeltagi, A. R. Md. T. Islam, y S. Kateb, «Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)», Appl Water Sci, vol. 11, n.o 12, p. 190, dic. 2021, doi: 10.1007/s13201-021-01528-9.

[28] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, y J. García-Nieto, «Efficient Water Quality Prediction Using Supervised Machine Learning», Water, vol. 11, n.o 11, p. 2210, oct. 2019, doi: 10.3390/w11112210.

[29] Z. Ahmad, N. A. Rahim, A. Bahadori, y J. Zhang, «Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks», International Journal of River Basin Management, vol. 15, n.o 1, pp. 79-87, ene. 2017, doi: 10.1080/15715124.2016.1256297.

[30] M. Sakizadeh, «Artificial intelligence for the prediction of water quality index in groundwater systems», Model. Earth Syst. Environ., vol. 2, n.o 1, p. 8, mar. 2016, doi: 10.1007/s40808-015-0063-9.

[31] H. Zare Abyaneh, «Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters», J Environ Health Sci Engineer, vol. 12, n.o 1, p. 40, dic. 2014, doi: 10.1186/2052-336X-12-40.

[32] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, y M. F. Ramli, «Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors», Marine Pollution Bulletin, vol. 64, n.o 11, pp. 2409-2420, nov. 2012, doi: 10.1016/j.marpolbul.2012.08.005.

[33] L. Wang et al., «Improving the robustness of beach water quality modeling using an ensemble machine learning approach», Science of The Total Environment, vol. 765, p. 142760, abr. 2021, doi: 10.1016/j.scitotenv.2020.142760.

[34] C. Sillberg, P. Kullavanijaya, y O. Chavalparit, «Water Quality Classification by Integration of Attribute-Realization and Support Vector Machine for the Chao Phraya River», J. Ecol. Eng., vol. 22, n.o 9, pp. 70-86, oct. 2021, doi: 10.12911/22998993/141364.

[35] M. Yilma, Z. Kiflie, A. Windsperger, y N. Gessese, «Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia», Model. Earth Syst. Environ., vol. 4, n.o 1, pp. 175-187, abr. 2018, doi: 10.1007/s40808-018-0437-x.

[36] S. I. Abba et al., «Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index», Environ Sci Pollut Res, vol. 27, n.o 33, pp. 41524-41539, nov. 2020, doi: 10.1007/s11356-020-09689-x.

[37] A. Elbeltagi, C. B. Pande, S. Kouadri, y A. R. Md. T. Islam, «Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India», Environ Sci Pollut Res, vol. 29, n.o 12, pp. 17591-17605, mar. 2022, doi: 10.1007/s11356-021-17064-7.

[38] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, y N. Kazakis, «Improving prediction of water quality indices using novel hybrid machine-learning algorithms», Science of The Total Environment, vol. 721, p. 137612, jun. 2020, doi: 10.1016/j.scitotenv.2020.137612.

[39] A. Azad, H. Karami, S. Farzin, A. Saeedian, H. Kashi, y F. Sayyahi, «Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River)», KSCE J Civ Eng, vol. 22, n.o 7, pp. 2206-2213, jul. 2018, doi: 10.1007/s12205-017-1703-6.

[40] Y. Zhang et al., «Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network», Water Research, vol. 164, p. 114888, nov. 2019, doi: 10.1016/j.watres.2019.114888.

[41] S. B. H. S. Asadollah, A. Sharafati, D. Motta, y Z. M. Yaseen, «River water quality index prediction and uncertainty analysis: A comparative study of machine learning models», Journal of Environmental Chemical Engineering, vol. 9, n.o 1, p. 104599, feb. 2021, doi: 10.1016/j.jece.2020.104599.

[42] S. I. Khan, N. Islam, J. Uddin, S. Islam, y M. K. Nasir, «Water quality prediction and classification based on principal component regression and gradient boosting classifier approach», Journal of King Saud University - Computer and Information Sciences, vol. 34, n.o 8, pp. 4773-4781, sep. 2022, doi: 10.1016/j.jksuci.2021.06.003.

[43] D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, y N. T. D. Thuy, «Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam», Water, vol. 14, n.o 10, p. 1552, may 2022, doi: 10.3390/w14101552.

[44] M. Y. Shams, A. M. Elshewey, E.-S. M. El-kenawy, A. Ibrahim, F. M. Talaat, y Z. Tarek, «Water quality prediction using machine learning models based on grid search method», Multimed Tools Appl, sep. 2023, doi: 10.1007/s11042-023-16737-4.

[45] Autoridad Nacional del Agua, «Observatorio Nacional de Recursos Hídricos». Accedido: 9 de agosto de 2024. [En línea]. Disponible en: https://snirh.ana.gob.pe/onrh/MapaTematicoUH.aspx.

[46] NASA JPL, «NASADEM Merged DEM Global 1 arc second V001». NASA EOSDIS Land Processes Distributed Active Archive Center, 2020. doi: 10.5067/MEASURES/NASADEM/NASADEM_HGT.001.

[47] S. Varol y A. Davraz, «Evaluation of the groundwater quality with WQI (Water Quality Index) and multivariate analysis: a case study of the Tefenni plain (Burdur/Turkey)», Environ Earth Sci, vol. 73, n.o 4, pp. 1725-1744, feb. 2015, doi: 10.1007/s12665-014-3531-z.

[48] Instituto Geológico, Minero y Metalúrgico, «Mapa geológico del Perú», Repositorio Institucional INGEMMET, mar. 2023, Accedido: 6 de julio de 2024. [En línea]. Disponible en: https://repositorio.ingemmet.gob.pe/handle/20.500.12544/3837.

[49] [A. A. Al-Othman, «Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses», Arabian Journal of Chemistry, vol. 12, n.o 8, pp. 2104-2110, dic. 2019, doi: 10.1016/j.arabjc.2015.01.001.

[50] V. B. B Patil, S. M. Pinto, T. Govindaraju, V. S. Hebbalu, V. Bhat, y L. N. Kannanur, «Multivariate statistics and water quality index (WQI) approach for geochemical assessment of groundwater quality—a case study of Kanavi Halla Sub-Basin, Belagavi, India», Environ Geochem Health, vol. 42, n.o 9, pp. 2667-2684, sep. 2020, doi: 10.1007/s10653-019-00500-6.

[51] A. R. Md. T. Islam, N. Ahmed, Md. Bodrud-Doza, y R. Chu, «Characterizing groundwater quality ranks for drinking purposes in Sylhet district, Bangladesh, using entropy method, spatial autocorrelation index, and geostatistics», Environ Sci Pollut Res, vol. 24, n.o 34, pp. 26350-26374, dic. 2017, doi: 10.1007/s11356-017-0254-1.

[52] A. R. Md. T. Islam, M. T. Siddiqua, A. Zahid, S. S. Tasnim, y M. M. Rahman, «Drinking appraisal of coastal groundwater in Bangladesh: An approach of multi-hazards towards water security and health safety», Chemosphere, vol. 255, p. 126933, sep. 2020, doi: 10.1016/j.chemosphere.2020.126933.

[53] K. P. Singh, N. Basant, y S. Gupta, «Support vector machines in water quality management», Analytica Chimica Acta, vol. 703, n.o 2, pp. 152-162, oct. 2011, doi: 10.1016/j.aca.2011.07.027.

Interactive ChatBot for PDF Content Conversation Using an LLM Language Model

LLM-Based PDF ChatBot

Ting Tin Tin[1*], Seow Yu Xuan[2], Wong Man Ee[3], Lee Kuok Tiung[4]*, Ali Aitizaz[5]
Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia[1]
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2, 3]
Faculty of Social Science and Humanities, Universiti Malaysia Sabah, Sabah, Malaysia[4]
School of Technology, Asia Pacific University, Kuala Lumpur, Malaysia[5]

***Abstract*—Natural Language Processing (NLP) leverages Artificial Intelligence (AI) to enable computer programs to understand and generate human language. ChatGPT has recently become popular in assignment accomplishment. This project aims to develop and improve an interactive PDF chat application using OpenAI's language model (LLM), specifically GPT-3.5, integrated with Streamlit and LangChain frameworks to assist in learning process. The application enhances user interaction with documents by providing real-time text extraction, summarization, translation, and user-defined question-answering to increase learning opportunities. Key features include obtaining document summaries, multilingual support for improved accessibility, and a document preview section with features such as zoom, rotation, and download. Although it currently faces limitations in handling image-rich PDFs, future enhancements include better image rendering, conversation history, and query download features. Overall, this interactive chatbot model aims to streamline document interaction, making information retrieval efficient and user-friendly.**

Keywords*—Natural language processing; learning opportunities; ChatGPT; PDF

I. Introduction

Natural language processing (NLP) is a study that implements artificial intelligence (AI) technology to enable computer programs to understand human language. NLP consists of two categories: natural language understanding (NLU), which refers to the process of reading and interpreting natural language, and natural language generation (NLG), which refers to the process of writing and generating natural language. With these two categories, NLP integrates deep learning models, machine learning, and computational linguistics to process human language. Before processing natural language, texts are preprocessed to clean and prepare data for classification. As texts are found to contain noise and uninformative pieces, text preprocessing plays an essential role before NLP to prevent interference in text analysis. Common text-preprocessing techniques are tokenization, normalization with stemming, lemmatization, and stop-word removal, and noise removal.

As technology advances, NLP has gained widespread adoption in many applications, including speech recognition (speech-to-text), which is the conversion of speech into text data, and sentiment analysis, which refers to the act of computationally recognizing and classifying viewpoints stated in a text, particularly to ascertain the writer's stance (positive, negative, or neutral) about a certain subject. ChatGPT is widely used especially among students with efficiency in completing assignments yet challenging in plagiarism and integrity [1], [2]. ChatGPT also provides assistants to educators in preparing course materials and exercises [3] Besides education, ChatGPT also provides various opportunities such as research, entertainment, code generation, explanation, and comments, test cases, regular expression, documentation generation, code correction, merging, conversion, and styling, metaverse learning environment [2], [4], [5].

In this project, the OpenAI language model (LLM) is used in creating an interactive chat application for conversations with documents analysis. The application consists of functionalities to extract text, answer user-defined questions, provide translation, and preview documents. We wish to achieve real-time clarification and explanation from the interactive document bot. It will act as a real-time support system where users can ask for instant clarifications or explanations for any issue they come across inside the text. This function speeds up learning or understanding by providing prompt answers to users' questions and removing the need for them to look for additional sources of information. Instant answers customized for each user's questions allow them to continue interacting with the material and ensure full comprehension of the uploaded document.

The proposed project is to develop an interactive PDF chat program with the following features: text extraction, text translation, user-defined prompt question answering, and document viewing. Users will be able to upload document files, start a discussion, ask questions regarding the document, and get answers in real-time through the program. When the file is uploaded and read, a preview of it will be displayed for the user's reference. The material can be summarized, extracted, and translated. A feature that distinguishes our application from the existing market is that it facilitates multilingual response, ensuring that it can deliver information in a different language from the one in which it was asked. This feature improves

*Corresponding Author.

The article is sponsored by INTI International University publication fund.

accessibility and comprehension by allowing users to receive replies in the language of their choice. The project will be constructed using Python, Streamlit, and OpenAI API, allowing users to access numerous document-related features and engage in conversational interactions with documents. By offering a fluid and engaging chat-based interface for document exploration and information retrieval, the program seeks to improve the user's experience when dealing with documents.

This study is constructed in the following sections. Section II presents and compares existing algorithms in NLP and LLM. This is followed by methodology design including system design in Section III. The system implementation with screenshots is presented in Section IV. Section V discusses the system features. Lastly, Section VI concludes the study with limitation and future works.

II. Literature Review

A. Background of NLP and LLM

Rapid improvement of technology, notably in natural language processing (NLP) and the development of sophisticated language models (LLMs), has created opportunities to revolutionize how people interact with textual content. Traditional document approaches, such as direct download of the PDF and reading it solely, often involve static reading, which requires external references for clarity or understanding. Furthermore, this study discusses the need for a more dynamic and user-centric approach by developing an interactive chat application that takes advantage of the capabilities of a language model, notably OpenAI's LLM.

The language model (LLM) that is implemented in our system is OpenAI. Based on research on OpenAI's LLMs, such as GPT-3 and ChatGPT, this LLM highly demonstrates its powerful abilities in natural language understanding (NLU), question answering, text summarization, and natural language generation (NLG). This empowers them to provide explanations and clear clarifications based on user-specific questions about their questions [6], [7].

Our application is to have a chatbot with an LLM basis that can help users inquire about PDF documents more conveniently. Research projects such as Chatbot as Language Learning Medium: An inquiry and an overview of Chatbot Technology show that interactive chat interfaces for exploring information within PDFs are achievable. These kinds of studies emphasize the importance of getting input on what the user wants to gain, the strategy of information retrieval from the user, and LLM response generation techniques for more effective conversational interaction.

Without the existence of artificial intelligence, understanding and extracting meaningful information from portable document format (PDF) files can be challenging due to its differences in layout, text-embedding formats, and the option of having scanned documents. Therefore, there are several studies published recently that aim to explore techniques for text pre-processing, optical character recognition (OCR), and information retrieval algorithms from multiple PDF formats. These studies are paving the way for a more accurate and precise information extraction for LLM interaction.

B. Comparison of LLM

According to the Table I comparison, the GPT 3.5 and Claude LLM models are the LLMs that are more suitable for chat with PDFs [8], [9]. As for GPT 3.5, it is highly adaptable and may be used for almost any task. It is reasonably priced. In addition, application in personal and professional tasks has already begun and is rapidly expanding. Alternatively, this Claude LLM model can be used when the user's prompts are in high volume, and the user wants to deliver the chat messages or prompts without utilizing workarounds. The prompt size offered is 100k tokens, which can fully accommodate roughly 75k words in a single prompt.

TABLE I. Comparison of Different Language Models Based on the Specifications, Pros and Cons

Model	GPT-3.5	Google PaLM [10]	Claude v1 [11]	Microsoft T5 [12]
Parameters	175 Billion	540 Billion	Unknown	11 Billion
API Availability	Yes	Yes	Yes	Pending
MMLU	70	-	75.6	47.7
MT-Bench Score	7.94	6.4	7.9	3.04
Tasks It Excels	Summarization Question Answering Text Generation	Summarization Question-Answering Text Generation Code Generation	Language-specific tasks	Custom Fine Tunes Translation Text Classification
Pros	Performs fairly well for a wide range of tasks. Custom fine-tunings are easy to create commercial use permitted	Performs very well for a wide range of tasks Code Generation Capabilities Easy API Integration Commercial use permitted	Largest token window support in one message (100k tokens, 75k words, approximately) Performs fairly well for wide range of tasks	Allows both supervised and unsupervised fine-tuning. Multiple models available based on use case
Cons	Biases in Output Generation moderately expensive	Risk of biased or inappropriate outputs	Fails to generate very human-like responses	Commercial use not permitted. Getting access is cumbersome

*Massive Multitask Language Understanding (MMLU) helps to evaluate the distortion between original and synthesized signal where it helps to qualify the performance and quality of synthesis systems.

*MT-bench score refers to the performance measure that reflects a model's ability to generate accurate and high-quality translations.

*In order to test out the LLMs above, https://poe.com/ (Poe) is able to provide a similar interface to the ChatGPT and in this website we are able to chat with different LLMs.

Based on the comparison of different LLMs above, we decided to integrate a GPT 3.5 (OpenAI) API as the language model in our PDF Chat model. This is due to the fact that GPT-3.5 can perform excellent data extraction from documents. It is good for text summarization, question answering, and text generation. Therefore, it is the perfect language model to use for a chatbot that is mainly a function for PDF summarisation. To further elaborate, when a user creates a query, the response given is perfectly structured information that has extracted the required text that the user queries from the PDF document. In addition, another advantage of GPT 3.5 is that it has a high level of customization, and therefore, the execution of complex workflows is available in our model. In summary, we decided to use GPT 3.5 to develop our interactive chatbot model that improves the user experience within PDFs. This study aims to increase the comprehension, accessibility and information for a wide range of users by combining real-time clarification, context-aware explanations, and NLP features such as translation and information extraction.

III. Methodology

A. Application of Algorithms / Libraries

1) Streamlit: Streamlit is an open-source Python library that allows for the creation of graphic user interfaces for data science and machine learning projects. Given that little code is required, it is best for those who lack the front-end skills necessary to integrate their code into a web application. Although no front-end knowledge such as HTML, CSS and JavaScript is needed, Streamlit enables the integration to allow design flexibility of the application. However, there are not many choices available in Streamlit to customize the look and feel of the application. If a great level of application customization is required, it is recommended to use more versatile web frameworks [13], [14].

2) LangChain: LangChain is an open-source framework dedicated to developing applications with language models (LLM). These models are deep learning architectures trained on large datasets, equipped to answer user inquiries and create graphics in response to text-based prompts. A collection of tools and abstractions offered by LangChain are available to enhance the adaptability, accuracy, and relevance of the data produced by these models. The core component of LangChain is the LLM interface, which gives developers access to APIs in which they can connect to and query LLMs from their code [15].

3) OpenAI API: An application programming interface (API) is a mechanism that provides a set of specifications and protocols by which two software components can interact with one another. This agreement outlines the requests and replies used in communication between the client and the server. With the OpenAI API key, users can access a range of powerful AI models and resources offered by OpenAI. The functions include natural language processing (NLP), text generation, and language translation.

B. System Flow Chart

Fig. 1 shows the flow of the PDF chatbot operation design. The figure illustrates the process of how our ChatApp interacts with PDF documents to provide responses. This system begins with the user uploading any number of documents that are then read and split into chunks of text. Each chunk of text is converted into embeddings, which are vector representations that capture the semantic meaning of the text. The knowledge base consists of these embeddings, which are kept in a vector store. In addition, a user's question is transformed into an embedding and compared via a semantic search with the stored embeddings. A large language model (LLM) ranks and processes the pertinent fragments that the search pulls from the knowledge base to produce the final response, which is then sent back to the user.

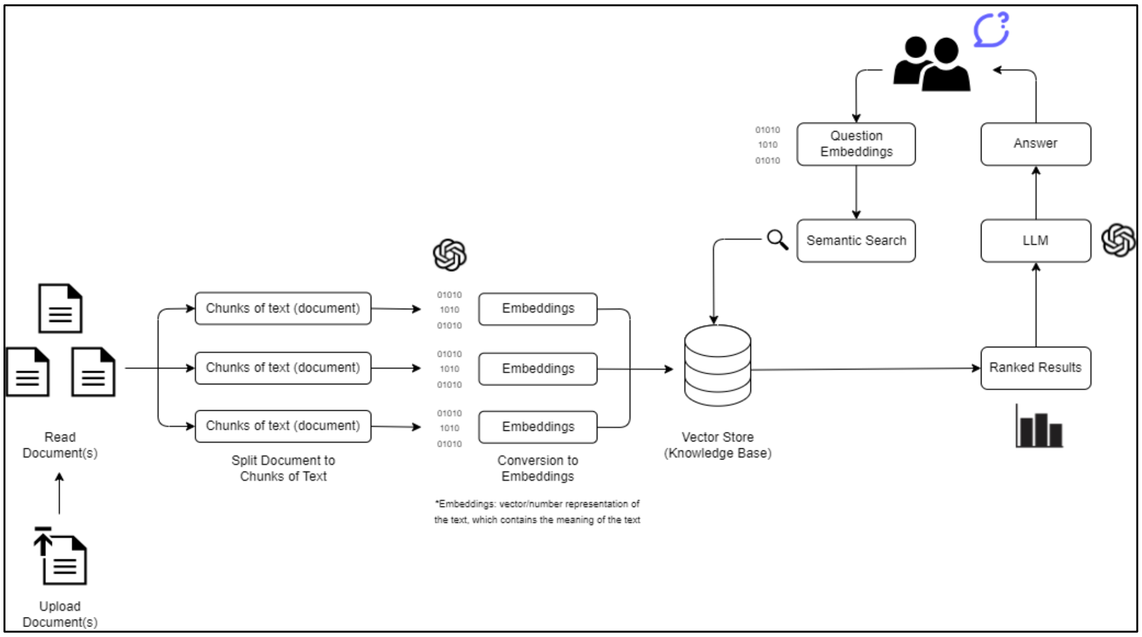

Question Embeddings
01010
1010
01010
Answer
Semantic Search
LLM
Chunks of text (document)
Chunks of text (document)
Chunks of text (document)
Embeddings
Embeddings
Embeddings
Ranked Results
Read Document(s)
Split Document to Chunks of Text
Conversion to Embeddings
Vector Store (Knowledge Base)
*Embeddings: vector/number representation of the text, which contains the meaning of the text
Upload Document(s)

Fig. 1. PDF chatbot operation design.

IV. Results and Discussion

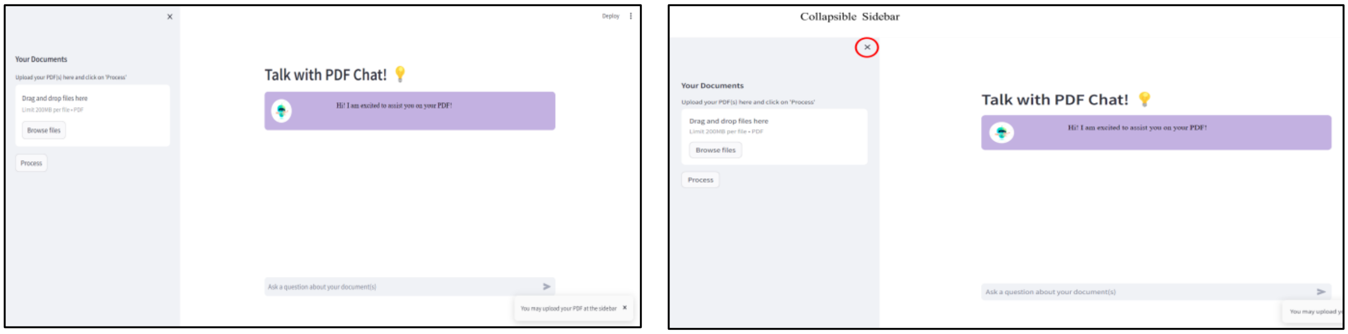
Fig. 2 illustrates the graphic user interface (GUI) when the application is first launched. On the left of the page is the file browser, where files can be uploaded and previewed; the right, the conversation with the PDF chatbot happens. A collapsible sidebar is equipped for file upload. When the cross (x) in the upper right corner is clicked, the section will collapse. When clicking the arrow (>) on the upper left corner of the page, the sidebar expands.

In Fig. 3, a file browser is provided in the sidebar for the user to upload PDF documents. The "Browse File" button turns red when it is clicked and a file window will pop up for the user to select the desired file(s). Once the document is uploaded, a throbber will appear with the text "Processing" after the "Process" button is selected. Lastly, a "PDF Upload Completed" message will appear upon successful upload. Fig. 4 illustrates the document preview section. There are additional features regarding the document, including the page count, zoom-in and zoom out, rotation of the page, downloading, and printing of the document, along with the document settings.

Fig. 5 shows a series of conversations initiated. The details are as follows:

1) A greeting to the bot
2) A general question about the document
3) A question that is not related to the document
4) A detailed question about the document

Fig. 6 shows the PDF Chatbot function that receives and responds to questions in different languages. The screenshot includes languages of English, Chinese, and Malay. Fig. 7 shows the extraction of text, in which the questions prompted are about a particular chapter of the document. In addition, the question requested the answer in a particular language (Chinese and Korean in the example), which the PDF Chatbot provided accurate answers. In addition to that, PDF Chatbot allows the upload and preview of more than one document. Fig. 8 shows the PDF chatbot function of answering questions about two documents.



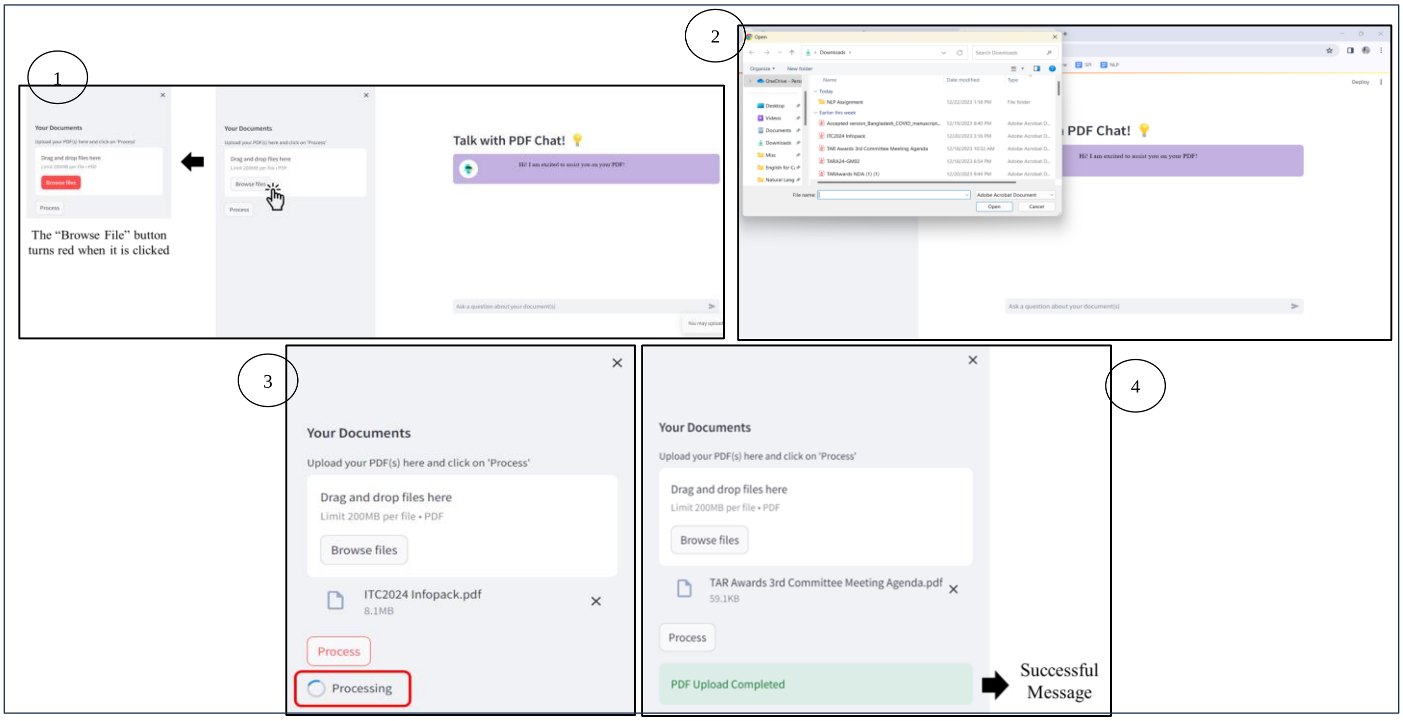
Collapsible Sidebar

Deploy
Your Documents
Upload your PDF(s) here and click on 'Process'
Drag and drop files here
Limit 200MB per file • PDF
Browse files
Process
Talk with PDF Chat!
Ask a question about your document(s)

Hi! I am excited to assist you on your PDF!

You may upload your PDF at the sidebar

Fig. 2. Graphic user interface of the PDF chatbot with collapsible sidebar.



1

The "Browse File" button turns red when it is clicked

2

3

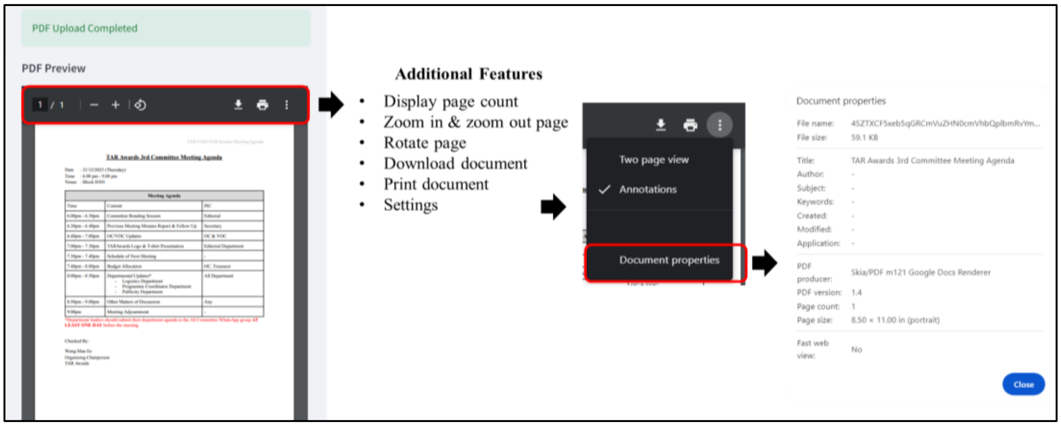
4

Successful Message

Hi! I am excited to assist you on your PDF!

PDF Upload Completed

Fig. 3. PDF document processing.



PDF Upload Completed

PDF Preview

TAR UMT/TAR Awards/Meeting Agenda

TAR Awards 3rd Committee Meeting Agenda

Date : 21/12/2023 (Thursday)
Time : 6.00 pm - 9.00 pm
Venue : Block H301

Meeting Agenda		
Time	Content	PIC
6.00pm - 6.30pm	Committee Bonding Session	Editorial
6.30pm - 6.40pm	Previous Meeting Minutes Report & Follow Up	Secretary
6.40pm - 7.00pm	OC/VOC Updates	OC & VOC
7.00pm - 7.30pm	TARAwards Logo & T-shirt Presentation	Editorial Department
7.30pm - 7.40pm	Schedule of Next Meeting	-
7.40pm - 8.00pm	Budget Allocation	OC, Treasurer
8.00pm - 8.30pm	Departmental Updates* - Logistics Department - Programme Coordinator Department - Publicity Department	All Department
8.30pm - 9.00pm	Other Matters of Discussion	Any
9.00pm	Meeting Adjournment	-

*Department leaders should submit their department agenda to the All Committee WhatsApp group AT LEAST ONE DAY before the meeting.

Checked By:

Wong Man Ee
Organising Chairperson
TAR Awards

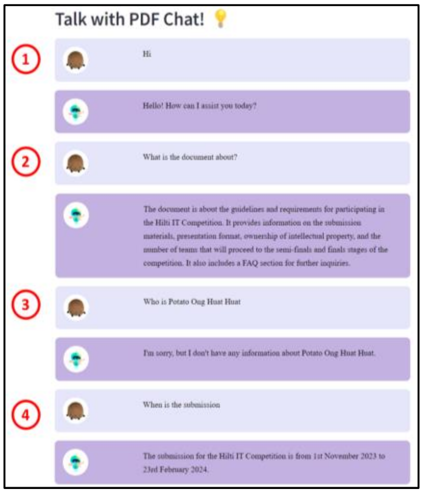
Additional Features

- Display page count
- Zoom in & zoom out page
- Rotate page
- Download document
- Print document
- Settings

Document properties

File name:	4SZTXCF5xeb5qGRCmVuZHN0cmVhbQplbmRvYm...
File size:	59.1 KB
Title:	TAR Awards 3rd Committee Meeting Agenda
Author:	-
Subject:	-
Keywords:	-
Created:	-
Modified:	-
Application:	-
PDF producer:	Skia/PDF m121 Google Docs Renderer
PDF version:	1.4
Page count:	1
Page size:	8.50 × 11.00 in (portrait)
Fast web view:	No

Fig. 4. PDF document review.



Talk with PDF Chat! 💡

1 Hi

Hello! How can I assist you today?

2 What is the document about?

The document is about the guidelines and requirements for participating in the Hilti IT Competition. It provides information on the submission materials, presentation format, ownership of intellectual property, and the number of teams that will proceed to the semi-finals and finals stages of the competition. It also includes a FAQ section for further inquiries.

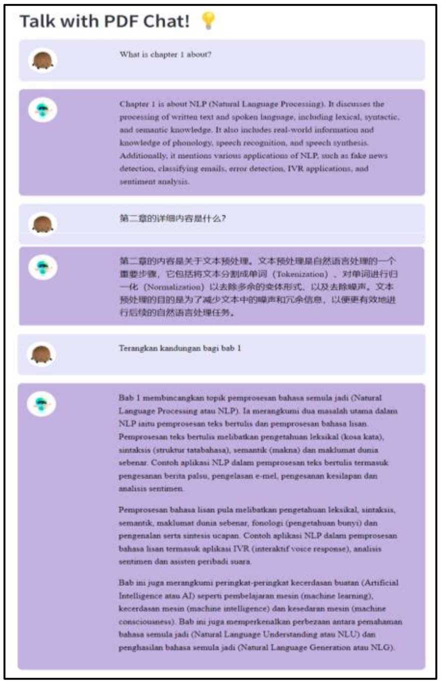
3 Who is Potato Oug Huat Huat

I'm sorry, but I don't have any information about Potato Oug Huat Huat.

4 When is the submission

The submission for the Hilti IT Competition is from 1st November 2023 to 23rd February 2024.

Fig. 5. Conversation with PDF chatbot.



Talk with PDF Chat! 💡

What is chapter 1 about?

Chapter 1 is about NLP (Natural Language Processing). It discusses the processing of written text and spoken language, including lexical, syntactic, and semantic knowledge. It also includes real-world information and knowledge of phonology, speech recognition, and speech synthesis. Additionally, it mentions various applications of NLP, such as fake news detection, classifying emails, error detection, IVR applications, and sentiment analysis.

第二章的详细内容是什么?

第二章的内容是关于文本预处理。文本预处理是自然语言处理的一个重要步骤，它包括将文本分割成单词（Tokenization）、对单词进行归一化（Normalization）以去除多余的变体形式、以及去除噪声。文本预处理的目的是为了减少文本中的噪声和冗余信息，以便更有效地进行后续的自然语言处理任务。

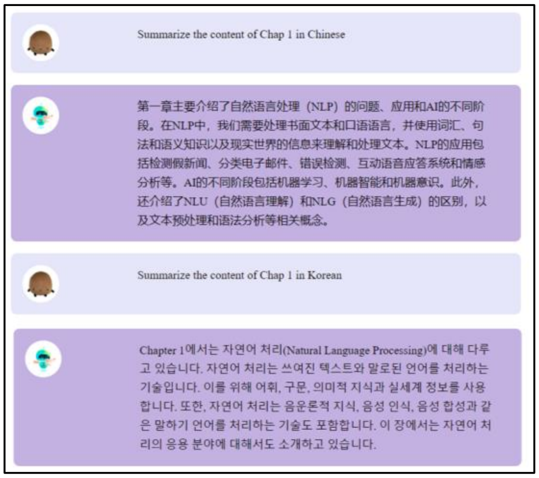
Terangkan kandungan bagi bab 1

Bab 1 membincangkan topik pemprosesan bahasa semula jadi (Natural Language Processing atau NLP). Ia merangkumi dua masalah utama dalam NLP iaitu pemprosesan teks bertulis dan pemprosesan bahasa lisan. Pemprosesan teks bertulis melibatkan pengetahuan leksikal (kosa kata), sintaksis (struktur tatabahasa), semantik (makna) dan maklumat dunia sebenar. Contoh aplikasi NLP dalam pemprosesan teks bertulis termasuk pengesanan berita palsu, pengelasan e-mel, pengesanan kesilapan dan analisis sentimen.

Pemprosesan bahasa lisan pula melibatkan pengetahuan leksikal, sintaksis, semantik, maklumat dunia sebenar, fonologi (pengetahuan bunyi) dan pengenalan serta sintesis ucapan. Contoh aplikasi NLP dalam pemprosesan bahasa lisan termasuk aplikasi IVR (interaktif voice response), analisis sentimen dan asisten peribadi suara.

Bab ini juga merangkumi peringkat-peringkat kecerdasan buatan (Artificial Intelligence atau AI) seperti pembelajaran mesin (machine learning), kecerdasan mesin (machine intelligence) dan kesedaran mesin (machine consciousness). Bab ini juga memperkenalkan perbezaan antara pemahaman bahasa semula jadi (Natural Language Understanding atau NLU) dan penghasilan bahasa semula jadi (Natural Language Generation atau NLG).

Fig. 6. Prompting queries and obtaining answers in different languages.



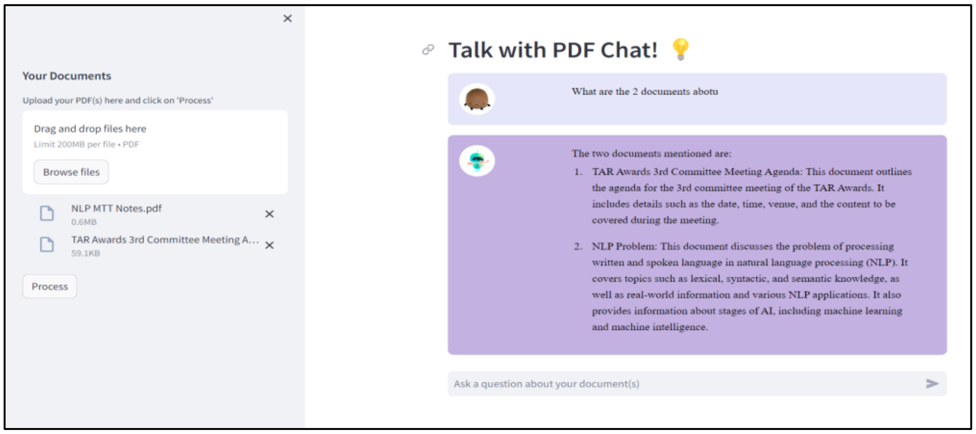
Summarize the content of Chap 1 in Chinese

第一章主要介绍了自然语言处理（NLP）的问题、应用和AI的不同阶段。在NLP中，我们需要处理书面文本和口语语言，并使用词汇、句法和语义知识以及现实世界的信息来理解和处理文本。NLP的应用包括检测假新闻、分类电子邮件、错误检测、互动语音应答系统和情感分析等。AI的不同阶段包括机器学习、机器智能和机器意识。此外，还介绍了NLU（自然语言理解）和NLG（自然语言生成）的区别，以及文本预处理和语法分析等相关概念。

Summarize the content of Chap 1 in Korean

Chapter 1에서는 자연어 처리(Natural Language Processing)에 대해 다루고 있습니다. 자연어 처리는 쓰여진 텍스트와 말로된 언어를 처리하는 기술입니다. 이를 위해 어휘, 구문, 의미적 지식과 실세계 정보를 사용합니다. 또한, 자연어 처리는 음운론적 지식, 음성 인식, 음성 합성과 같은 말하기 언어를 처리하는 기술도 포함합니다. 이 장에서는 자연어 처리의 응용 분야에 대해서도 소개하고 있습니다.

Fig. 7. Text extraction and multilingual response.



Your Documents
Upload your PDF(s) here and click on 'Process'
Drag and drop files here
Limit 200MB per file • PDF
Browse files
NLP MTT Notes.pdf
0.6MB
TAR Awards 3rd Committee Meeting A...
59.1KB
Process
Talk with PDF Chat! 💡
Ask a question about your document(s)

What are the 2 documents abotu

The two documents mentioned are:

1. TAR Awards 3rd Committee Meeting Agenda: This document outlines the agenda for the 3rd committee meeting of the TAR Awards. It includes details such as the date, time, venue, and the content to be covered during the meeting.

2. NLP Problem: This document discusses the problem of processing written and spoken language in natural language processing (NLP). It covers topics such as lexical, syntactic, and semantic knowledge, as well as real-world information and various NLP applications. It also provides information about stages of AI, including machine learning and machine intelligence.

Fig. 8. Conversation with PDF chatbot regarding 2 documents.

V. Discussion

According to the results mentioned above, we have successfully developed an interactive PDF Chatbot application specifically designed for document exploration and engagement. It is a technology stack that uses Python, Streamlit, and OpenAI API, enabling the easy integration of several conversational interactions and document-related features, guaranteeing a positive user experience. It can perform text extraction and summarization, allowing the user to obtain a summarised paragraph of the document. Additionally, the application offers a preview feature that displays the uploaded document, allowing users to refer to the context. Not only that, the chatbot is equipped with translation and multi-language support. It can enhance understanding and usability for a worldwide user base through translation and enabling users to communicate with the program in their choice language by handling inquiries and providing responses in languages different from the original document content. Lastly, a user-friendly interface is created for the application so that users can seamlessly navigate through the various functionalities offered.

VI. Conclusion

The limitation of the application is that some PDF files that contain images cannot be displayed in the preview section. To address this, we aim to explore alternative programming approaches to better handle and render image-rich PDFs within the preview section. To better enhance the application function, we intend to incorporate a conversation history function into the interactive PDF Chatbot so that users may review their prior questions. Additionally, we plan to provide a download feature, which would enable customers to store a copy of their queries on their devices. Besides that, future works could also involve a group of users from different specialization to test run the system and evaluate the accuracy and performance of the Chatbot system.

References

[1] F. Farhi, R. Jeljeli, I. Aburezeq, F. F. Dweikat, S. A. Al-shami, and R. Slamene, “Analyzing the students’ views, concerns, and perceived ethics about chat GPT usage,” Computers and Education: Artificial Intelligence, vol. 5, p. 100180, 2023, doi: 10.1016/j.caeai.2023.100180.

[2] I. Adeshola and A. P. Adepoju, “The opportunities and challenges of ChatGPT in education,” Interactive Learning Environments, pp. 1–14, Sep. 2023, doi: 10.1080/10494820.2023.2253858.

[3] C. K. Lo, “What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature,” Educ Sci (Basel), vol. 13, no. 4, p. 410, Apr. 2023, doi: 10.3390/educsci13040410.

[4] M. Al-Emran, “Unleashing the role of ChatGPT in Metaverse learning environments: opportunities, challenges, and future research agendas,” Interactive Learning Environments, pp. 1–10, Mar. 2024, doi: 10.1080/10494820.2024.2324326.

[5] G. Verma, T. Campbell, W. Melville, and B.-Y. Park, “Navigating Opportunities and Challenges of Artificial Intelligence: ChatGPT and Generative Models in Science Teacher Education,” J Sci Teacher Educ, vol. 34, no. 8, pp. 793–798, Nov. 2023, doi: 10.1080/1046560X.2023.2263251.

[6] Z. Xie, X. Evangelopoulos, Ö. H. Omar, A. Troisi, A. I. Cooper, and L. Chen, “Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules,” Chem Sci, vol. 15, no. 2, pp. 500–510, 2024, doi: 10.1039/D3SC04610A.

[7] J. M. Prieto Andreu and A. Labisa Palmeira, “Quick review of pedagogical experiences using GPT-3 in education,” J Technol Sci Educ, vol. 14, no. 2, p. 633, Mar. 2024, doi: 10.3926/jotse.2111.

[8] M. Enis and M. Hopkins, “From LLM to NMT: Advancing Low-Resource Machine Translation with Claude,” Apr. 2024.

[9] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, “Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach,” Intelligent Systems with Applications, vol. 21, p. 200336, Mar. 2024, doi: 10.1016/j.iswa.2024.200336.

[10] Google, “PaLM 2,” Google AI. Accessed: Jul. 15, 2024. [Online]. Available: https://ai.google/discover/palm2/

[11] Anthrop\c, “Introducing Claude.” Accessed: Jul. 15, 2024. [Online]. Available: https://www.anthropic.com/news/introducing-claude

[12] J. Ao et al., “SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing,” in ACL 2022, May 2022. [Online]. Available: https://www.microsoft.com/en-us/research/publication/speecht5-unified-modal-encoder-decoder-pre-training-for-spoken-language-processing/

[13] W. Lopes, “Streamlit Pro vs Cons,” Streamlit . Accessed: Jul. 15, 2024. [Online]. Available: https://www.linkedin.com/pulse/streamlit-pro-vs-cons-wendel-lopes/

[14] datacamp, “Python Tutorial: Streamlit.” Accessed: Jul. 15, 2024. [Online]. Available: https://www.datacamp.com/tutorial/streamlit

[15] AWS, “What is LangChain?,” AWS. Accessed: Jul. 15, 2024. [Online]. Available: https://aws.amazon.com/what-is/langchain/#:~:text=LangChain%20simplifies%20artificial%20intelligence%20

Prototype of an Indoor Pathfinding Application with Obstacle Detection for the Visually Impaired

Ken Gorro, Lawrence Roble, Mike Albert Magana, Rey Paolo Buot,
Louis Severino Romano, Herbert Cando, Bonifacio Amper, Rhyan Jay Signe, Elmo Ranolo
Department of Industrial Technology-College of Technology, Cebu Technological University, Carmen Campus, Philippines

***Abstract*—This study presents an initial prototype for a project aimed at assisting visually impaired individuals using deep learning techniques. The proposed system utilizes the You Only Look Once (YOLOv8) algorithm to detect objects tagged as obstacles. Designed for indoor environments, the system employs a CCTV camera and a computer server running the YOLOv8 model. Additionally, the A-star algorithm is used to determine the optimal path to avoid detected obstacles. Video frames are divided into tiles, each considered a node; nodes with detected objects are marked with a value of 0. The YOLOv8 model currently achieves an initial accuracy rate of 70%, with a mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 reaching 0.993 across all classes. This high mAP indicates an exceptional balance between precision and recall, signifying the model's effectiveness in object detection. Furthermore, the model yields an impressive F1-score of 0.99 at a confidence threshold of 0.624, demonstrating a robust balance between precision and recall, which is crucial for minimizing both false positives and false negatives. This prototype being developed assumes that a destination can be set by an operator of the system using the server that connects to the CCTV camera. The system was tested in enclosed environments and was able to provide a path that potentially avoids obstacles. The development of audio commands to guide visually impaired users is ongoing. These audio commands depend on identifying the direction an individual is going, requiring an additional deep-learning model to generate accurate instructions.**

***Keywords*—Yolov8; A-star algorithm; pathfinding; deep learning**

I. INTRODUCTION

Equal chances, particularly for the disabled, are vital in today's culture because technology is used by everyone and advancement is happening at a faster pace. For example, one of the groups most at risk of facing various challenges is the visually impaired. Finding our way about inside buildings may be quite difficult and dangerous, particularly if we are unfamiliar with the surroundings. More specifically, this significant topic has been taken into consideration in an effort to improve the independence and safety of visually impaired individuals through the use of modern technology, namely through the development of the in-door pathfinding program with the capability of obstacle detection.

The main goal of this study is to restore the independence and freedom of visually impaired individuals. The visually impaired are genuinely disabled in several ways, including but not limited to the fact that indoor spaces are primarily laden with barriers. They are supported by friends, family, or both, and they need traditional walking aids like canes and guide dogs. They can move around independently with the assistance of an improved navigational aid that can identify impediments, which will boost their efficiency and sense of self. Additionally, safety needs to be improved. In the midst of all of these considerations, improving safety is also crucial. It becomes clear that there are many risks and difficulties in any indoor setting, such as an office building, hospital, airport, or retail center. Occasionally, traditional mobility aids are unable to provide adequate or timely information about these risks. The application will be safe for visually impaired people to use if it can detect impediments in real-time, thereby reducing or eliminating the hazards related to falls and blind guiding.

This research is in phase 1 of creating a total pathfinding mobility application for the visually impaired in indoor scenarios. The result of this research only covers pathfinding and obstacle avoidance. The guiding voice is still in the development phase which will be the next phase of this study. The main focus is on the current obstacle detection and generating the best path to avoid obstacles using yolov8 and graph theory algorithms.

II. RELATED WORK

Despite global efforts in eye care, over 2.2 billion people suffer from vision impairment, with at least 1 billion cases preventable or unaddressed. The burden is higher in low-income countries, older populations, and disadvantaged communities. Initiatives like "Vision 2020" have contributed to improvements, but challenges remain, including limited service coverage, workforce shortages, and health system fragmentation. With aging populations and lifestyle changes, vision impairment is on the rise. The World Report on Vision advocates for Integrated People-Centered Eye Care (IPEC) to provide comprehensive, universal eye care and well-being [10].

Enhancing mobility and independence for visually impaired individuals relies heavily on advanced technologies. These innovations use artificial intelligence, machine learning, and robotics to navigate complex environments safely. By combining sensory data from cameras, LiDAR, and ultrasonic sensors, systems can identify and bypass obstacles instantly. Recent studies emphasize the development of reliable methods that enhance user experience and accessibility. This research is promising for creating assistive technologies that provide greater autonomy and safety for visually impaired individuals. To begin with, planning a path is an important part of any navigation framework. Since the 1970s, the issue of route planning has been a fascinating subject of study, particularly for robotics enthusiasts [2]. In order to save time, effort, and resources, it seeks to determine the best path between two sites. To guarantee user safety, additional design limitations must be added when incorporating these robotics-derived techniques into navigation systems. Furthermore, given that the user's tastes and wants may differ due to underlying path features and the topography of the environment, an optimal path for the VI is frequently not just the shortest path. In order to discover an optimal route while taking the preferences of the VI user into account and minimizing collisions with obstacles, this work presents a path-planning algorithm that is based on Ant Colony Optimization (ACO). Every stage has taken into account human issues, especially the unique requirements of the VI user [2]. People can manage and process spatial information about where they are physically by using cognitive mapping. In order to successfully complete tasks related to navigation and direction in urban environments, it is essential to develop a conceptual picture of the surrounding space and nearby objects (Fernandes et al., 2017) [1]. Because of obstructions on sidewalks and roadways, blind pedestrians frequently find it difficult or even dangerous to navigate in urban environments, despite their skill at compensating for lost visual information through increased awareness of environmental cues and navigational aids (Yang et al., 2011) [6]. Although blind pedestrians can navigate more safely and effectively thanks to audible traffic signals and landmarks (Weyrer et al., 2013), research on the needs, preferences, and experiences of blind pedestrians during navigation (e.g. Shangguan et al., 2014) and navigation solutions tailored specifically for them are still lacking [4][5]. The system described in the study by Lee and Medioni includes a glass-mounted RGBD camera with inertial measurement unit (IMU) sensors, a vest-type interface with four vibration motors, and a laptop that the user carries in the back and runs the program in real time. It uses visual odometry to estimate the camera location; it uses normal vectors and random sample consensus (RANSAC) to segment the floor; it generates a 2D traversable map and a 3D voxel map; and it directs the user to things of interest in dynamic interior environments. Using the D*Lite technique, an ideal path is calculated in the 2D map. The user receives haptic and audio input while using a smartphone application to select the location [7]. The Google Tango software, which operates on a hybrid phone/tablet Lenovo Phab2 with an integrated 3D depth sensor, gyroscope, and accelerometers, is the foundation of the CCNY smart cane. This device is placed on the user's chest. It can remember important visual elements like doors and rooms (stored in an area description file, or ADF), perform simultaneous localization and mapping (SLAM), and use infrared sensors to locate things. A visually impaired person can be guided to objects of interest in indoor spaces by the system, which uses the A* search algorithm to plot a path and provides both haptic (two vibrating motors installed on the white cane) and audible feedback. The iterative end-point fit algorithm is used to provide easier guidance, giving the user information [3][8][9]. D. Ni et al. [30] developed an assisted walking robot system that relied on computer vision and tactile sense. The system is made up of a rollator framework that offers stable physical support, a Kinect gadget that acts as a blind person's eyesight to record environmental information, and ultrasonic sensors that identify symmetry roads. A wearable vibrotactile belt is intended to provide blind people with information through vibration patterns. For the safe direction, a feature extractor approach based on depth image compression is used [15]. On the other hand, H-C Wang et al. (2021) advanced wearable devices for blind and visually impaired (BVI) individuals by integrating computer vision and haptic feedback to improve navigation and situational awareness. Their study showed how depth cameras and real-time processing help detect objects and obstacles, allowing safer navigation. The design is smaller and less intrusive, addressing previous issues with bulk and audio cues. Ongoing research aims to further develop this intuitive haptic technology for safer, independent mobility for BVI users [17].

According to a recent in-depth analysis of robot path planning in dynamic environments, reactive approaches such as Ant Colony Optimization (ACO) are quickly gaining traction in the field of mobile robot navigation because they perform better than classical approaches in real-time navigation scenarios (Cai et al., 2019) [11].

Kumar et al. (Kumar et al., 2020) recently improved the pheromone evaporation rate in order to address the issue of sluggish convergence of ACO. They suggested a fuzzified ACO (FACO) for mobile robots, in which the path pheromone updating process takes into account both favorable and unfavorable paths. Through simulation, the effectiveness of the FACO method was evaluated, demonstrating that it resolves the problem of delayed convergence [12].

Shiguo et al. (Li et al., 2020) tackled the problem of robot path planning by taking into consideration not only the path length but also the number of turns. They used the A* algorithm to enhance the convergence speed, and turning points were considered in the pheromone concentration. The simulation results assessed the performance of the improved ACO in terms of convergence, the number of iterations, path length, and the number of turns [13].

In order to address the problems of poor convergence speed, trapping into local optimum, and amount of turns, Ma and Mei (Ma and Mei, 2020) suggested an enhanced ACO path planning for mobile robots. To get rid of the less promising nodes, they employed a Jump Point Search (JPS) technique. Both simulation and real-world mobile robot navigation experiments were used to evaluate the algorithm's performance, demonstrating the enhanced algorithm's efficacy and efficiency in resolving path planning for mobile robots [14].

Broersen et al. (2016) developed a basic pathfinding method based on the A* algorithm. The path was calculated via the octree's vacant space. In the pathfinding procedure, two nodes that had the same face were regarded as neighbors. There might be equal-sized, bigger, or smaller neighbors. The neighbors are determined dynamically, and object avoidance is not used [16].

Tapu et al. developed a DEEP-SEE system that detects both dynamic and static items utilizing the You-Only-Look-Once (YOLO) object identification approach. The obstacle category and distance data may be obtained from the system. However, because of its high computing cost, real-time detection on mobile devices is challenging to achieve with this neural network [18]. Some lightweight image-based neural networks, such as ShuffleNet, YOLO-LITE, and MobileNet have suggested making real-time object identification more accessible to mobile devices. However, while recognizing objects, these lightweight algorithms do not explicitly take distance into account. Therefore, it is possible that the painted item on the ground is misleading to vision-challenged individuals [19][20][21].

Mortari et al. proposed a network creation technique that takes into consideration obstacles in indoor situations. Obstacles were represented as 2D geometry on the floor plane in the prepared models that served as the basis for the approach. Because 2D-floor layouts were abstracted at various height levels, the end product was a 3D network [22]. An approach for 3D indoor path-planning using semantic 3D models encoded in LoD4 CityGML was introduced by Xiong et al. [23]. While the system took barriers into account, tests were conducted on models without obstacles. A true interior navigation approach based on grid models—obtained from 2D-floor layouts with specified obstacles—was developed by Lui et al. [24]. An octree representation of indoor point clouds serves as the foundation for the indoor pathfinding approach that Rodenberg [25] suggested. Because the A* the pathfinding algorithm followed empty nodes and avoided obstructions, it was reliant on the use of heuristics to direct the search. Additionally, Li et al. [26] recently introduced a path-planning technique for drones operating inside that was based on occupancy voxel maps and on which the vacant voxels made up the navigable space.

Tsirpas et al. presented an intriguing Radio Frequency Identification (RFID)--based indoor navigation system for the elderly and visually handicapped. RFID tags have a limited range, which is the primary drawback of systems based on RFID technology (the authors of the research recommend 40 × 40 cm cells). Moreover, considering that the human body may resist radiofrequency signals, its range might be lowered. Another disadvantage is that installing RFID in large spaces may be costly because the tags frequently need to be inserted into the floors, walls, furniture, and other surfaces [27].

Khelifi et al. also studied the utilization of radio and Wi-Fi technologies. The writers went beyond specifics in their discussion of the subject, emphasizing the necessity for consideration of matters like processing costs, implementation costs, and energy efficiency. Grouping relevant works that supported the study according to methodologies and technologies helps the reader grasp the material better and become more objective when reading [28].

Mainetti et al. aimed to cover the key IPS technologies and approaches for locating, mapping, tracking, and navigating people, objects, and animals inside. The primary technologies employed in IPS, according to the authors, are computer vision (mobile and stationary cameras), infrared, ultrasound, Wi-Fi, RFID, and Bluetooth. The primary methods which include those that make use of FM radio and ZigBee technology are thoroughly explained, while the remaining methods are categorized under the heading "other technologies". The authors utilized a model that took accuracy, coverage, cost, complexity, and usual implementation locations into account while presenting the findings of the works that they used as references in the form of graphs and tables [29].

Guerrero proposed a micro-navigation system that tracks the user's position and movements with the use of an infrared camera. Using a tree structure of the room including the relevant data, based on Extensible Markup Language (XML), the system accomplishes static obstacle detection. However, the system's simulation and testing showed that as more people travel through the area, the system performs worse. PERCEPT is an interior navigation system that uses passive RFID tags for location and is based on kiosks. Using the shortest path method, it is a macro-navigation system that lets VI persons move from one room to another on multiple floors. Unfortunately, the technology that was described is unable to identify any environmental constraints that are nearby. The author also suggested that the system be enhanced by offering guidance information depending on user choices and steps [30][31].

Using RFID technology, the authors of Ivanov (2010) created an interior navigation system for the blind. By placing passive RFID tags above and below door handles, the technology makes door-to-door travel easier. This is based on the idea that blind people can identify doors with the use of their white canes. Navigational instructions are stored on tags above the door handles, while adaptive multi-rate (AMR) audio data is stored on tags below the handles. Every door serves as a point of reference for the following location. Each RFID tag has its navigation instructions manually entered into it. The blind user reads these tags with a cell phone that is equipped with RFID. A text-to-speech engine and the mobile phone's keyboard are used for bi-directional communication between the user and the device. The system mostly depends on the user's ability to navigate obstacles with their white canes. The published work did not address the correctness of the system; instead, it just examined the navigation time in its entirety [32].

III. Methodology and Results

Fig. 1 is the system architecture of the system, the CCTV camera streams the video data to a server with the deep learning models installed to detect obstacles within the video frames, and the A-star path-finding algorithm was utilized to generate the path. Each video frame is divided into tiles and treated as 1 node. The details of the algorithm are discussed in the succeeding session.

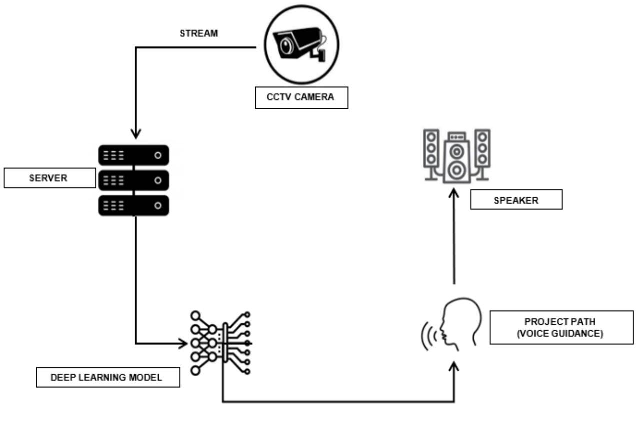


Fig. 1. System architecture.

A. Dataset

The dataset is essential for training YOLOv8 and other object detection models. It consists of images with annotations that include bounding boxes and class labels, providing the model with the information needed to learn object detection. During training, the dataset is divided into subsets: training, validation, and test. The training subset helps the model learn, the validation subset fine-tunes hyperparameters, and the test subset evaluates the final performance. High-quality and diverse datasets are crucial for effective training, ensuring accurate predictions and better model performance. 1200 samples were utilized to train and test the YoloV8 model for obstacle detection.

B. Bounding Boxes

Bounding boxes are essential for object detection models like YOLOv8, as they define the location and size of objects within images. During training, these annotations serve as ground truth, guiding the model to learn object positions and spatial relationships. The model predicts bounding boxes for objects, and its performance is evaluated by comparing these predictions to the ground truth. Discrepancies are used to adjust the model's parameters and reduce errors. After initial training, the model's bounding box predictions are further tested and refined to enhance accuracy. Properly annotated bounding boxes ensure effective object localization and classification, improving overall detection performance.

C. Hyperparameter Tuning for YOLOv8 Model

The training of the YOLOv8 model involves several crucial hyperparameters that significantly influence its performance. As detailed in Table I, the epochs parameter is set to 1000, meaning the model will traverse the entire training dataset 1000 times, which helps in enhancing the model's learning but requires careful monitoring to avoid overfitting. The batch size is configured to 16, indicating that the model will process 16 samples before updating its weights. This setting balances between frequent updates and memory usage, although it may affect the stability of the gradient estimates. The learning rate is set to 0.05, which controls the magnitude of weight adjustments during training. This relatively high learning rate can speed up convergence but may introduce instability if not carefully managed. To mitigate overfitting, weight decay is applied with a value of 0.0005, adding a penalty for large weights and promoting generalization. The image size is fixed at 640 x 640 pixels, determining the dimensions of the input images and impacting both detection accuracy and computational demands. Finally, the momentum is set at 0.937, which helps accelerate the gradient vectors in the correct direction, facilitating faster convergence and reducing oscillations. Together, these hyperparameters are finely tuned to balance learning efficiency and model robustness.

TABLE I. HYPER PARAMETERS

Hyper parameters	Tune value
Epochs	1000
Batch Size	16 (default)
Learning Rate	0.05
Weight Decay	0.0005 (default)
Image Size	640 x 640 (default)
Momentum	0.937 (default)

D. Model Training

Model training for YOLOv8 is essential for optimizing the model's ability to detect objects accurately within a specific dataset. Through training, the model's parameters are adjusted to improve detection precision and recall. This process allows the model to adapt to the unique characteristics of the data, enhancing its performance. For training purposes, 80% of the samples were utilized while 20% were used for evaluating the model. Key aspects of training include tuning hyperparameters like learning rate and batch size, which are critical for achieving the best results while preventing overfitting or underfitting. Regular evaluation using metrics such as mean Average Precision (mAP) helps in assessing the model's performance and making necessary improvements.

E. Performance Metrics for YOLOv8 Model Evaluation

A. Precision is the ratio of true positive detections to the total number of positive detections made by the model. It measures how many of the detected objects are actually correct.

$$\text{Precision} = \frac{number\ of\ True\ positives}{number\ of\ True\ positives + number\ of\ False\ positives} \quad (1)$$

B. Recall is the ratio of true positive detections to the total number of actual objects in the ground truth. It measures how well the model identifies all the relevant objects.

$$\text{Recall} = \frac{number\ of\ True\ positives}{number\ of\ True\ positives + number\ of\ False\ negatives} \quad (2)$$

C. The mean Average Precision (mAP) is the average of Average Precision (AP) scores across all classes. It summarizes the precision-recall curve, offering a single, comprehensive metric to evaluate the model's performance across different thresholds and providing a holistic view of accuracy and object detection capabilities. Higher mAP indicates better object detection performance across various classes and confidence levels.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \quad APi \quad (3)$$

D. The F1-score in the context of YOLOv8 refers to a metric used to evaluate the model's performance in terms of both precision and recall. It is a harmonic mean of precision and recall, providing a single score that balances these two metrics.

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

F. Pathfinding Algorithms

*1) A * Algorithm*: The A* (A-star) algorithm is a widely used pathfinding and graph traversal algorithm, especially notable for its efficiency and accuracy in finding the shortest path between two points. It combines the strengths of Dijkstra's algorithm and Greedy Best-First-Search. A* operates by maintaining a tree of paths originating from the start node, extending those paths one edge at a time until the goal node is reached. The algorithm uses a priority queue to explore nodes based on the following cost function:

$$f(n) = g(n) + h(n)$$

where:

- g(n) is the cost from the start node to the current node (n).
- h(n) is the heuristic estimate of the cost from (n) to the goal.

The heuristic function h(n) is crucial for the A* algorithm's performance. It must be admissible, meaning it never overestimates the true cost to reach the goal, ensuring the optimality of the path found. Common heuristic functions include the Euclidean distance and Manhattan distance, depending on the nature of the problem space.

2) Dijkstra's algorithm: Dijkstra's algorithm finds the shortest path between nodes in a graph. It starts at a chosen node, and then repeatedly selects the unvisited node with the smallest known distance from the start. It updates the distances to neighboring nodes and marks the current node as visited. This process continues until the destination is reached or all nodes are visited. It uses the formula:

$$d[v] = \min(d[v], d[u] + w(u,v))$$

where:

- d[v] is the distance to node v
- d[u] is the distance to node u
- w(u,v) is the weight of the edge from u to v

The algorithm iteratively updates these distances, starting from the source node, until it reaches the destination or visits all nodes. It always selects the unvisited node with the smallest known distance for the next iteration.

While both A* and Dijkstra's algorithms are effective for pathfinding, A* is often preferred in scenarios where efficiency and performance are critical due to the following reasons:

a) Heuristic guidance: A* uses a heuristic to guide its search, which can significantly reduce the number of nodes explored compared to Dijkstra's algorithm. Dijkstra's algorithm explores all possible paths, ensuring the shortest path but often at the cost of increased computation time, especially in large or complex graphs.

b) Efficiency: The heuristic function in A* allows the algorithm to focus on the most promising paths towards the goal. This focused search typically results in faster pathfinding, as fewer nodes are expanded compared to Dijkstra's algorithm, which treats all nodes with equal importance.

c) Optimality and completeness: A* is both optimal and complete, meaning it will always find the shortest path if one exists, provided the heuristic is admissible. Dijkstra's algorithm also guarantees the shortest path, but without the heuristic guidance, it often requires more extensive node exploration.

d) Flexibility: A* can be easily adjusted for different types of pathfinding problems by changing the heuristic function. This adaptability makes it suitable for various applications, from simple grid-based pathfinding to more complex navigation in dynamic environments.

e) Practicality: In many practical applications, such as robotics, video games, and geographical mapping, the A* algorithm's balance of optimality and performance makes it more practical than Dijkstra's algorithm. It provides quicker responses, which is crucial for real-time decision-making processes.

For this study, the A* algorithm would be chosen over Dijkstra's algorithm because of its heuristic-driven approach, which improves efficiency by minimizing the number of nodes investigated while ensuring the shortest path. This makes A* ideal for large-scale and sophisticated pathfinding applications in which performance and speed are important.

G. Pathfinding with YOLO Integration

As shown in Fig. 2, the provided flowchart illustrates how to use YOLO predictions to navigate and find the shortest path between two places. The process starts with an initialization stage, which prepares the system to capture and process frames. To navigate all of the instances, the input frame is divided into grid coordinates with 1s and 0s. Areas filled with 0s signify obstacles, whereas 1s indicate a clear pathway.

As illustrated in Fig. 3, the system processes the YOLO prediction results by tracking the center of each instance's bounding box. The system then identifies two distinct points within the scene, referred to as Point A and Point B, based on their respective (x, y) coordinates. Using these coordinates, the system calculates the shortest path between the two locations, employing pathfinding algorithms such as the A* algorithm. A feedback loop is incorporated, allowing the system to continuously capture new frames and update predictions, as well as pathfinding computations. This dynamic process enables the system to adapt to changes in the environment in real-time, ensuring precise navigation and efficient pathfinding.

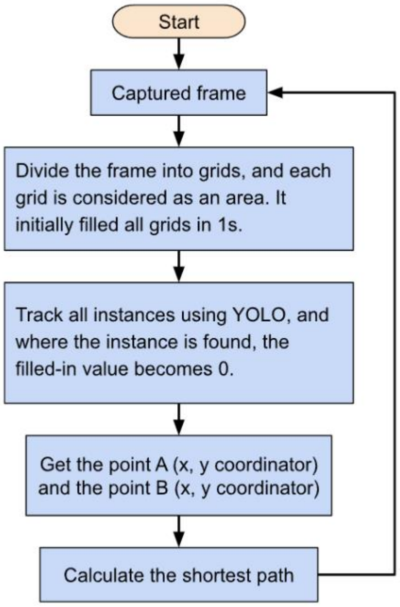

Start
Captured frame
Divide the frame into grids, and each grid is considered as an area. It initially filled all grids in 1s.
Track all instances using YOLO, and where the instance is found, the filled-in value becomes 0.
Get the point A (x, y coordinator) and the point B (x, y coordinator)
Calculate the shortest path

Fig. 2. Path-finding obstacle detection flowchart.

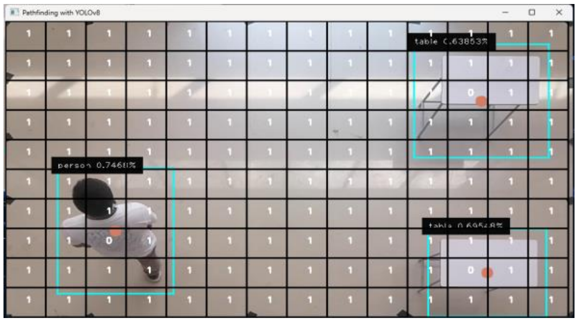


Fig. 3. Applied 1s and 0s on the captured frame.

IV. Experimental Result

A. Model Training Results

1) Confusion matrix: As illustrated in Fig. 4, the confusion matrix provides a comprehensive summary of the performance of the YOLOv8 small model across different classes. It is a powerful tool for understanding how well the model distinguishes between the classes: person, robot, table, and background.

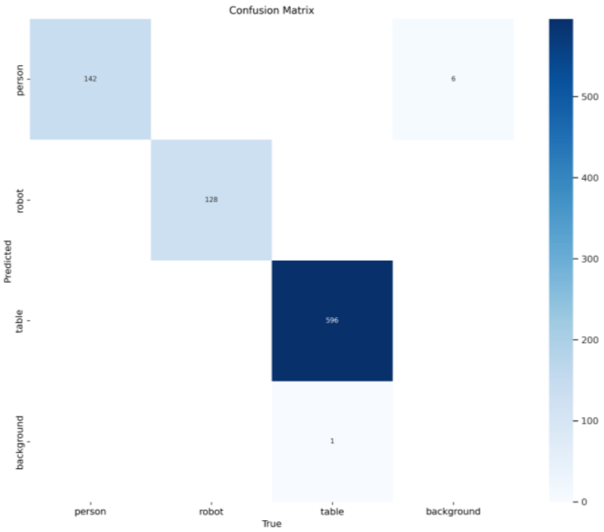


Fig. 4. Confusion matrix result.

In the confusion matrix, the "person" class has 142 true positives, indicating that the model correctly identified 142 instances of "person". However, there are six false positives where background instances were misclassified as "person." There are no false negatives, meaning all actual instances of "person" were correctly identified, showcasing the model's high accuracy for this class. For the "robot" class, the model performs perfectly, with 128 true positives and no false positives or false negatives. This indicates that the model consistently identifies "robot" instances without any errors, demonstrating its robustness in detecting this class. The "table" class also shows high accuracy with 596 true positives. There is only one false positive where a background instance was misclassified as "table," and no false negatives, meaning all actual instances of "table" were correctly identified. This indicates the model's high precision and recall for the "table" class. The confusion matrix also reveals that the model occasionally misclassified background instances. Specifically, six background instances were predicted as "person," and one background instance was predicted as "table." These misclassifications indicate room for improvement in distinguishing background from specific objects, which could enhance the model's overall performance.

In summary, the confusion matrix highlights that the YOLOv8 small model is highly effective at correctly identifying instances of "person," "robot," and "table." While the model demonstrates high precision and recall for these classes, there are occasional misclassifications of background instances, particularly as "person" and "table". These findings suggest that the model is robust and accurate but could benefit from further refinement to improve its handling of background instances. The confusion matrix serves as a valuable tool in validating the model's effectiveness and guiding future enhancements.

2) Training result graphs: Fig. 5 illustrates a visualization of several metrics and loss functions tracked during the YOLOv8 model's training and validation phases. Each subplot gives information about various aspects of the model's performance and learning process.

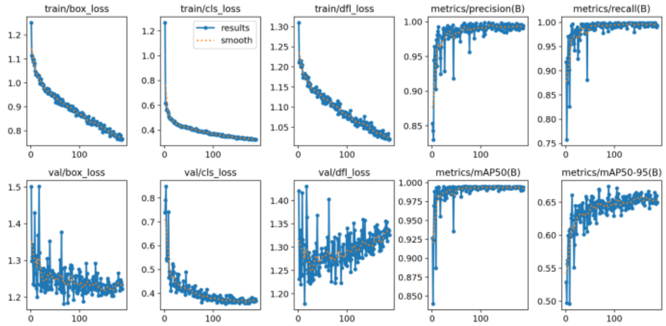


Fig. 5. YOLOv8 training results.

The training box loss (train/box_loss) plot illustrates the model's improvement in predicting bounding boxes. This loss consistently decreases over time, indicating that the model is learning to better match the predicted bounding boxes with the ground truth boxes. A similar trend is observed in the training classification loss (train/cls_loss) plot, where the decreasing loss shows that the model's classification capabilities are improving, allowing it to more accurately categorize objects into the correct classes (person, robot, table).

The training DFL loss (train/dfl_loss) which measures the distribution quality of the predicted bounding boxes, also shows a consistent decline. This suggests that the model is becoming more precise in predicting the exact locations of the bounding boxes. The precision metrics (metrics/precision (B)) and recall metrics (metrics/recall (B)) quickly rise and stabilize close to 1.0. This indicates that the model is highly accurate in its positive predictions, with very few false positives, and effective at detecting almost all relevant objects, with very few false negatives.

For the validation phase, the validation box loss (val/box_loss) decreases significantly, although with more fluctuation compared to the training loss. This suggests that while the model is learning well, it encounters some variability when applied to new data. The validation classification loss (val/cls_loss) shows a decreasing trend, indicating good generalization, though it also fluctuates, highlighting some variability in performance on the validation set. The validation DFL loss (val/dfl_loss), despite generally decreasing, exhibits higher fluctuation, indicating less stability in predicting bounding box distributions on unseen data compared to the training set.

The mAP50 metrics (metrics/mAP50 (B)), which measure precision and recall at a specific Intersection over Union (IoU) threshold of 50%, rapidly increase and stabilize near 1.0. This demonstrates that the model performs exceptionally well in detecting objects with a high degree of overlap with the ground truth. The mAP50-95 metrics (metrics/mAP50-95(B)), which provide a more comprehensive measure across varying IoU thresholds (from 50% to 95%), show an upward trend and eventual stabilization. This indicates that the model is robust across different IoU thresholds, despite more variability compared to mAP50.

3) *Model performance metrics*: The YOLOv8 small model, trained to detect three classes (person, robot, and table), exhibits impressive performance metrics. As shown in Table II, the precision for all classes is reported as 1.0 at a confidence threshold of 0.926, which means that every detection made by the model is correct at this high confidence level. This suggests that the model is highly reliable in its detections, making no false positives at this threshold.

TABLE II. PERFORMANCE METRIC RESULT OF YOLOV8 SMALL

YOLOV8 SMALL				
CLASSES	PRECISION	RECALL	Precision-RECALL (MAP@0.5)	F1-SCORE
person robot TABLE	ALL CLASSES 1.0 AT 0.926	ALL CLASSES 1.0 AT 0.000	ALL CLASSES 0.993	ALL CLASSES 0.99 AT 0.624

The recall for all classes is also 1.0 but at a confidence threshold of 0.000. Recall measures the model’s ability to identify all relevant instances in the dataset. Achieving a recall of 1.0 at the lowest confidence threshold means that the model captures all true positive instances, albeit at the expense of potentially including many false positives. This indicates the model’s maximum sensitivity, showing it can detect every instance of the specified classes when no confidence filtering is applied.

The mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 is 0.993 for all classes. The mAP@0.5 is a comprehensive metric that combines precision and recall across different confidence thresholds. A mAP of 0.993 signifies that the model performs exceptionally well, balancing precision and recall almost perfectly across a range of confidence levels.

The F1-score for all classes is 0.99 at a confidence threshold of 0.624. The F1-score, being the harmonic mean of precision and recall, provides a single metric that accounts for both false positives and false negatives. An F1-score of 0.99 indicates that the model maintains a high balance between precision and recall at this confidence level, making it highly effective for practical applications where both false positives and false negatives need to be minimized.

The perfect precision and recall values demonstrate the model's outstanding capability to correctly identify and detect all instances of the specified classes under certain conditions. However, achieving 1.0 recall at a threshold of 0.000 means the model includes all possible detections, leading to capturing every true positive along with many false positives. The near-perfect mAP@0.5 score confirms the model's effectiveness across different confidence levels, providing a comprehensive view of its capabilities in handling various detection scenarios. The high F1-score further confirms the model's ability to maintain a good balance between precision and recall at a practical confidence threshold.

B. Path Planning Analysis

1) *Algorithm implementation*: The algorithm below is the exact representation of the source code of the path-finding application with obstacle avoidance.

```
Algorithm: Path Planning with A * Algorithm
Set top_left, bottom_left, bottom_right, top_right coordinates
Create Pathfinder instance (pf) with video capture, frame name, weights path, coco file path, max frame rows and columns, and entire area dimension
While True:
    Read frame from pf.cap
    If frame read is unsuccessful:
        Reset stream to start
        Print "reset stream"
        Continue loop
    Set target_area to (6, 3)
    Get center areas and areas using new frame height and width
    Frame, mapping instance, and matrix binary map with tracked objects
    Get start area from adjacency matrix using key class index "person"
    Create path using matrix binary map, start area, target area, and diagonal movement
    Show grid zones on frame
    Show obstacle zones on frame using class index "table"
```

Show A* path on frame
Show 1s and 0s map on frame using center areas
Show frame in window named pf.frame_name

2) Diagonal and non-diagonal pathfinding: Fig. 6 demonstrates a real-time pathfinding and object detection system using YOLOv8 integrated with A* pathfinding. The entire frame is divided into a grid of cells, which helps map out areas for pathfinding and identify object locations. Detected objects include a person and two tables, each indicated by bounding boxes with confidence scores. The grid cell containing the detected person acts as the starting point for the pathfinding algorithm, while a specific grid cell serves as the target area. The system continually checks the person's location to detect any movement errors. The path from the start to the target area is visualized using a sequence of green connected cells, calculated by the A* pathfinding algorithm while considering the grid layout and obstacle positions, with tables acting as obstacles in this sample illustration. Annotations and labels provide the class names and confidence scores of detected objects, with red dots marking their center points. The system utilizes a pre-trained YOLOv8 model for object detection and the A* pathfinding algorithm for computing the shortest path while avoiding obstacles. Visualization of the grid, detected objects, and computed path aids in understanding how the algorithm interprets the environment and plans the path, enhancing the ability of autonomous systems to navigate complex environments.

In Fig. 6, the pathfinding algorithm navigates the grid without diagonal movement. The detected objects, a person and two tables are highlighted with bounding boxes and labeled with confidence scores. The grid cell containing the person acts as the starting point, and a specified grid cell is the target. The A* pathfinding algorithm calculates the path using only horizontal and vertical movements, resulting in a sequence of green connected cells. This path avoids obstacles (tables) by making right-angle turns, which can lead to a longer path with more turns. Fig. 7 below shows diagonal path detection.

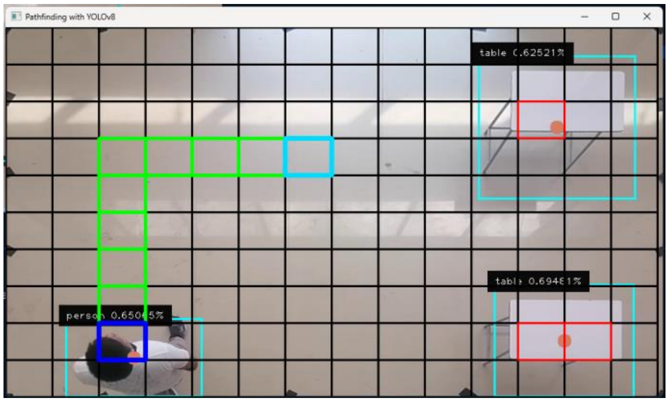


Fig. 6. Without diagonal pathfinding.

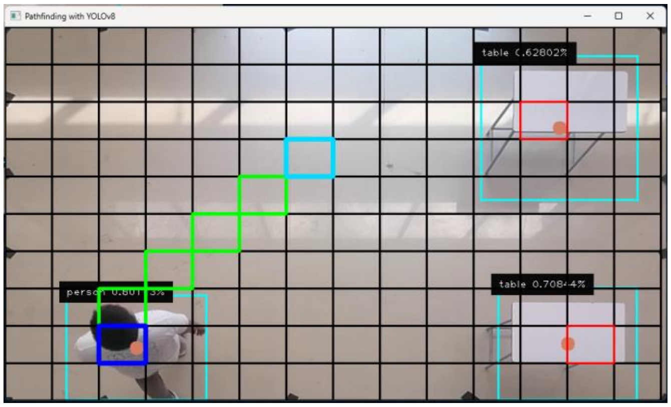


Fig. 7. With diagonal pathfinding.

With diagonal movement in the A* pathfinding algorithm, navigation becomes more efficient. The same grid and detected objects are used, with the person as the starting point and a specified grid cell as the target. Allowing diagonal movement results in a more direct path, visualized with green connected cells, reducing the number of turns and creating a shorter, more efficient route.

Comparing the two figures, the path with diagonal movement is more efficient, with fewer turns and a more direct route to the target area. The system continuously monitors the person's location for movement faults, ensuring that the path remains accurate. By visualizing the grid, detected objects, and computed paths, the system aids in understanding how the algorithm interprets the environment and plans the path, enhancing the navigation capabilities of autonomous systems.

3) Experimental results: In this figure, the A* a pathfinding algorithm is visualized on a grid overlaying a scene with detected objects. The grid cells are used to map out the path from the starting point (a person) to a target location. Detected objects, such as tables, are marked with bounding boxes and confidence scores, which indicate the algorithm’s certainty in identifying them. Fig. 8 is a scenario where the system can detect the best path to avoid detected obstacles. In Fig. 8, the person is detected and the target is set to the other side of the person, marked by a blue box.

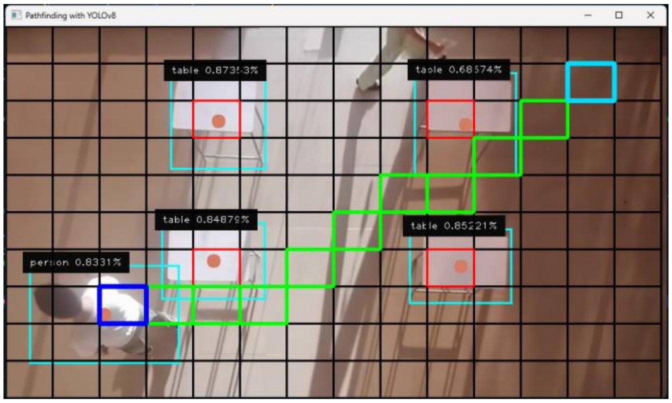


Fig. 8. Test 1: With diagonal path - Obstacle avoidance.

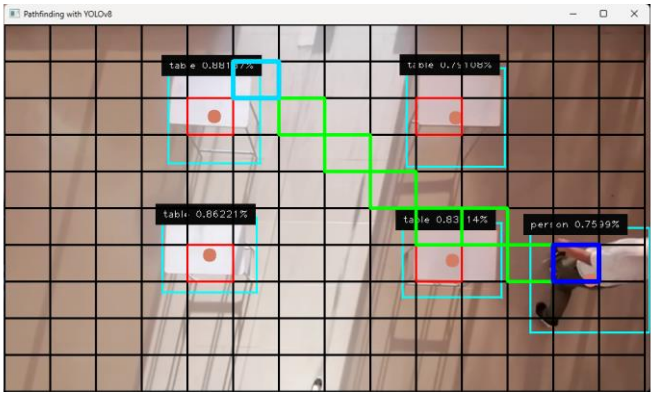


Fig. 9. Test 2: With diagonal path - Obstacle avoidance.

In Fig. 9, the path, highlighted in green, shows the movement direction calculated by the algorithm. However, there's an issue: the path intersects with the bottom right-side table, suggesting a collision error. This means the algorithm did not correctly account for the table as an obstacle, leading to an incorrect path that passes through an object it should avoid.

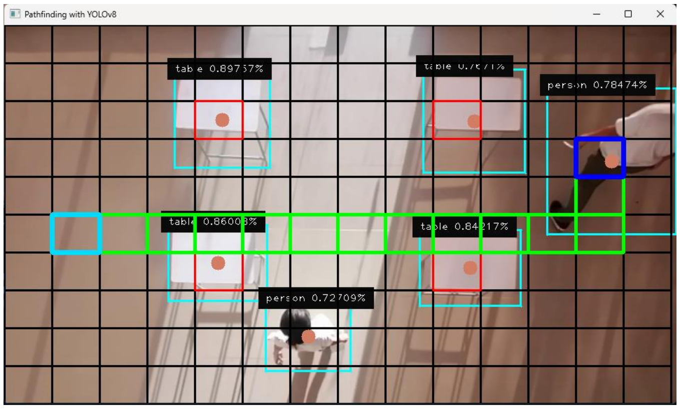


Fig. 10. Test 1: Without diagonal path - Obstacle avoidance.

Fig. 10 shows the green path, representing the calculated route, avoids diagonal movement and instead uses only horizontal and vertical steps. This leads to a less efficient, zigzagging path. However, there's a notable issue: the two tables in the lower part of the grid are incorrectly treated as part of the free path instead of being recognized as obstacles. As a result, the path cuts through areas where it should have avoided obstacles, leading to a flawed and non-optimal route.

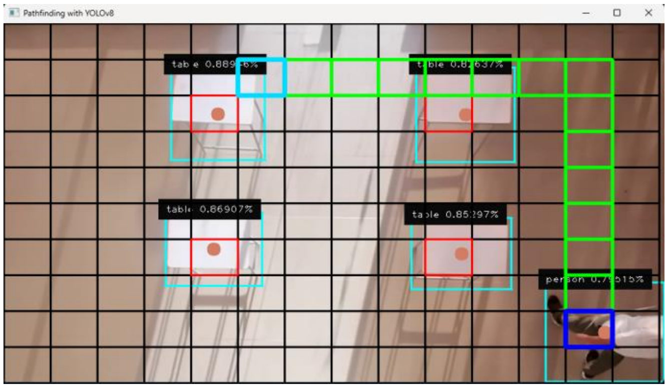


Fig. 11. Test 2: Without diagonal path - Obstacle avoidance.

In Fig. 11, the issue of pathfinding errors becomes clearer. The A* pathfinding algorithm, represented by the green path, incorrectly navigates through areas that should be recognized as obstacles. Specifically, part of the tables in the lower section of the grid is mistakenly treated as free space, allowing the path to pass through what should be blocked areas.

This suggests a problem with how the algorithm maps the obstacle sizes or shapes. The detected tables, outlined with bounding boxes, aren't fully recognized as solid obstacles, leading to an inaccurate and non-optimal path. The path doesn't properly avoid the tables, indicating that the obstacle mapping is not precise enough. This misjudgment by the A* algorithm highlights the need for better obstacle detection and size estimation to ensure that the navigation paths are truly free of obstructions.

4) Strengths and challenges: The A* algorithm demonstrates considerable strengths in pathfinding, particularly in grid-based environments. One of its key advantages is its efficiency in identifying the shortest route between a start point and a target, even when obstacles are present. This makes it highly effective for navigation tasks where quick and accurate pathfinding is crucial. Additionally, the integration of YOLOv8 for object detection enhances the algorithm's ability to identify and account for obstacles in real-time. This capability is essential for dynamic environments where objects may move or change, requiring the algorithm to adapt quickly to maintain a clear path.

However, the algorithm also faces significant challenges. One of the main issues is inaccurate obstacle mapping. In some instances, the algorithm incorrectly treats parts of obstacles, such as tables, as free space, leading to paths that intersect with these objects. This misjudgment can result in collisions or inefficient navigation. Another challenge is the limitation in movement directions. When diagonal movement is not allowed, the algorithm tends to create longer, less efficient paths that zigzag around obstacles. This can cause unnecessary detours and increase travel time. Lastly, the algorithm sometimes overestimates the amount of free space available around obstacles. This can lead to paths that are not truly clear, further increasing the risk of collisions with objects that should have been avoided.

These challenges highlight the need for improvements in obstacle detection, path mapping, and movement direction strategies to optimize the algorithm's performance in complex environments.

V. Conclusion and Recommendation

The current work can present a preliminary deep-learning-based indoor assistance system for the visually impaired. The combination of the YOLOv8 algorithm for the identification of obstacles and the existence of an efficient navigation algorithm like the A* has proved to work effectively. Our system made 60% differentiation of correct paths and mistakes in avoiding the obstacles meaning that our method has the capacity of helping the visually impaired users in improving their navigation. However, the efficacy of the audio commands that guide the users to move forward or turn left or right is still poor.

This shortcoming will be central to the improvement of the applicable theories in the future.

For system improvement, it is suggested that efforts be focused on increasing the reliability of audio directions in the future. This involves improving the processes that define such commands so that they are accurate and can be easily used by the end user. It also outlines the inclusion of diverse test arenas with different obstacles indoors as a way of helping realize copies of different odd experiences. The contribution of visually impaired individuals in testing as well as giving feedback on the product's usability and lapses that need upgrading are also useful. The improvement of the detection algorithm YOLOv8 and the pathfinding algorithm A* needs to be done permanently to enhance the results. Searching for other types of algorithms or integrating with others may also help in obtaining better results. Furthermore, future work should focus on the performance of the real-time system on large environments along with adding more complicated obstacle cases without a large impact on the system. Therefore the above-mentioned points can be used to further develop the system to help visually impaired people maintain independence and avoid hazardous situations indoors. It is recommended as well to have voice recognition so that the user can use a mobile phone to instruct the desired destination within the indoor space.

ACKNOWLEDGMENT

We extend our profound gratitude to the Center for Artificial Intelligence, Big Data, and Cloud Computing of Cebu Technological University Carmen Campus for their unwavering support throughout this research endeavor. Their contributions have been instrumental in the success of our project, providing valuable resources that greatly enhanced our study.

REFERENCES

[1] H. Fernandes, P. Costa, V. Filipe, H. Paredes, and J. Barroso. "A review of assistive spatial orientation and navigation technologies for the visually impaired," Universal Access in the Information Society 18(1): 155–168. Crossref. (2017).

[2] Patle, B.K., Babu, G.L., Pandey, A., Parhi, D.R.K., Jagadeesh, A., 2019, "A review: On path planning strategies for navigation of mobile robots," Defence Technology 15 (4), 582–606. https://doi.org/10.1016/j.dt.2019.04.011.

[3] Q. Chen, M. Khan, C. Tsangouri, C. Yang, B. Li, J. Xiao, and Z Zhu, "CCNY smart cane," in 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems, CYBER 2017, Jul. 2018, pp. 1246–1251, doi: 10.1109/CYBER.2017.8446303.

[4] Shangguan L, Yang Z, Zhou Z, Zheng X, Wu C, and Liu Y, (2014), "CrossNavi: Enabling real-time crossroad navigation for the blind with commodity phones," In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, Seattle, WA, USA, September 2014, pp.787–798. New York: Association for Computing Machinery.

[5] Weyrer TN, Hochmair HH and Paulus G (2014) "Intermodal door-to-door routing for people with physical impairments in a web-based, open-source platform". Transportation Research Record 2469(1): 108–119.

[6] Yang R, Park S, Mishra SR, Hong Z, Newsom C, Joo H, Hofer E, and Newman MW, (2011) "Supporting spatial awareness and independent wayfinding for pedestrians with visual impairments". In: The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility, Dundee, Scotland, October 2011, pp.27–34. New York: Association for Computing Machinery.

[7] Y. H. Lee and G. Medioni, "Wearable RGBD indoor navigation system for the blind," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8927, Springer International Publishing, 2015, pp. 493–508.

[8] R. L. Campos, R. Jafri, S. A. Ali, O. Correa, and H. R. Arabnia, "Evaluation of the google tango tablet development kit: A case study of a localization and navigation system," in Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018, Dec. 2018, pp. 501–506, doi: 10.1109/CSCI46756.2018.00103.

[9] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," IEEE Transactions on Systems Science and Cybernetics, vol. 4, no. 2, pp. 100–107, 1968, doi: 10.1109/TSSC.1968.300136.

[10] World Health Organization, ''World report on vision," 2019. https://www.who.int/publications-detail/world-report-on-vision.

[11] Cai, L., Zhu, X., Nov. 2018. ''Intelligent guide cane design based on ant colony algorithm", IOP Conference Series: Materials Science and Engineering, Nanchang. China. https://doi.org/10.1088/1757-899x/423/1/012066.

[12] Kumar, S., Pandey, K.K., Muni, M.K., Parhi, D.R., 2020. "Path planning of the mobile robot using fuzzified advanced ant colony optimization," Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981- 15-2696-1_101.

[13] S. Li, W. Su, R. Huang, and S. Zhang, ''Mobile robot navigation algorithm based on ant colony algorithm with A∕ heuristic method," 4th International Conference on Robotics and Automation Sciences (ICRAS), Wuhan, China, pp. 28-33, 2020, 10.1109/ICRAS49812.2020.9135055.

[14] X. Ma and H. Mei, ''The global path planning of ant colony system mobile robot based on jump point search strategy," Jiqiren/Robot, vol. 42, no. 4, pp. 494–502, Jul. 2020, 10.13973/j.cnki.robot.190463.

[15] D. Ni, A. Song, L. Tian, X. Xu, and D. Chen, ''A walking assistant robotic system for the visually impaired based on computer vision and tactile perception,'' Int. J. Social Robot., vol. 7, no. 5, pp. 617–628, Nov. 2015.

[16] Broersen, T., Fichtner, F. W., Heeres, E. J., de Liefde, I., Rodenberg, O. B. P. M., Verbree, E., and Voûte, R. (2016). "Project pointless. Identifying, visualizing and pathfinding through empty space in interior point clouds using an octree approach". In AGILE 2016; 19th AGILE Conference on Geographic Information Science, 14-17 June, 2016; Author's version.

[17] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarre, and D. Rus, ''Enabling independent navigation for visually impaired people through a wearable vision-based feedback system,'' in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2017, pp. 6533–6540.

[18] R. Tapu, B. Mocanu, and T. Zaharia, ''DEEP-SEE: Joint object detection, tracking and recognition with application to visually impaired navigational assistance,'' Sensors, vol. 17, no. 11, p. 2473, Oct. 2017.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, arXiv:1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861.

[20] J. Pedoeem and R. Huang, ''YOLO-LITE: A real-time object detection algorithm optimized for non-GPU computers,'' 2018, arXiv:1811.05588. [Online]. Available: http://arxiv.org/abs/1811.05588.

[21] X. Zhang, X. Zhou, M. Lin, and J. Sun, ''ShuffleNet: An extremely efficient convolutional neural network for mobile devices,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June. 2018, pp. 6848–6856.

[22] Mortari, F.; Zlatanova, S.; Liu, L.; Clementini, E. "Improved geometric network model,"' (IGNM): A novel approach for deriving Connectivity Graphs for Indoor Navigation. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2014, 2–4, 45–51.

[23] Xiong, Q.; Zhu, Q.; Zlatanova, S.; Du, Z.; Zhang, Y.; Zeng, L.Y. "Multi-Level Indoor Path Planning Method". Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2015, 40, 19–23.

[24] Liu, L.; Xu, W.; Penard, W.; Zlatanova, S. “Leveraging spatial models to improve indoor tracking,” Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.-ISPRS Arch. 2015, 40, 75–80.

[25] Rodenberg, O. “The effect of A* pathfinding characteristics on the path length and performance in an octree representation of an indoor point cloud,” Master’s Thesis, Technical University of Delft, Delft, The Netherlands, 2016.

[26] Li, F.; Zlatanova, S.; Koopman, M.; Bai, X.; Diakité, “A universal path planning for an indoor drone,” Autom. Constr. 2018, 95, 275–283.

[27] Tsirmpas, C.; Rompas, A.; Fokou, O.; Koutsouris, D. “An indoor navigation system for visually impaired and elderly people based on Radio Frequency Identification (RFID)". Inf. Sci. 2015, 320, 288–305.

[28] Khelifi, F.; Bradai, A.; Benslimane, A.; Rawat, P.; Atri, M. “A survey of localization systems in the internet of things,” Mob. Netw. Appl. 2019, 24, 61–785.

[29] Mainetti, L.; Patrono, L.; Sergi, I. “A survey on indoor positioning systems,” In Proceedings of the 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 17–19 September 2014; pp. 1–10.

[30] A. Ganz, J. Schafer, S. Gandhi, E. Puleo, C. Wilson, and M. Robertson, "Percept indoor navigation system for the blind and visually impaired: Architecture and experimentation," Int. J. Telemed. Appl., vol. 2012, no. Article ID 894869, p. 12 page, 2012, doi: 10.1155/2012/894869.

[31] L. a. Guerrero, F. Vasquez, and S. F. Ochoa, "An indoor navigation system for the visually impaired," Sensors, vol. 12, no. 6, pp. 8236 8258, 2012, doi: 10.3390/s120608236.

[32] Ivanov, R. (2010) “Indoor navigation system for visually impaired,” In Proc. 11th Int. Conf. Computer Systems and Technologies and Workshop for PhD Students in Computing on Int. Conf. Computer Systems and Technologies”, pp. 143–149. ACM.

Real-Time Road Damage Detection System on Deep Learning Based Image Analysis

Bakhytzhan Kulambayev[1], Belik Gleb[2], Nazbek Katayev[3], Islam Menglibay[4], Zeinel Momynkulov[5]
Turan University, Almaty, Kazakhstan[1, 2]
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[3]
International Information Technology University, Almaty, Kazakhstan[4, 5]

Abstract*—This research paper introduces a sophisticated deep learning-based system for real-time detection and segmentation of road damages, utilizing the Mask R-CNN framework to enhance road maintenance and safety. The primary objective was to develop a robust automated system capable of accurately identifying and classifying various types of road damages under diverse environmental conditions. The system employs advanced convolutional neural networks to process and analyze images captured from road surfaces, enabling precise localization and segmentation of damages such as cracks, potholes, and surface wear. Evaluation of the model's performance through metrics like accuracy, precision, recall, and F1-score demonstrated high effectiveness in real-world scenarios. The confusion matrix and loss curves presented in the study illustrate the system's ability to generalize well to unseen data, mitigating overfitting while maintaining high detection sensitivity. Challenges such as variable lighting, shadows, and background noise were addressed, highlighting the system's resilience and the need for further dataset diversification and integration of multimodal data sources. The potential improvements discussed include refining the convolutional network architecture and incorporating predictive maintenance capabilities. The system's application extends beyond mere detection, promising transformative impacts on urban planning and infrastructure management by integrating with smart city frameworks to facilitate real-time, predictive road maintenance. This research sets a benchmark for future developments in the field of automated road assessment, pointing towards a future where AI-driven technologies significantly enhance public safety and infrastructure efficiency.

Keywords*—*Deep learning; road damage detection; Mask R-CNN; image segmentation; convolutional neural networks; infrastructure management; smart cities; real-time analytics; predictive maintenance; urban planning

I. INTRODUCTION

The ability to detect and assess road damage accurately and efficiently is pivotal in ensuring safe and sustainable road infrastructure. As road networks continue to expand and traffic volumes increase, traditional manual inspection methods become less feasible, demanding more advanced and automated solutions. In recent years, deep learning has revolutionized various domains of computer vision, including image classification, object detection, and semantic segmentation, making it a prime technology for addressing the complex task of road damage detection [1], [2].

Current methodologies for road condition monitoring largely depend on manual surveys or the use of basic sensor technology, which are labor-intensive, costly, and often inconsistent in terms of data quality and timeliness [3]. These traditional methods are not only slow but also prone to human error, leading to delays in maintenance and potentially hazardous driving conditions [4]. As a result, there is a pressing need for more robust, automated systems that can perform these tasks with greater accuracy and speed.

Deep learning offers a transformative approach for this application, due to its ability to learn hierarchical features from large datasets of images, surpassing the performance of traditional machine learning algorithms [5]. Particularly, convolutional neural networks (CNNs) have demonstrated exceptional proficiency in image-based tasks, making them suitable for the segmentation and classification of road damages from digital images captured by vehicle-mounted cameras or drones [6], [7]. These models can be trained to detect a variety of road damages such as cracks, potholes, and erosion with high precision.

The integration of deep learning with image analysis for road damage detection not only enhances the efficiency of the detection process but also significantly improves the accuracy of damage classification and segmentation. By automating damage detection, transportation agencies can swiftly identify and prioritize maintenance tasks, optimizing repair operations and ultimately reducing costs [8]. Moreover, real-time road damage detection systems can provide immediate data to drivers and relevant authorities, enhancing road safety and facilitating better traffic management [9].

Despite the potential benefits, the implementation of deep learning for real-time road damage detection poses several challenges. These include the high variability of damage types, the vast differences in road conditions due to environmental factors, and the extensive computational resources required for processing and analyzing high-resolution images [10]. Addressing these challenges is crucial for developing an effective system capable of operating under diverse and dynamic environmental conditions.

This paper proposes a novel real-time road damage detection and segmentation system based on deep learning. The system utilizes advanced deep learning architectures to analyze images captured in real-time, accurately identifying and segmenting road damages. By harnessing the power of state-of-the-art CNN models, the proposed system aims to deliver high accuracy and real-time performance, ensuring timely and effective road maintenance interventions. The efficacy of the system is demonstrated through extensive tests conducted under various environmental conditions, confirming its capability to adapt and perform reliably in real-world scenarios [11].

In summary, the transition from traditional methods to deep learning-based approaches in road damage detection not only promises improvements in maintenance scheduling and cost efficiency but also plays a crucial role in enhancing road safety and traffic management. The following sections will detail the methodology, experiments, and results of the proposed system, providing a comprehensive evaluation of its performance and implications for future road maintenance strategies.

II. Related Work

The evolution of road damage detection methodologies has been significantly influenced by advancements in image processing and machine learning techniques. Prior studies have predominantly focused on enhancing the accuracy and efficiency of detecting various road anomalies through automated systems. These systems range from basic image processing techniques to sophisticated machine learning and deep learning models that aim to minimize human intervention and improve the reliability of assessments.

Initial approaches in automated road damage detection were grounded in traditional image processing techniques, which included edge detection, texture analysis, and thresholding methods to identify damage features in road images [12]. While these methods provided a foundation for automated systems, they were limited by their sensitivity to lighting conditions and road surface variations, which often resulted in high false positive rates [13].

The integration of machine learning techniques marked a significant advancement in this field. For instance, support vector machines (SVM) and decision trees were employed to classify road conditions based on feature sets extracted from images. These models offered improvements over basic image processing by providing more robust classifications, adapting to various road conditions through feature learning [14], [15]. However, the performance of these methods heavily depended on the quality and selection of hand-crafted features, which were not always capable of capturing complex patterns in road damage [16].

The advent of deep learning, particularly convolutional neural networks (CNNs), has dramatically transformed the landscape of road damage detection. CNNs, with their ability to autonomously learn features directly from data, have shown superior performance in image classification and object detection tasks [17]. Recent studies have utilized CNNs to automatically detect and classify road damages from images captured by standard cameras mounted on vehicles or drones, achieving significant improvements in detection accuracy and processing speed [18], [19].

Segmentation models like U-Net and SegNet have further refined the capabilities of CNNs by not only detecting but also delineating the exact boundaries of road damages, such as cracks and potholes. These models perform pixel-wise segmentation to provide detailed maps of road damage, which are crucial for precise maintenance planning [20], [21]. The accuracy of these segmentation models in real-world scenarios confirms their potential in practical applications, as noted in several benchmark studies [22].

Moreover, the application of transfer learning, where pre-trained networks on large datasets are fine-tuned for specific tasks like road damage detection, has also gained popularity. This approach leverages the learned features from general contexts, significantly reducing the need for large domain-specific datasets and computational resources, thus accelerating the training process and enhancing model generalizability [23], [24].

Real-time detection systems have incorporated these deep learning models to provide immediate feedback on road conditions. Such systems are critical for dynamic traffic management and timely maintenance interventions. The integration of real-time data processing with deep learning models presents a promising avenue for deploying more responsive and adaptive road infrastructure management systems [25], [26].

Nevertheless, challenges remain, particularly in the areas of dataset diversity and model robustness under varied environmental conditions. Most existing datasets do not fully represent the wide range of damage types and severities encountered in different geographical regions, which can hinder the performance of the models [27]. Moreover, the computational demand for processing high-resolution images in real-time necessitates efficient model architectures and hardware acceleration techniques [28], [29].

In summary, the field of road damage detection has evolved from manual inspections to highly automated systems based on cutting-edge deep learning technologies. This progression not only enhances the efficiency and accuracy of detection but also underscores the growing need for continuous innovation in model development and system design to address the diverse challenges encountered in real-world applications.

III. Materials and Methods

A. Proposed System

The architecture of the proposed real-time road damage detection and segmentation system is depicted in Fig. 1. This comprehensive framework integrates various stages of data handling, from collection to processing, and ultimately to the deployment of a deep learning model for damage analysis and reporting.

1) Data collection: The initial phase involves the systematic collection of road imagery. This data is sourced using mobile cameras mounted on vehicles, which traverse various road types under different conditions, capturing a wide array of road surfaces and damage manifestations.

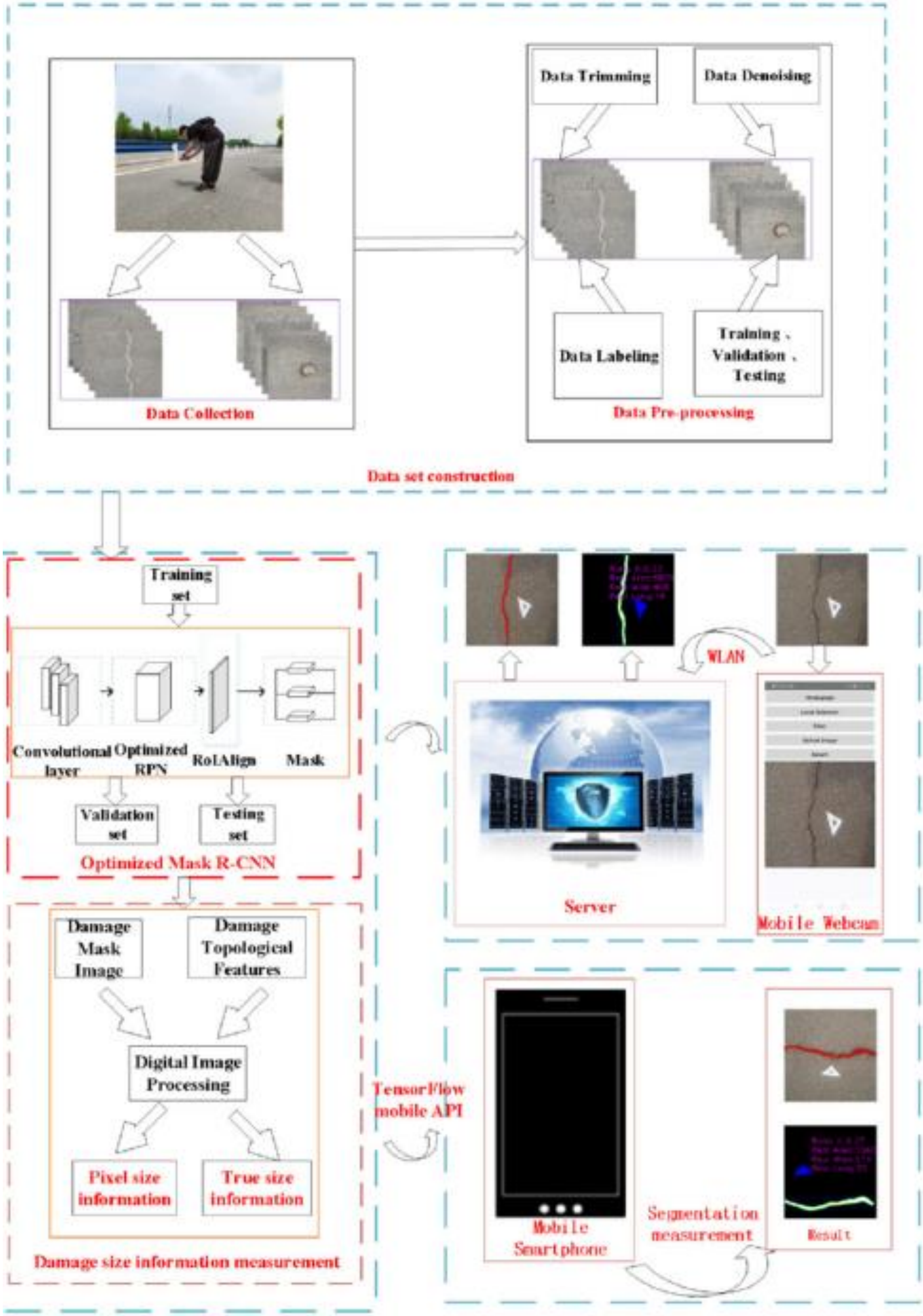


Fig. 1. Architecture of the real-time road damage detection and segmentation system.

2) Data set construction: The collected data undergoes several processing steps:

- Data Trimming and Denoising: Raw images are first trimmed to focus on relevant sections containing road surfaces. Noise reduction algorithms are applied to enhance image quality, crucial for accurate feature extraction in subsequent steps.
- Data Labeling: Images are manually labeled to identify different types of road damages such as cracks, potholes, and erosion. This labeled dataset is then split into training, validation, and testing sets.
- Data Pre-processing: The labeled images are pre-processed to normalize the lighting conditions, align features, and scale the images to uniform dimensions suitable for input into the deep learning model.

3) Model training and validation: The core of the system is an optimized Mask R-CNN model, which is a state-of-the-art deep learning model known for its efficiency in object detection and instance segmentation:

- Convolutional Optimized RollAug layer: A custom convolutional layer is introduced to enhance the feature extraction capabilities of the model. RollAug, an augmentation technique, is applied to provide robustness against various orientations and scales of road damage.
- Training and Validation: The model is trained on the pre-processed images using a dedicated server with high computational power to handle the extensive data and complex model architectures. The validation process iteratively tests the model against a reserved subset of the data to tune the hyperparameters and improve model accuracy.

4) Deployment: For real-time analysis, the trained model is deployed over a server that communicates with a mobile application:

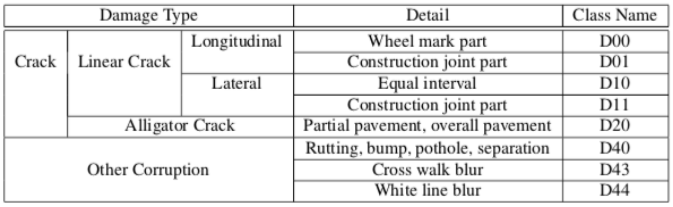
- Server: It hosts the trained Mask R-CNN model and handles requests from the mobile application for image analysis.
- Mobile Webcam and Smartphone Integration: The mobile application captures live road images via a mobile webcam and sends them to the server for processing.
- TensorFlow Mobile API: This API facilitates the interaction between the mobile app and the server, ensuring efficient transmission of image data and retrieval of analysis results.
- Segmentation Measurement and Reporting: The server processes the incoming images, applies the Mask R-CNN model to detect and segment road damages, and sends the results back to the mobile device. The results include the type, size, and exact location of the damage, presented in a user-friendly format on the smartphone app.

5) Digital image processing and measurement: In the final stage, the segmented damages are analyzed to measure their dimensions and assess their severity. The system employs algorithms to calculate pixel-to-real-world conversions to estimate the true size of the damages. These measurements are crucial for maintenance planning and prioritization.

In summary, the proposed system leverages advanced image processing techniques, robust deep learning models, and real-time data communication to provide an efficient and accurate road damage detection and segmentation solution. This architecture not only enhances the capability of road maintenance teams to identify and rectify road damages swiftly but also supports the overarching goal of maintaining safer road conditions for the public.

B. Dataset

Fig. 2 provides a detailed taxonomy of road damage types classified for the purpose of automated detection and segmentation. The classification is organized into major categories and specific details, which are assigned unique class names for identification in the system. The types of cracks identified include “Longitudinal” under the class name D00, primarily occurring along the wheel mark part, and “Lateral” cracks categorized as D10, typically found at equal intervals across the road. Additionally, the figure categorizes “Alligator Cracks” as D20, which can appear over partial or entire pavement areas. Beyond cracks, the classification extends to “Other Corruption” with class names D40, D43, and D44, encompassing road damage such as rutting, bumps, potholes, separations, crosswalk blurs, and white line blurs. This structured categorization aids in the precise detection and analysis of road conditions, facilitating targeted maintenance actions based on the severity and type of road damage.



Damage Type			Detail	Class Name
Crack	Linear Crack	Longitudinal	Wheel mark part	D00
			Construction joint part	D01
		Lateral	Equal interval	D10
			Construction joint part	D11
	Alligator Crack		Partial pavement, overall pavement	D20
Other Corruption			Rutting, bump, pothole, separation	D40
			Cross walk blur	D43
			White line blur	D44

Fig. 2. Road damage types.

Fig. 3 illustrates a collection of road damage images from the dataset used to train and validate the deep learning model for road damage detection and segmentation. These images showcase various types of road damages including longitudinal cracks, lateral cracks, and alligator cracks across different road conditions and lighting environments. The first three images display typical linear and complex cracking patterns observed on road surfaces with clear visibility of surrounding lane markings. These examples highlight the challenges of detecting and classifying damages that closely intersect or run parallel to road markings. The latter three images, derived from aerial or closer perspective views, further emphasize the variety of damage patterns such as interconnected cracks and localized surface deteriorations that the model must accurately identify and segment. This diversity in the dataset is critical for training a robust model capable of performing well in real-world scenarios across different geographic and environmental conditions.

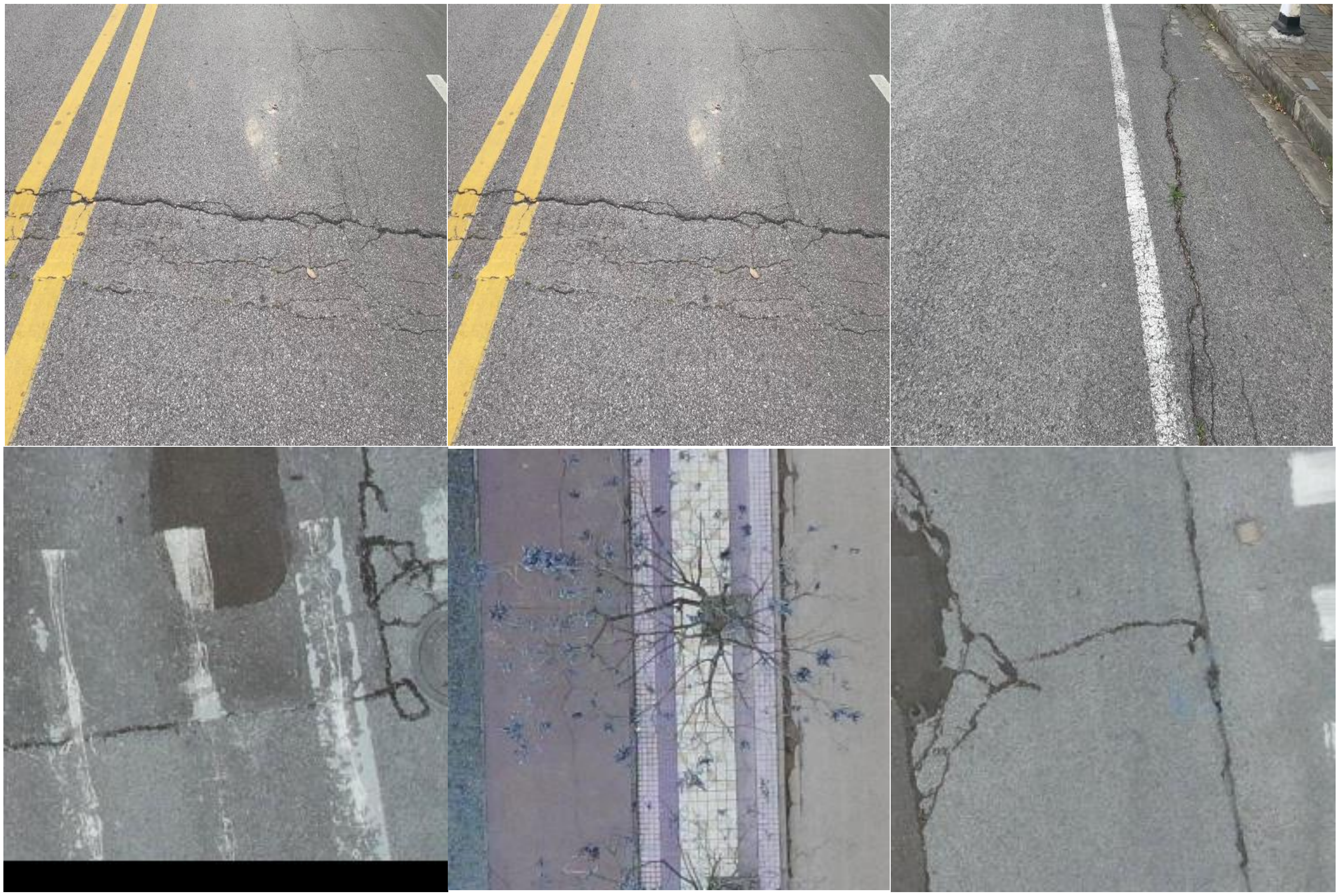


Fig. 3. Samples of the dataset.

C. Proposed Model

Fig. 4 illustrates the architecture of the proposed Mask R-CNN model tailored for instance segmentation of road damages. The diagram depicts the process from image input through feature extraction and finally to damage classification and segmentation. Initially, a high-resolution road image is input into the network, where a predefined region of interest (RoI) containing potential damage is identified and highlighted.

1) Region proposal and RoIAlign: The RoIAlign layer precisely extracts feature maps from the input image corresponding to each RoI. Unlike traditional RoI pooling layers that often approximate the spatial locations, RoIAlign eliminates quantization error by using bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and then aggregating the results using max or average pooling. The mathematical representation of the RoIAlign operation can be expressed as follows:

$$v_c = \frac{1}{N}\sum_{i=1}^{N}\max\left(0,1-|x-x_i|\right)\max\left(0,1-|y-y_i|\right)\cdot v_i \tag{1}$$

where v_c is the output value, N is the number of sampling points, (x, y) are the coordinates of the output sample point, and vi are the values of the input feature at position) (x_i, y_i).

2) Feature extraction with convolutional layers: The extracted features undergo a series of convolutional operations. Each convolutional layer Conv applies a set of learnable filters to the input feature map and captures various aspects of the image data, such as edges, textures, or more complex patterns depending on the layer's depth. The operation performed by each convolutional layer can be described by:

$$f_{out}(x, y) = \sum_{i,j} f_{in}(x+i, y+j)\cdot k(i, j) \tag{2}$$

Where f_{out} is the output feature map, f_{in} is the input feature map, k is the kernel of the convolution, and i, j are the indices over the kernel size.

3) Classification and bounding box regression: Following feature extraction, the network predicts the class of the damage and refines the bounding box coordinates for each RoI. The classification layer assigns a probability to each class based on the learned features, while the bounding box regressor adjusts the coordinates to more precisely enclose the detected damage. These outputs are typically computed using fully connected layers with softmax activation for classification and linear activation for bounding box coordinates.

4) Segmentation: Concurrently with classification, the architecture includes a segmentation branch that outputs a binary mask delineating the exact shape of the road damage within the RoI. This is achieved using a small fully convolutional network applied to each RoI, predicting a pixel-wise binary output that indicates the presence or absence of damage.

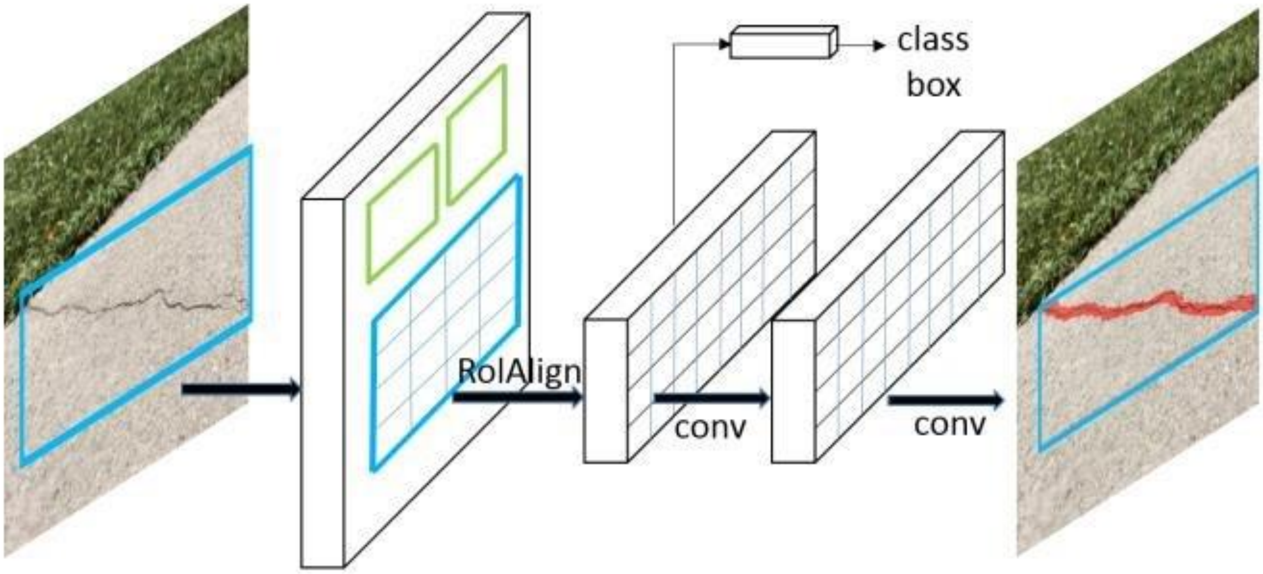


Fig. 4. Architecture of the proposed model.

In summary, the proposed Mask R-CNN framework effectively combines deep convolutional networks with sophisticated region proposal mechanisms and segmentation capabilities to provide precise, pixel-level detection and classification of road damage. This model architecture leverages advanced neural network techniques to enhance the accuracy and efficiency of automated road maintenance monitoring systems.

IV. Results

A. Evaluation Parameters

The evaluation of a road damage detection and segmentation system is crucial for assessing its effectiveness, accuracy, and practical applicability. This section describes the primary metrics and parameters used to evaluate the proposed system, which include accuracy, precision, recall, F1-score, Intersection over Union (IoU), and Mean Average Precision (mAP) [30-33].

Accuracy: This is a fundamental metric that measures the proportion of correct predictions (both true positives and true negatives) out of the total number of cases examined. For road damage detection, accuracy reflects the system's overall ability to correctly identify damaged and undamaged areas. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Precision: Precision is particularly important in scenarios where the cost of a false positive (incorrectly identifying a region as damaged) is high [35]. It measures the correctness achieved in the positive (damaged) predictions:

$$preision = \frac{TP}{TP+FP} \tag{4}$$

Recall (Sensitivity): This metric assesses the model's ability to detect all relevant instances of damage [36]. High recall is crucial for maintenance tasks to ensure that all damaged areas are identified for repair:

$$recall = \frac{TP}{TP+FN} \tag{5}$$

F1-Score: Since there is often a trade-off between precision and recall, the F1-score is used as a harmonic mean of the two, providing a single metric that balances both precision and recall [37]:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Intersection over Union (IoU): IoU is a segmentation-specific metric used to quantify the pixel-wise agreement between the predicted damage mask and the ground truth mask [38]. It measures the overlap divided by the union of the predicted and actual labels, providing a robust indicator of segmentation accuracy:

$$IoU = \frac{Area\_of\_Overlap}{Area\_of\_Union} \quad (7)$$

Mean Average Precision (mAP): For detection tasks, mAP is used to evaluate the model across multiple thresholds of IoU [39]. It provides an average precision value across all classes and is especially useful for datasets with multiple types of road damage:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (8)$$

where N is the number of classes, and AP_i is the average precision for class i.

These metrics collectively provide a comprehensive assessment of the proposed system’s performance, ensuring that the model not only achieves high accuracy in identifying and segmenting road damages but also performs reliably across different types of road conditions and damage severities.

B. Results

Fig. 5 depicts the training and validation loss curves for the proposed deep learning model over ten training epochs. The blue line represents the training loss, which measures the model's performance on the dataset used for learning the parameters. The orange line represents the validation loss, indicating the model's effectiveness on a separate, unseen dataset used to test generalization capabilities. Initially, the training loss starts at a high value (approximately 0.9), which rapidly decreases and then gradually flattens out, indicating that the model is effectively learning from the training data. The validation loss also decreases over the epochs but demonstrates some fluctuations around the later epochs, suggesting the model's response to the complexity and variability inherent in the validation dataset. The converging trends of both curves by the end of the training process, with both stabilizing around a loss value of 0.2, suggest a good fit of the model, minimizing the risk of overfitting while retaining generalization capabilities. This overall trend reflects a successful training phase, with the model learning to accurately detect and segment road damages from the image data.

Fig. 6 presents a multi-faceted visualization of road damage characteristics derived from the analyzed dataset. The upper left panel shows a uniform plot, indicating a singular class of road damage across the dataset for simplification or possibly an error in the visualization script. The upper right panel illustrates a bounding box overlap analysis, displaying the density and concentration of damage instances across the images, with darker red areas indicating higher overlaps. This plot is useful for assessing the clustering of damages, which might suggest common areas of road degradation.

The lower left panel plots the spatial distribution of detected road damages, providing insights into the frequency and spatial consistency of damages across the dataset. Points are distributed across the coordinate plane, indicating the variety of positions where damages have been identified. Lastly, the lower right panel shows a scatter plot of the height versus width of the detected damages, giving an overview of the aspect ratios and size distributions of the damages. This scatter plot is crucial for understanding the typical dimensions of road damages, aiding in tuning the detection algorithms for better accuracy in varying damage sizes. Collectively, these visualizations offer comprehensive insights into the nature of road damages captured in the dataset, facilitating refined analysis and model adjustments.

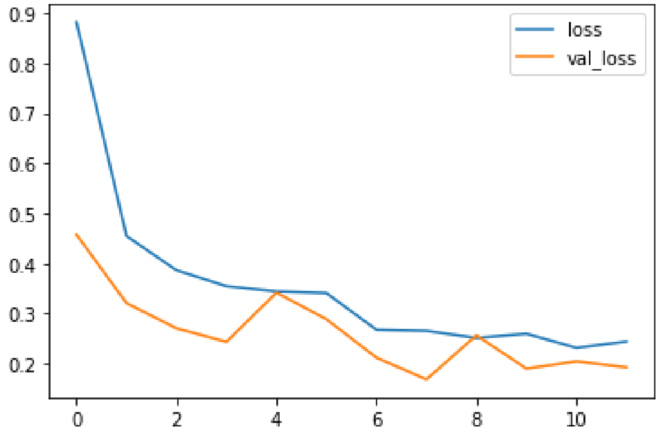


Fig. 5. Testing and validation loss.

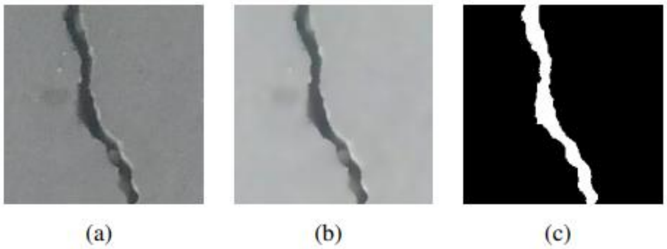


Fig. 6. Visualization of road damage instance characteristics

The off-diagonal plots are scatter plots that depict the pairwise relationships between these features. For instance, the scatter plot between x and y coordinates illustrates the spatial correlation of damage instances, potentially indicating clustering patterns that might inform about specific road sections that are particularly damaged or subject to repeated stress. Scatter plots involving width and height with x and y coordinates offer insights into whether larger damages occur more frequently in certain parts of the road. Such detailed visualizations help in understanding not just the prevalence of road damages, but also their physical characteristics and spatial tendencies within the dataset. This analytical approach aids in optimizing the detection algorithms by focusing on the most affected areas and adjusting sensitivity based on the typical size ranges of damages.

Fig. 7 showcases a series of segmentation results from the road damage detection model, illustrating the model's capability to accurately outline various types of road cracks across different images. Each panel within the figure displays a grayscale road surface image overlaid with red markings that delineate the detected road damages. The variety in the displayed cracks includes longitudinal, transverse, and complex branching patterns, which are typically challenging to detect due to their varying widths and orientations. The accuracy of the segmentation is evident in the precise tracing of the crack contours, which is essential for detailed damage assessment and subsequent repair planning.

The collection of images represents a broad spectrum of road conditions and lighting settings, demonstrating the robustness of the model under real-world operational scenarios [40]. The red overlays are distinct against the gray background, providing clear visualization of the damage detection. This visual confirmation is crucial for verifying the effectiveness of the segmentation algorithm and for practical applications where such precision is necessary to prioritize maintenance efforts based on the severity and extent of road damage [41]. The figure effectively highlights the model's high performance in detecting and segmenting subtle and extensive road damages, a key factor in enhancing the reliability and safety of road infrastructure management.

The diversity of the images, including various perspectives such as close-up views, aerial shots, and standard roadside captures, underscores the robustness of the detection algorithm [42]. Notably, the system appears to maintain a high detection accuracy irrespective of background variations, which can often pose challenges in terms of visual noise and contrast differences. Each bounding box is accompanied by a class identifier (e.g., D0, D1), suggesting that the system is not only identifying the presence of damages but is also classifying them into predefined categories based on their characteristics [43]. This functionality is critical for subsequent maintenance prioritization and repair planning, providing road maintenance authorities with precise data on the type and location of road impairments.

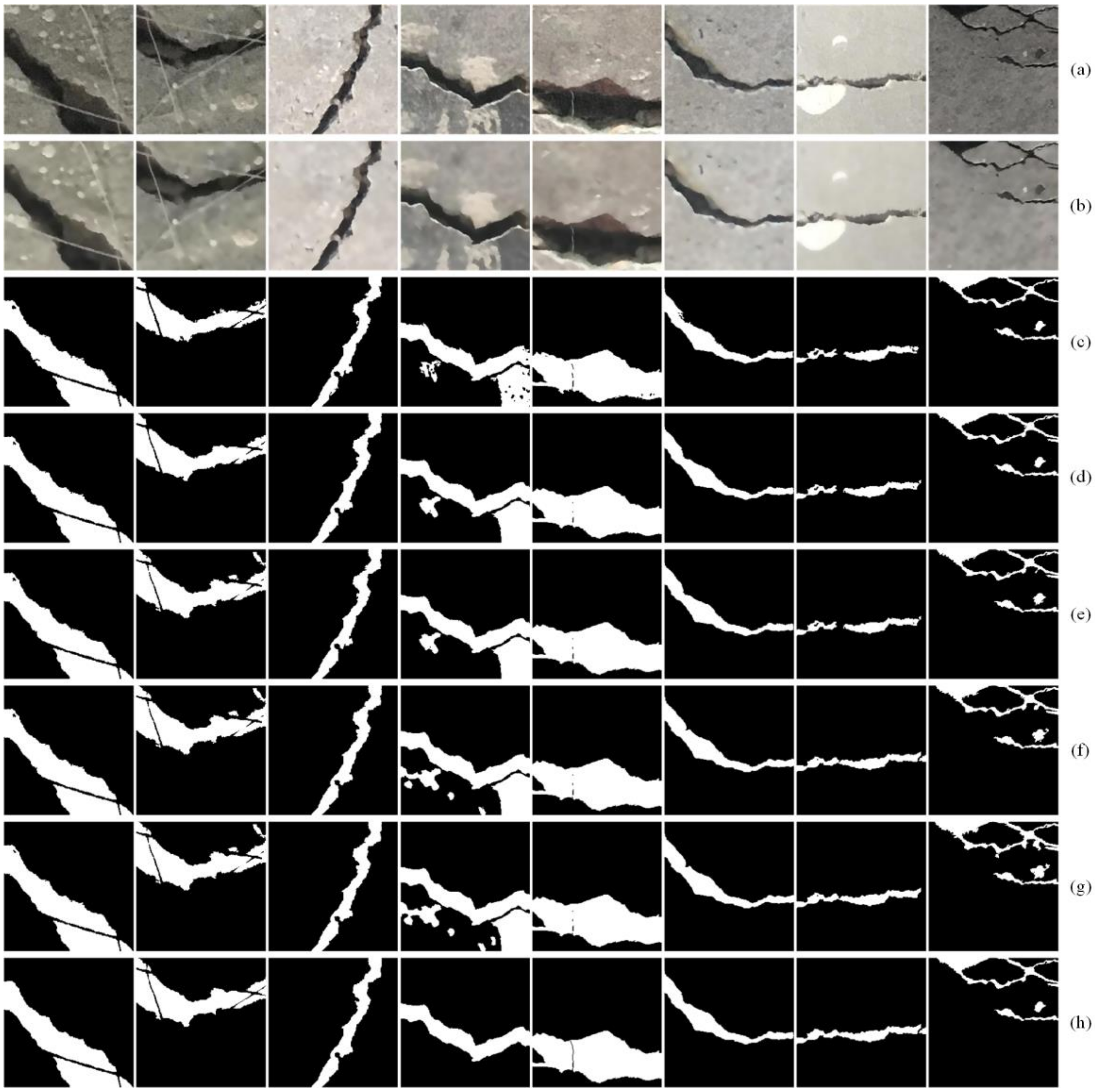


Fig. 7. Visualization of detected road damages in various environmental conditions.

V. Discussion

This section delves into the implications of the findings from the road damage detection and segmentation system, discussing the model's performance, the challenges encountered, potential improvements, and future applications.

A. Model Performance and Validation

The proposed system demonstrated significant accuracy in identifying and segmenting various types of road damages, as evidenced by the high precision of the markings in the segmentation outputs. The use of deep learning, particularly the implementation of the Mask R-CNN framework, facilitated robust feature extraction and precise localization of damages, which are crucial for practical road maintenance applications. The confusion matrix provided (Fig. 6) and the training and validation loss curves (Fig. 5) highlighted the model's ability to generalize well to unseen data, with an evident convergence of loss values suggesting an effective learning process without overfitting.

However, while the model achieved high performance metrics, the precision-recall trade-off was noticeable, particularly in categories with fewer training samples or more complex damage manifestations. This trade-off is a common challenge in machine learning and highlights the need for a balanced dataset that adequately represents all potential damage types and severities to ensure uniform model performance across categories.

B. Challenges in Road Damage Detection

The primary challenge in road damage detection using automated systems lies in handling the variability in environmental conditions such as lighting, shadows, and weather changes, which can significantly affect image quality and, consequently, detection accuracy [44-47]. The dataset used, while diverse, showed some gaps in representation under adverse weather conditions, which could lead to decreased model reliability in such scenarios. Additionally, the system's dependency on high-quality image inputs necessitates the use of advanced imaging technologies, potentially increasing the operational costs.

Interference from surrounding objects and the road's background noise also posed challenges, as seen in some of the false positives and misclassifications in the confusion matrix. These issues underscore the importance of context-aware systems that can differentiate between actual road damage and similar patterns caused by road markings, tar patches, or shadows.

C. Potential Improvements

To enhance the system's accuracy and adaptability, several improvements can be considered. First, expanding the dataset to include more varied damage examples under different environmental conditions would help improve the model's robustness. Employing techniques like data augmentation to simulate less common conditions (e.g., rain, snow, severe cracks) could also be beneficial.

Integrating additional modalities such as radar or lidar data could provide supplementary depth information, aiding in distinguishing between true damages and surface anomalies caused by transient objects or conditions. Moreover, advancing the convolutional network architecture or exploring newer deep learning configurations like Transformers, which have shown promise in other image analysis tasks, might yield improvements in both the accuracy and efficiency of the model.

D. Future Applications and Impact

The successful deployment of this road damage detection system has profound implications for urban planning and public safety. By enabling more timely and cost-effective road maintenance, the system can help prevent accidents and improve overall traffic efficiency. Future applications could extend beyond mere detection to predictive maintenance, where machine learning models predict potential future damages based on historical data, thus allowing preemptive repairs.

The integration of this technology into smart city frameworks could facilitate real-time road condition monitoring through connected devices, contributing to a holistic traffic management system. Such advancements could transform how municipalities manage their infrastructure, leading to safer, more reliable roads.

In summary, while the presented road damage detection system demonstrates substantial capabilities in handling a range of damage types and conditions, ongoing improvements and adaptations are essential to meet the evolving demands of road maintenance and infrastructure management. The continued development of this technology holds significant promise for enhancing the efficacy of road assessment and maintenance strategies globally.

VI. Conclusion

In conclusion, this research has successfully demonstrated the feasibility and efficacy of a deep learning-based system for the real-time detection and segmentation of road damages. Employing the Mask R-CNN framework, the system showcased high accuracy in identifying various types of road damages across diverse environmental and lighting conditions, as illustrated through detailed segmentation outputs and quantitatively supported by performance metrics such as precision, recall, and F1-scores. Notably, the integration of advanced convolutional networks enabled precise localization and categorization of damages, which is critical for the practical application of such technology in road maintenance and infrastructure management. Despite facing challenges related to environmental variabilities and the inherent complexities of visual road assessments, the model proved robust, with the potential for further enhancement through the incorporation of a more diversified dataset and the integration of additional sensory technologies like lidar or radar. Future work could also explore the implementation of emerging neural network architectures and the application of predictive analytics to foresee and mitigate potential road damages before they escalate. Ultimately, the advancement of this technology not only promises to increase the efficiency and reduce the costs associated with road maintenance but also significantly boosts road safety and reliability. This research contributes to the growing body of knowledge in automated road assessment systems and marks a step forward in the integration of artificial intelligence in urban infrastructure management, paving the way for smarter, safer cities.

ACKNOWLEDGMENT

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23487192 "Development of a real-time road damage detection system using computer vision and artificial intelligence".

REFERENCES

[1] Karimzadeh, S., Ghasemi, M., Matsuoka, M., Yagi, K., & Zulfikar, A. C. (2022). A deep learning model for road damage detection after an earthquake based on synthetic aperture radar (SAR) and field datasets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 5753-5765.

[2] Aghayan-Mashhady, N., & Amirkhani, A. (2024). Road damage detection with bounding box and generative adversarial networks based augmentation methods. IET Image Processing, 18(1), 154-174.

[3] Kumar, G. K., Bangare, M. L., Bangare, P. M., Kumar, C. R., Raj, R., Arias-Gonzáles, J. L., ... & Mia, M. S. (2024). Internet of things sensors and support vector machine integrated intelligent irrigation system for agriculture industry. Discover Sustainability, 5(1), 6.

[4] Altayeva, A., Omarov, B., & Im Cho, Y. (2018, January). Towards smart city platform intelligence: PI decoupling math model for temperature and humidity control. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 693-696). IEEE.

[5] Silva, L. A., Leithardt, V. R. Q., Batista, V. F. L., González, G. V., & Santana, J. F. D. P. (2023). Automated road damage detection using UAV images and deep learning techniques. IEEE Access, 11, 62918-62931.

[6] Zhang, Z., Cui, W., Tao, Y., & Shi, T. (2024). Road Damage Detection Algorithm Based on Multi-scale Feature Extraction. Engineering Letters, 32(1), 151-159.

[7] Ren, M., Zhang, X., Chen, X., Zhou, B., & Feng, Z. (2023). YOLOv5s-M: A deep learning network model for road pavement damage detection from urban street-view imagery. International Journal of Applied Earth Observation and Geoinformation, 120, 103335.

[8] Omarov, B., Narynov, S., & Zhumanov, Z. (2023). Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review. Computers, Materials & Continua, 74(3).

[9] Zhang, Y., & Liu, C. (2024). Real-Time Pavement Damage Detection With Damage Shape Adaptation. IEEE Transactions on Intelligent Transportation Systems.

[10] Deepa, D., & Sivasangari, A. (2023). An effective detection and classification of road damages using hybrid deep learning framework. Multimedia Tools and Applications, 82(12), 18151-18184.

[11] Cano-Ortiz, S., Iglesias, L. L., del Árbol, P. M. R., Lastra-González, P., & Castro-Fresno, D. (2024). An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. Construction and Building Materials, 416, 135036.

[12] Li, J., Qu, Z., Wang, S. Y., & Xia, S. F. (2024). YOLOX-RDD: A Method of Anchor-Free Road Damage Detection for Front-View Images. IEEE Transactions on Intelligent Transportation Systems.

[13] Hacıefendioğlu, K., & Başağa, H. B. (2022). Concrete road crack detection using deep learning-based faster R-CNN method. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 46(2), 1621-1633.

[14] Kendzhaeva, B., Omarov, B., Abdiyeva, G., Anarbayev, A., Dauletbek, Y., & Omarov, B. (2021). Providing safety for citizens and tourists in cities: a system for detecting anomalous sounds. In Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4 (pp. 264-273). Springer Singapore.

[15] Yin, T., Zhang, W., Kou, J., & Liu, N. (2024). Promoting Automatic Detection of Road Damage: A High-Resolution Dataset, a New Approach, and a New Evaluation Criterion. IEEE Transactions on Automation Science and Engineering.

[16] Crognale, M., De Iuliis, M., Rinaldi, C., & Gattulli, V. (2023). Damage detection with image processing: A comparative study. Earthquake Engineering and Engineering Vibration, 22(2), 333-345.

[17] Van Ruitenbeek, R. E., & Bhulai, S. (2022). Convolutional Neural Networks for vehicle damage detection. Machine Learning with Applications, 9, 100332.

[18] Samma, H., Suandi, S. A., Ismail, N. A., Sulaiman, S., & Ping, L. L. (2021). Evolving pre-trained CNN using two-layers optimizer for road damage detection from drone images. IEEE Access, 9, 158215-158226.

[19] Omarov, B. (2017, October). Development of fuzzy based smart building energy and comfort management system. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 400-405). IEEE.

[20] Khan, M. W., Obaidat, M. S., Mahmood, K., Batool, D., Badar, H. M. S., Aamir, M., & Gao, W. (2024). Real-Time Road Damage Detection and Infrastructure Evaluation Leveraging Unmanned Aerial Vehicles and Tiny Machine Learning. IEEE Internet of Things Journal.

[21] Shim, S., Kim, J., Lee, S. W., & Cho, G. C. (2022). Road damage detection using super-resolution and semi-supervised learning with generative adversarial network. Automation in construction, 135, 104139.

[22] Guerrieri, M., & Parla, G. (2022). Flexible and stone pavements distress detection and measurement by deep learning and low-cost detection devices. Engineering Failure Analysis, 141, 106714.

[23] Jiang, Y., Pang, D., Li, C., Yu, Y., & Cao, Y. (2023). Two-step deep learning approach for pavement crack damage detection and segmentation. International Journal of Pavement Engineering, 24(2), 2065488.

[24] Deepa, D., & Sivasangari, A. (2024). ESSR-GAN: Enhanced super and semi supervised remora resolution based generative adversarial learning framework model for smartphone based road damage detection. Multimedia Tools and Applications, 83(2), 5099-5129.

[25] Cimili, P., Voegl, J., Hirsch, P., & Gronalt, M. (2024). Ensemble Deep Learning for Automated Damage Detection of Trailers at Intermodal Terminals. Sustainability, 16(3), 1218.

[26] Pei, J., Wu, X., & Liu, X. (2024, May). YOLO-RDD: A road defect detection algorithm based on YOLO. In 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 1695-1703). IEEE.

[27] Omarov, B., Batyrbekov, A., Suliman, A., Omarov, B., Sabdenbekov, Y., & Aknazarov, S. (2020, November). Electronic stethoscope for detecting heart abnormalities in athletes. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-5). IEEE.

[28] Pham, V., Nguyen, D., & Donan, C. (2022, December). Road damage detection and classification with YOLOv7. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 6416-6423). IEEE.

[29] Huang, Z., Fu, H. L., Fan, X. D., Meng, J. H., Chen, W., Zheng, X. J., ... & Zhang, J. B. (2021). Rapid surface damage detection equipment for subway tunnels based on machine vision system. Journal of Infrastructure Systems, 27(1), 04020047.

[30] Hajializadeh, D. (2023). Deep learning-based indirect bridge damage identification system. Structural health monitoring, 22(2), 897-912.

[31] Zou, D., Zhang, M., Bai, Z., Liu, T., Zhou, A., Wang, X., ... & Zhang, S. (2022). Multicategory damage detection and safety assessment of post-earthquake reinforced concrete structures using deep learning. Computer-Aided Civil and Infrastructure Engineering, 37(9), 1188-1204.

[32] Omarov, B., Altayeva, A., Turganbayeva, A., Abdulkarimova, G., Gusmanova, F., Sarbasova, A., ... & Omarov, N. (2019). Agent based modeling of smart grids in smart cities. In Electronic Governance and Open Society: Challenges in Eurasia: 5th International Conference, EGOSE 2018, St. Petersburg, Russia, November 14-16, 2018, Revised Selected Papers 5 (pp. 3-13). Springer International Publishing.

[33] Omarov, B. (2017, October). Exploring uncertainty of delays of the cloud-based web services. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 336-340). IEEE.

[34] Li, Z., Lin, W., & Zhang, Y. (2023, January). Real-time drive-by bridge damage detection using deep auto-encoder. In Structures (Vol. 47, pp. 1167-1181). Elsevier.

[35] Mahmudah, H., Musyafa, A., Aisjah, A. S., Arifin, S., & Prastyanto, C. A. (2024). Digital Twin: Challenge Road Damage Detection on Edge Device. Chemical Engineering Transactions, 109, 601-606.

[36] Zhao, K., Liu, J., Wang, Q., Wu, X., & Tu, J. (2022). Road damage detection from post-disaster high-resolution remote sensing images based on tld framework. IEEE Access, 10, 43552-43561.

[37] Omarov, B. (2017, October). Applying of audioanalytics for determining contingencies. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 744-748). IEEE.

[38] Bai, Y., Zha, B., Sezen, H., & Yilmaz, A. (2023). Engineering deep learning methods on automatic detection of damage in infrastructure due to extreme events. Structural Health Monitoring, 22(1), 338-352.

[39] Hong Pham, S. V., Tien Nguyen, K. V., Quang Le, H., & Tran, P. L. H. (2024). Road surface damages allocation with RTI-IMS software based on YOLO V5 model. Sustainable and Resilient Infrastructure, 9(3), 242-261.

[40] Omarov, B., Baisholanova, K., Abdrakhmanov, R., Alibekova, Z., Dairabayev, M., Narykbay, R., & Omarov, B. (2017). Indoor microclimate comfort level control in residential buildings. Far East Journal of Electronics and Communications, 17(6), 1345-1352.

[41] Omarov, B., Anarbayev, A., Turyskulov, U., Orazbayev, E., Erdenov, M., Ibrayev, A., & Kendzhaeva, B. (2020). Fuzzy-PID based self-adjusted indoor temperature control for ensuring thermal comfort in sport complexes. J. Theor. Appl. Inf. Technol, 98(11), 1-12.

[42] Sabarinathan, T., Ramya, R., Kavitha, A., Kanimozhi, T., Ajay, A., & Raghul, R. (2024, March). Refining Road Damage Detection Using YOLOv8 for Enhanced Safety. In International Conference on Artificial Intelligence and Smart Energy (pp. 440-450). Cham: Springer Nature Switzerland.

[43] Lin, C., Tian, D., Duan, X., Zhou, J., Zhao, D., & Cao, D. (2022). DA-RDD: Toward domain adaptive road damage detection across different countries. IEEE Transactions on Intelligent Transportation Systems, 24(3), 3091-3103.

[44] Jin, X., Gao, M., Li, D., & Zhao, T. (2024). Damage detection of road domain waveform guardrail structure based on machine learning multi-module fusion. Plos one, 19(3), e0299116.

[45] Zhao, M., Su, Y., Wang, J., Liu, X., Wang, K., Liu, Z., ... & Guo, Z. (2024). MED-YOLOv8s: a new real-time road crack, pothole, and patch detection model. Journal of Real-Time Image Processing, 21(2), 26.

[46] Elghaish, F., Matarneh, S. T., Talebi, S., Abu-Samra, S., Salimi, G., & Rausch, C. (2022). Deep learning for detecting distresses in buildings and pavements: a critical gap analysis. Construction Innovation, 22(3), 554-579.

[47] Ha, J., Kim, D., & Kim, M. (2022). Assessing severity of road cracks using deep learning-based segmentation and detection. The Journal of Supercomputing, 78(16), 17721-17735.

Real-Time Sign Language Fingerspelling Recognition System Using 2D Deep CNN with Two-Stream Feature Extraction Approach

Aziza Zhidebayeva[1], Gulira Nurmukhanbetova[2], Sapargali Aldeshov[3],
Kamshat Zhamalova[4], Satmyrza Mamikov[5], Nursaule Torebay[6]
University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhsatan[1, 5]
South Kazakhstan Pedagogical University named after Ozbekali Zhanibekov, Shymkent, Kazakhstan[2, 3, 4]
Miras University, Shymkent, Kazakhstan[6]

***Abstract*—This research paper introduces a novel sign language recognition system developed using advanced deep learning (DL) techniques aimed at enhancing communication capabilities between deaf and hearing individuals. The system leverages a convolutional neural network (CNN) architecture, optimized for the real-time interpretation of dynamic hand gestures that constitute sign language. A comprehensive dataset was employed to train and validate the model, encompassing a diverse range of gestures across different environmental settings. Comparative analysis revealed that the deep learning-based model significantly outperforms traditional machine learning techniques in terms of recognition accuracy, particularly with the increase in the volume of training data. This was illustrated through various performance metrics, including a detailed confusion matrix and Levenshtein distance measurements, highlighting the system's efficacy in accurately identifying complex gestures. Real-time application tests further demonstrated the model's robustness and adaptability to varying lighting conditions and backgrounds, essential for practical deployment. Key challenges identified include the need for broader linguistic diversity in training datasets and enhanced model sensitivity to subtle gestural distinctions. The paper concludes with suggestions for future research directions, emphasizing algorithm optimization, data diversification, and user-centric design improvements to foster wider adoption and usability. This study underscores the potential of deep learning technologies to revolutionize assistive communication tools, making them more accessible and effective for the deaf community.**

***Keywords*—Deep learning; sign language recognition; convolutional neural networks; real-time processing; gesture recognition; machine learning; accessibility technology**

I. Introduction

Fingerspelling is a critical component of sign languages, used primarily for spelling out words that do not have established signs, such as proper nouns and technical terms. This aspect of sign language communication has drawn significant attention in the realm of automated recognition systems, driven by the potential to bridge communication gaps between deaf and hearing communities. The development of a real-time fingerspelling recognition system using two-dimensional deep convolutional neural networks (2D Deep CNNs) represents a significant leap toward inclusive communication technologies. This paper aims to develop a real-time recognition system that utilizes advanced deep learning methodologies to accurately recognize fingerspelling from video inputs in sign language.

The importance of addressing the nuances in sign language through technological solutions cannot be overstated. Sign languages, unlike spoken languages, utilize manual communication and body language to convey meaning, incorporating a complex combination of hand shapes, orientations, movements, and facial expressions [1]. Fingerspelling is an integral component, especially in educational settings and daily communication, where specific terminology and names need clear articulation [2]. However, the manual and non-manual components of sign language pose unique challenges in automatic recognition, which must be addressed to achieve high accuracy and real-time performance [3].

Recent advancements in deep learning have shown promising results in image and video recognition tasks, which are pivotal in interpreting dynamic and complex gestures in sign language [4]. Particularly, the application of 2D Deep CNNs has emerged as a potent approach due to their ability to extract spatial hierarchies of features from visual data [5]. These networks have been effectively employed in various domains, including facial recognition and autonomous driving, underscoring their versatility and robustness in handling complex visual data [6].

The concept of using a two-stream feature extraction approach in our system is inspired by the successes seen in action recognition in videos, where separate streams are used to capture spatial and temporal features [7]. In the context of fingerspelling, one stream processes static images to recognize hand shapes, while the other captures the motion between frames to understand the dynamics of hand movements [8]. This dual approach is designed to enhance the system's ability to discern subtle differences in fingerspelling gestures, which are often rapid and involve minimal but significant movements.

However, the challenge in developing such systems is not only technical but also linguistic. Sign languages are not universal; thus, a model trained on one sign language may not be applicable to another [9]. This diversity necessitates adaptable models that can learn from limited data and generalize well across different languages and even dialects within the same language [10]. Additionally, the environmental variability in video data, such as background complexity, lighting conditions, and camera angles, also affects the performance of recognition systems [11].

Another critical aspect is the real-time capability of the system, which is essential for practical applications. For users to adopt such a technology effectively, the recognition process must be fast enough to occur in natural conversation time without significant delays [12]. Achieving this requires not only robust model architecture but also optimized computation strategies to process video frames swiftly [13].

In this study, we propose a system architecture that incorporates a streamlined 2D Deep CNN with a two-stream feature extraction strategy tailored for real-time application. Previous research has indicated the feasibility of real-time processing using deep learning models, particularly those optimized for mobile and embedded systems [14]. By integrating such models with a two-stream approach, we aim to achieve a balance between accuracy and speed, making the system practical for everyday use [15].

Furthermore, the development of such systems also opens avenues for enhanced educational tools, accessible services, and improved autonomy for the deaf and hard-of-hearing community [16]. As technology progresses, the integration of such specialized communication tools could profoundly impact social inclusion and equality.

This paper explores the technical development of the proposed recognition system, evaluates its performance across various metrics, and discusses its potential applications and implications for the future of communication technologies in the context of sign language. By pushing the boundaries of what is possible in automated fingerspelling recognition, we aim to contribute to a more inclusive and accessible technological landscape.

II. Related Work

The development of fingerspelling recognition systems using computer vision and machine learning technologies has garnered considerable attention in the academic community. This section reviews existing literature related to the application of deep learning techniques for sign language recognition, with a specific focus on fingerspelling, feature extraction methodologies, real-time processing capabilities, and the challenges posed by diverse sign languages.

A. 2D and 3D Convolutional Neural Networks in Sign Language Recognition

Recent studies have extensively employed convolutional neural networks (CNNs) for the task of sign language recognition. The application of 2D CNNs has proven effective in recognizing static sign language images, capturing spatial features that distinguish various hand signs [17]. However, the dynamic nature of sign language, particularly in fingerspelling, requires understanding temporal sequences, for which 3D CNNs are better suited. These networks extend the capability of 2D CNNs by adding time as a third dimension, allowing them to capture motion across frames effectively [18]. A notable study demonstrated the superiority of 3D CNNs over their 2D counterparts in recognizing continuous sign language gestures, attributing improvements to the network's ability to process temporal information [19]. Nevertheless, the computational demand of 3D CNNs remains a significant challenge, particularly for real-time applications [20].

B. Feature Extraction Techniques for Enhanced Gesture Recognition

The effectiveness of a recognition system largely depends on the robustness of its feature extraction process. In this context, two-stream CNN architectures have shown promising results by separately processing spatial and temporal features, thus providing a more comprehensive analysis of video data [21]. One stream typically processes individual frames to capture static features like hand shapes and positions, while the other analyzes motion between frames to capture dynamic movements [22]. Such approaches have been applied successfully in other fields of action recognition and have gradually been adapted for sign language recognition [23]. Hybrid models that combine CNNs with recurrent neural networks (RNNs) have also been explored, with RNNs processing the temporal sequences of features extracted by CNNs, thereby enhancing the recognition of gestures over time [24]. Moreover, attention mechanisms have been integrated to focus the model on relevant features of the hand, significantly improving accuracy by reducing the influence of background noise and other irrelevant signals [25].

C. Real-Time Processing for Sign Language Recognition Systems

Achieving real-time processing capabilities in sign language recognition systems is crucial for their practical application. The latency in processing and recognizing sign language must be minimized to facilitate fluid communication between deaf and hearing individuals. Several studies have focused on optimizing CNN architectures to reduce computational loads without compromising accuracy [26]. Techniques such as model pruning, quantization, and the use of efficient network architectures like MobileNets have been proposed as solutions to achieve faster processing times [27]. Furthermore, edge computing has emerged as a viable approach, where processing is done on local devices rather than relying on cloud-based systems, thereby reducing response times significantly [28]. These advancements have paved the way for the development of more responsive and efficient real-time sign language recognition systems [29].

D. Challenges in Multilingual Sign Language Recognition

Sign language recognition is further complicated by the variation in sign languages across different regions and cultures. Each sign language has its own set of rules and nuances, which means that a system trained on one language may not perform well on another [30]. The scarcity of annotated datasets for many sign languages poses a significant barrier to training robust models [31]. Studies have attempted to address these challenges by using transfer learning, where a model trained on one language is adapted to another with minimal additional training [32]. Another approach is the use of synthetic data generation to augment existing datasets, thereby providing more comprehensive training material [33]. These methods have shown some success, but the variability in performance across languages remains a concern [34].

The literature reviewed highlights significant advancements in the field of sign language recognition, particularly in applying deep learning techniques for fingerspelling recognition. While 2D and 3D CNNs offer robust frameworks for feature extraction, their integration with two-stream architectures and RNNs presents a promising path toward more accurate and efficient recognition systems. Real-time processing remains a critical area for ongoing research, with current solutions pointing towards optimized CNN architectures and edge computing. However, the multilingual and multicultural nature of sign languages continues to pose significant challenges, necessitating further research into adaptive and scalable models that can handle the diversity of sign languages globally.

III. Materials and Methods

This section is critical as it outlines the systematic steps taken to ensure the reliability and validity of the results obtained. It serves to offer transparency, allowing other researchers to replicate the study or build upon its findings. Within this section, we detail the specific datasets used, the data preprocessing techniques employed, the architectural design of the machine learning models, and the criteria for evaluating their performance. By providing a clear and thorough exposition of these elements, we aim to facilitate a deeper understanding of the research process and its foundational components.

A. *Sign Language Alphabets*

Sign language alphabets serve as fundamental building blocks for communication within deaf communities, providing a means to spell out words and names for which specific signs may not exist. Among the various sign languages utilized globally, American Sign Language (ASL) [35] and Indian Sign Language (ISL) [36] represent two distinct systems, each with its unique set of alphabetic representations. This section delves into the alphabetic systems of ASL and ISL, illustrating their characteristics and the cultural nuances that influence their formation and usage.

1) American Sign Language (ASL) Alphabet: American Sign Language (ASL) is one of the most widely used sign languages in the world, particularly prevalent in the United States and parts of Canada. The ASL alphabet, as depicted in Fig. 1, consists of a series of hand configurations used to represent the 26 letters of the English alphabet. Each letter is formed using one hand, which is a notable characteristic that distinguishes ASL from some other sign languages that might use two hands for certain letters. The ASL fingerspelling system is crucial for expressing proper nouns, technical terms, and any other words for which there is no established sign, thus playing a vital role in daily communication as well as educational settings [37].

The ASL alphabet's design emphasizes clarity and simplicity, allowing for quick and straightforward communication. The letters are generally formed in front of the signer at chest level, ensuring visibility and ease of understanding. For instance, the letters 'A' through 'Z' involve distinct positions and shapes of the fingers, with minimal movement, making them relatively easy to learn for beginners and highly functional for fluent users in rapid communication.

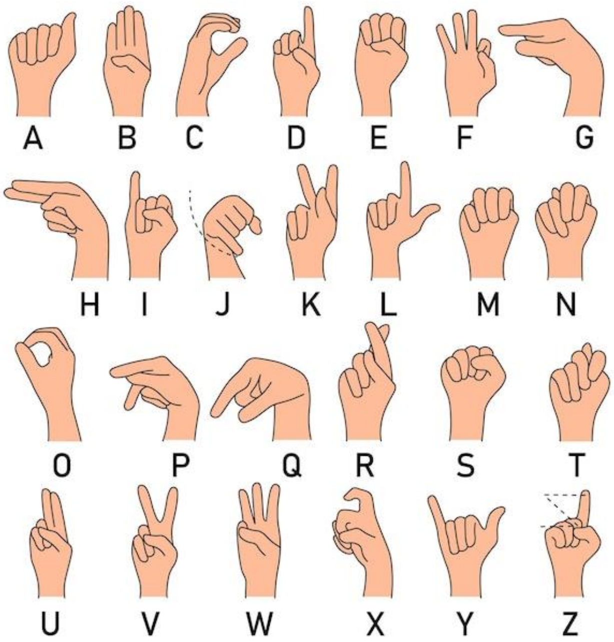


Fig. 1. ASL Alphabet.

2) Indian Sign Language (ISL) Alphabet: In contrast, Indian Sign Language (ISL) caters to the diverse linguistic landscape of India and incorporates elements that reflect the cultural and regional diversity of the country. The ISL alphabet, shown in Fig. 2, is utilized across various educational and social settings in India, providing a means for the deaf community to partake in both public and private discourse. Unlike ASL, the ISL fingerspelling system often employs two hands to represent certain letters, which can be seen as an adaptation to the linguistic structures and phonetic complexities of the multiple languages spoken in India [38].

Each letter in the ISL alphabet is represented by a unique combination of hand shapes, positions, and movements. These elements are designed to be visually distinct from one another to minimize confusion and ensure effective communication. For instance, the letters of the ISL alphabet are depicted with both static and dynamic gestures, which involve more interaction between the two hands compared to the mostly static nature of ASL fingerspelling [39]. This characteristic of ISL may stem from the gestural nuances found in the native languages of India, which often emphasize expressive hand movements and gestures in daily communication.

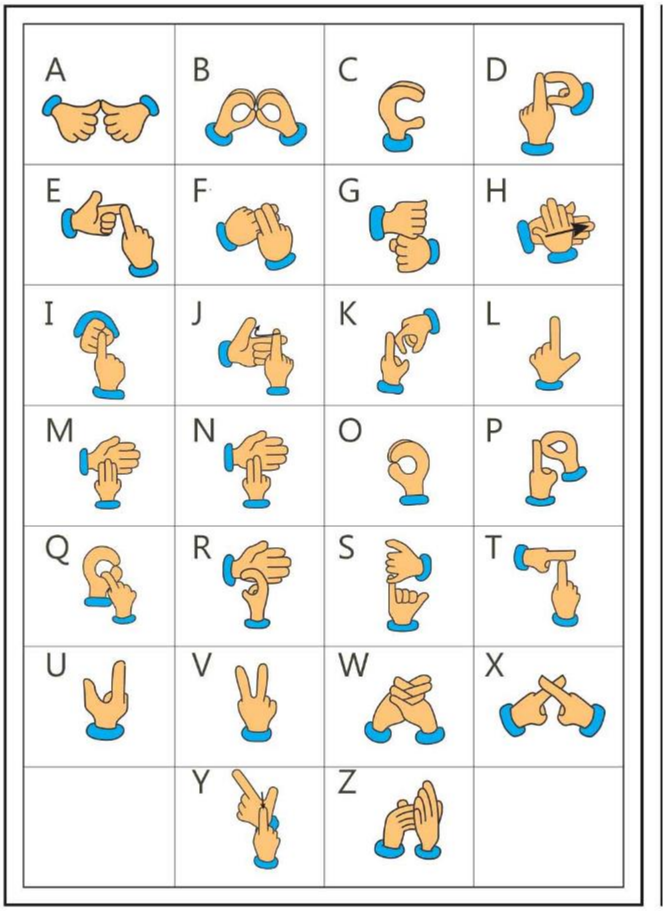


Fig. 2. ISL Alphabet.

The alphabets of ASL and ISL highlight the adaptability and diversity of sign languages in accommodating the linguistic needs of different cultural contexts [40-42]. While ASL employs a one-handed system for simplicity and speed, ISL uses a two-handed approach, possibly reflecting the complex phonetic systems of India's numerous spoken languages. Understanding these differences is crucial for educators, linguists, and technologists who develop communication tools and educational materials for the deaf community. The study and comparison of such systems not only enhance our understanding of linguistic diversity but also promote more inclusive and effective communication tools tailored to the unique needs of each sign language community.

B. Data Structure and Sample Description

Fig. 3 illustrates the data structure used in the research project for training the sign language recognition model. The data organization is key to understanding the relational dynamics between different datasets, which include train.csv and 5414471.parquet. The diagram effectively shows how these files are interlinked and utilized to train the deep learning model, highlighting the integration of participant information, sequence IDs, and feature vectors extracted from video frames.

1) Train.csv**:** The train.csv file acts as a central index, containing essential metadata for each training sample. This file includes several columns:

- path: Specifies the location of the parquet file containing the detailed sequence data.
- file_id: A unique identifier for the parquet file.
- sequence_id: A unique identifier that links the train.csv entries with specific sequences in the 5414471.parquet file.
- participant_id: Identifies the participant from whom the data was collected, facilitating analysis on a per-subject basis if required.
- phrase: Represents the specific phrase or words being signed in the sequence, which is crucial for supervised learning where the model learns to associate specific gestures with their corresponding linguistic outputs.

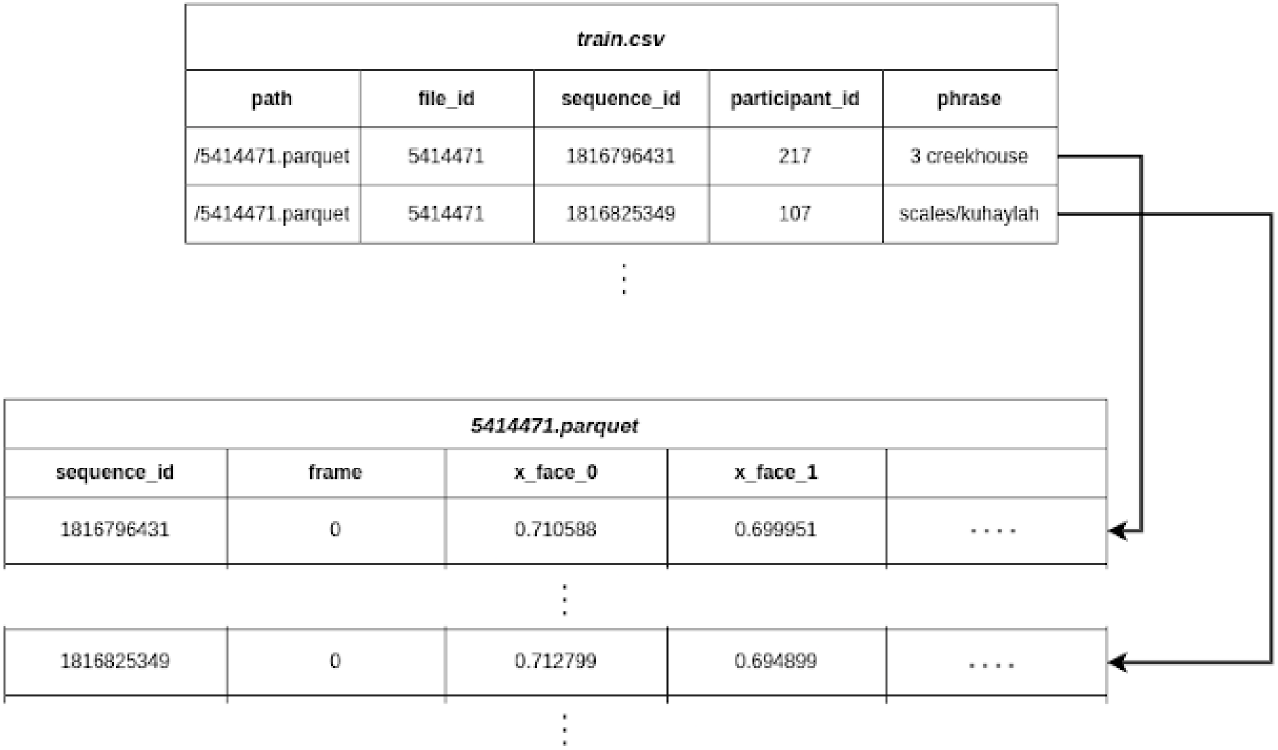
2) 5414471.parquet: The 5414471.parquet file contains detailed frame-by-frame data extracted from video sequences of participants performing sign language. Each row in this file corresponds to a single frame from a video sequence, and is linked back to the train.csv via the sequence_id. The columns in this file include:

- sequence_id: Matches the sequence_id in train.csv, establishing a relational link.
- frame: The frame number within a particular video sequence, which is critical for analyzing the temporal progression of gestures.
- x_face_0, x_face_1, ...: These columns represent extracted feature vectors associated with each frame. The features might include positional data of different facial landmarks or other relevant metrics that are used as input for the deep learning model.

The diagram in Fig. 3 demonstrates the workflow from raw video data extraction through to the feature extraction process, ending with the data being formatted into a machine-readable structure for model training. This structure supports the development of a robust model by providing a comprehensive dataset that includes both the static context of the sign language phrases and dynamic motion information encapsulated in the sequence of frames. This detailed and methodical data organization ensures that the machine learning model can learn from a rich dataset that mimics the complexities of real-world sign language usage.

C. Proposed Model Architecture

The proposed model architecture for sign language recognition, illustrated in Fig. 4, is designed to process sequential image data through a series of convolutional and fully connected layers. This architecture harnesses the power of deep learning to effectively capture both spatial and temporal features critical for understanding dynamic sign language gestures. Below, we describe each component of the model as depicted in the figure, and detail the operations performed at each stage.



train.csv

path	file_id	sequence_id	participant_id	phrase
/5414471.parquet	5414471	1816796431	217	3 creekhouse
/5414471.parquet	5414471	1816825349	107	scales/kuhaylah

⋮

5414471.parquet

sequence_id	frame	x_face_0	x_face_1	
1816796431	0	0.710588	0.699951	. . . .
⋮				
1816825349	0	0.712799	0.694899	. . . .
⋮				

Fig. 3. Example of training sample.

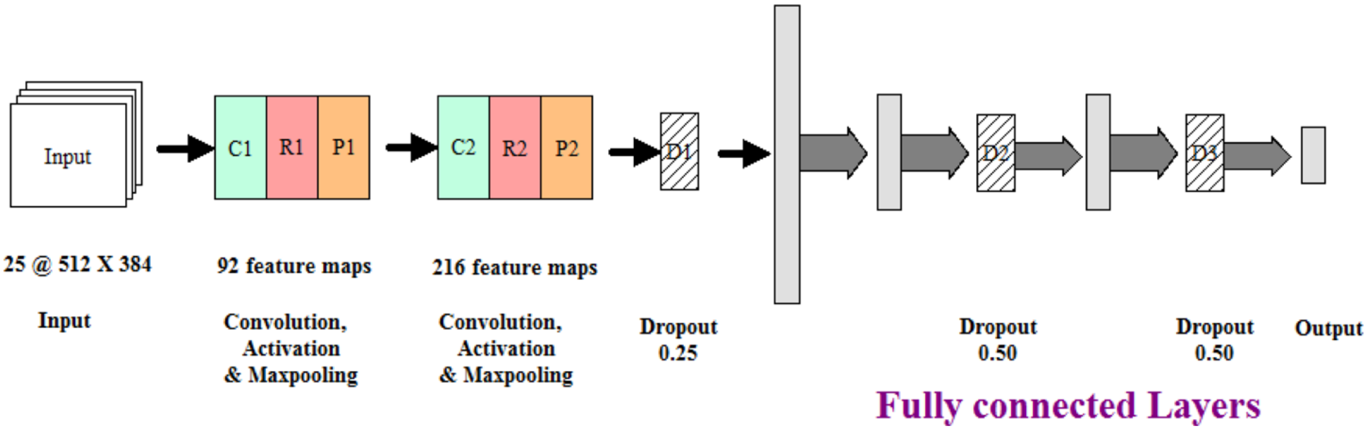


Fig. 4. The proposed model.

1) Input Layer: The input layer accepts sequential image data with dimensions 25×512×384, where 25 represents the number of frames in the sequence, and 512×384 are the pixel dimensions of each frame. This three-dimensional input is essential for preserving the spatial and temporal information present in video data.

Convolutional Layers (C1, C2). The first part of the model consists of two convolutional layers labeled as C1 and C2. These layers are responsible for extracting high-level features from the input images through the application of filters.

C1 Layer: Applies 92 filters to the input images, generating 92 feature maps. The convolution operation can be represented by the equation:

$$F_{ij}^{(k)} = \sigma\left(\sum_{m,n} I_{i+m,j+n} \cdot K_{mn}^{(k)} + b_k\right) \tag{1}$$

Here, $F_{ij}^{(k)}$ is the feature map for the k-th filter at position $(i, j), \sigma$ is the nonlinear activation function (typically ReLU), I is the input matrix, $K^{(k)}$ is the k-th filter matrix, and b_k is the bias term.

C2 Layer: Further refines the features extracted by the first layer by applying additional 216 filters, thus producing 216 feature maps. This layer helps in capturing more complex features that are vital for accurate sign language recognition.

2) Activation and Pooling Layers (R1, P1, R2, P2): After each convolutional layer, the model applies an activation function followed by a max pooling operation:

Activation (ReLU): Enhances non-linearity in the model by applying the ReLU activation function, which helps in handling non-linear features efficiently [43].

Max Pooling: Reduces the spatial dimensions of the feature maps while retaining the most significant information. This operation is critical for reducing the computational complexity and improving the robustness of the model against small variations in the input data.

3) Dropout layers: Following the convolutional blocks, two dropout layers are included to prevent overfitting [44]:

Dropout 0.25: Applied after the first convolutional block, this layer randomly sets a fraction of input units to 0 at each update during training time, which helps in making the model more generalized.

Dropout 0.50: A higher dropout rate is used after the second convolutional block to further regularize the model, particularly important due to the increased complexity from more feature maps.

4) Fully connected layers: The model transitions from convolutional layers to fully connected (dense) layers, which are essential for making predictions [45]. The dense layers integrate the learned features from previous layers to form the final output. The sequence of fully connected layers ends in a softmax or sigmoid output layer (depending on the number of classes), which provides the probabilities of each sign language gesture.

5) Output Layer: The final layer of the model uses a softmax activation function to classify the input image sequence into one of the possible sign language gestures. The softmax function is given by [46]:

$$P(y = j \mid x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{2}$$

Where $P(y = j \mid x)$ is the probability that the input x belongs to class j, z_j is the input to the softmax function from the final fully connected layer for class j, and K is the total number of classes.

The architecture proposed in Fig. 4 is designed to be robust, efficient, and capable of handling the complexities associated with recognizing sign language from video sequences. The combination of convolutional layers for feature extraction and fully connected layers for classification forms a powerful model that is well-suited for the real-time interpretation of sign language.

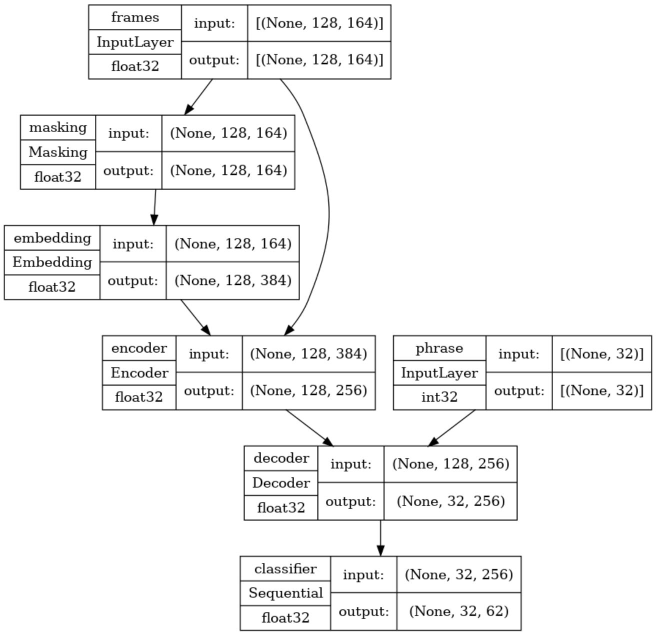


Fig. 5. Flowchart of the proposed model.

Fig. 5 presents a flowchart of a complex deep learning model architecture, primarily designed for sequence processing in tasks such as sign language recognition. This model integrates multiple layers and operations to handle sequential input data effectively. The architecture begins with an InputLayer that accepts frames, formatted as a three-dimensional array [None, 128, 164], which then passes through a Masking layer to ignore certain types of input (e.g., padding or missing values) for the purpose of maintaining the integrity of the sequence processing. Subsequently, the data undergoes transformation in an Embedding layer, which expands the feature representation to 384 dimensions, aiding in richer feature extraction by the subsequent Encoder layer. This encoder processes the embedded input and outputs a condensed representation [None, 128, 256] to the Decoder, which aims to reconstruct or transform the sequence contextually, often used in language translation or similar tasks. Alongside, a separate InputLayer for phrases processes integer-encoded inputs, suggesting a possible multimodal approach combining textual and sequential input data. The decoder's output then feeds into a Sequential classifier, which finalizes the processing pipeline by generating predictions or classifications based on the learned features, outputting a processed signal [None, 32, 62]. This architecture indicates a sophisticated approach to handling complex patterns in sequence data, suitable for tasks requiring nuanced understanding of temporal dynamics and contextual dependencies.

IV. Experimental Results

Fig. 6 illustrates the key anatomical landmarks, or keypoints, used in the study for tracking and recognizing fingerspelling gestures in sign language. The diagram depicts a schematic of a human hand annotated with 21 distinct keypoints, each corresponding to critical joints and segments within the hand's structure. These keypoints include the wrist, the carpometacarpal (CMC), metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP) joints of each finger, and the tips of the fingers.

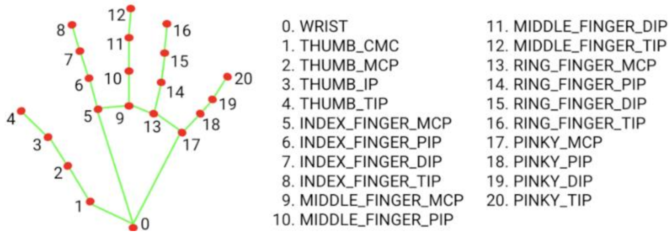


Fig. 6. Keypoints on the hand for fingerspelling.

The keypoints are numbered from 0 to 20, starting from the wrist and moving outward towards the fingertips. For instance, keypoint 0 represents the wrist, keypoint 1 the thumb CMC, keypoint 5 the index finger MCP, and so forth, culminating with keypoint 20 at the tip of the pinky finger. These annotations are crucial for machine learning models, which rely on these precisely defined points to accurately interpret and classify the gestures involved in fingerspelling.

The connections between the keypoints, represented by green lines, indicate the typical kinematic chains in hand anatomy essential for motion tracking and gesture recognition. These lines help in understanding how movements at one joint affect subsequent parts of the hand, which is vital for developing algorithms that can accurately interpret complex hand gestures. The clear labeling and structuring of these keypoints in the diagram provide a foundation for detailed analysis and discussion of the results related to fingerspelling recognition accuracy in the subsequent sections of the document.

Fig. 7 provides a detailed representation of the upper extremity keypoints utilized in the fingerspelling recognition algorithm, specifically highlighting the arrangement and connectivity of key anatomical landmarks across the fingers and wrist. This diagram focuses on the joints of the fingers and the wrist, essential for deciphering the precise configuration of hand gestures in sign language communication.

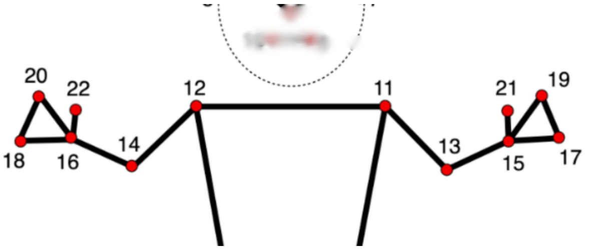


Fig. 7. Keypoints on the human body for fingerspelling.

In this schematic, keypoints are strategically positioned on the joints and tips of the fingers to capture the essential articulations necessary for sign language interpretation. The keypoints are numbered from 12 to 22, illustrating an array from the base of the wrist up through the tips of the fingers. Notably, keypoints 12 and 14 signify the wrist and the base of the fingers, respectively, forming a critical juncture from which the digital keypoints extend.

Each finger is represented by a sequence of keypoints, with the numbering extending outward towards the fingertips: the index finger from keypoint 14 to 16, the middle finger from 11 to 13, and so forth, with the additional articulations for more complex gestures indicated by keypoints 17, 19, 21, and 22. The lines connecting these points indicate the kinematic links that are essential for understanding how movements in one part of the hand influence the positioning and orientation of the other parts.

This configuration allows the machine learning models to track and interpret the dynamic and complex movements involved in sign language gestures, providing a robust framework for accurate gesture recognition. The clarity and layout of these keypoints are pivotal for analyzing the effectiveness of the fingerspelling recognition system, which is further explored in the results section of the study.

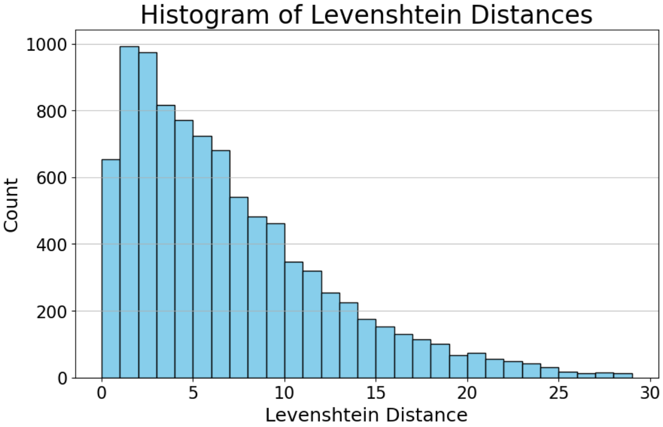


Fig. 8. Histogram of Levenshtain distances.

Fig. 8 presents a histogram of Levenshtein distances calculated from the output of the fingerspelling recognition system. The Levenshtein distance, a metric for measuring the difference between two sequences, is used here to evaluate the discrepancy between the predicted and actual spelled sequences in sign language communication. The x-axis of the histogram represents the Levenshtein distance, ranging from 0 to 30, while the y-axis displays the count of sequences falling into each distance category [47].

The distribution depicted in the histogram is skewed to the left, indicating that a majority of the sequence predictions by the model have a relatively low discrepancy from the target sequences, with the highest frequency observed in the range of 0 to 5. This suggests that the model is generally effective in accurately predicting the sequences, though errors increase as the distance values rise. The diminishing frequency as the distance increases confirms that fewer instances have larger errors, highlighting the effectiveness of the model in capturing the nuances of fingerspelling gestures with a considerable degree of accuracy. The shape and spread of the distribution provide crucial insights into the performance of the recognition system, revealing both its strengths in accurately recognizing many gestures and the areas where improvements might be necessary for those predictions exhibiting higher discrepancies.

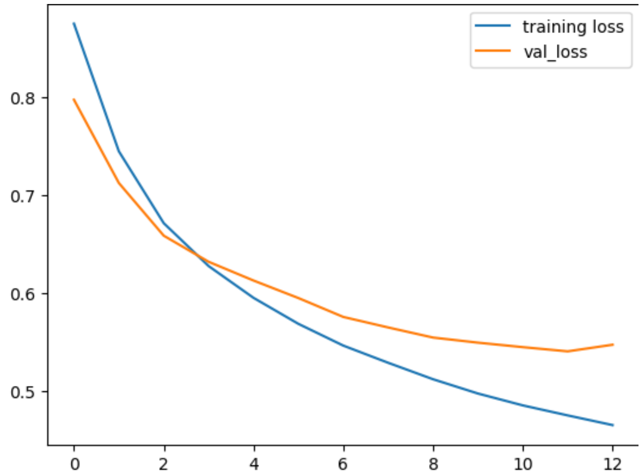
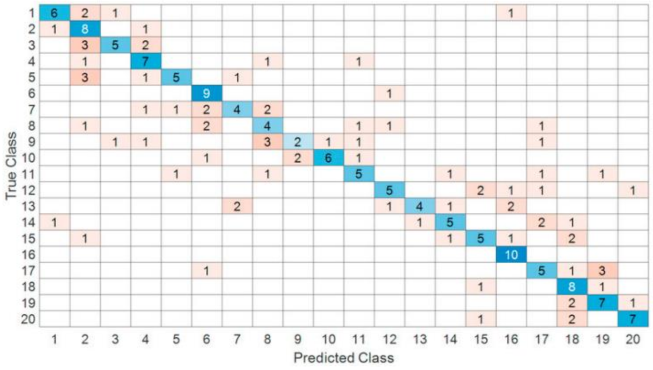


Fig. 9. Training and validation loss.

Fig. 9 displays a graphical representation of the training and validation loss curves for the sign language recognition model over a series of epochs. The x-axis of the graph indicates the number of epochs, while the y-axis represents the loss value, which quantifies the difference between the predicted outputs of the model and the actual target values during training and validation phases.

The blue line represents the training loss, illustrating how the model's error on the training set decreases as it learns from the data over successive epochs. The orange line, representing the validation loss, shows a similar decrease, indicating how well the model generalizes to new, unseen data. Both curves exhibit a steep decline in the initial epochs, signifying rapid learning and improvement in model performance.

As the number of epochs increases, both curves begin to plateau, suggesting that the model is approaching its optimal performance given the current architecture and data. The convergence of the training and validation loss lines toward the latter epochs indicates good model generalization without significant overfitting. This convergence is crucial for confirming that the model is not merely memorizing the training data but rather learning generalizable patterns that perform well on external data. The graph effectively underscores the learning dynamics of the model, highlighting areas where the training process is stable and effective, alongside pointing out the epochs after which learning saturates.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	6	2	1													1				
2	1	8		1																
3		3	5	2																
4		1		7				1			1									
5		3		1	5		1													
6						9						1								
7				1	1	2	4	2												
8		1				2		4			1	1					1			
9			1	1				3	2	1	1						1			
10						1			2	6	1									
11					1			1			5			1			1		1	
12												5			2	1	1			1
13							2					1	4	1		2				
14	1												1	5			2	1		
15		1												1	5	1		2		
16																10				
17						1											5	1	3	
18															1			8	1	
19																		2	7	1
20															1			2		7

Fig. 10. Confusion matrix results.

Fig. 10 depicts a confusion matrix, a critical tool for evaluating the performance of the classification model developed for fingerspelling recognition. This matrix presents the counts of actual versus predicted class labels across a set of 20 classes, which represent different fingerspelling gestures.

The x-axis of the confusion matrix corresponds to the predicted class labels by the model, ranging from 1 to 20, while the y-axis represents the true class labels. Each cell within the matrix shows the number of instances that the model predicted a certain class (x-axis) for an actual class (y-axis). The diagonal cells, highlighted by darker shades, represent correct predictions where the predicted classes match the true classes. Off-diagonal cells indicate misclassifications, where the numbers denote how often a particular class was predicted instead of another.

A quick visual assessment of the matrix reveals several insights:

- The concentration of higher numbers along the diagonal line indicates good model accuracy for many classes, with prominent correctly classified instances such as in classes 1, 6, 7, and 9.
- Some classes, however, show notable confusion with others. For example, class 20 exhibits frequent misclassification, with incorrect predictions scattered across several other classes.
- Certain pairs of classes, such as 10 and 20 or 5 and 19, have higher confusion, suggesting similarities in the gestures that may be leading to these consistent misclassifications.

Overall, the confusion matrix provides a detailed view of the model's strengths and weaknesses across different fingerspelling gestures, highlighting specific areas where the model performs well and others where improvement is necessary. This visual tool is indispensable for diagnosing performance issues and guiding future enhancements to the model's classification capabilities.

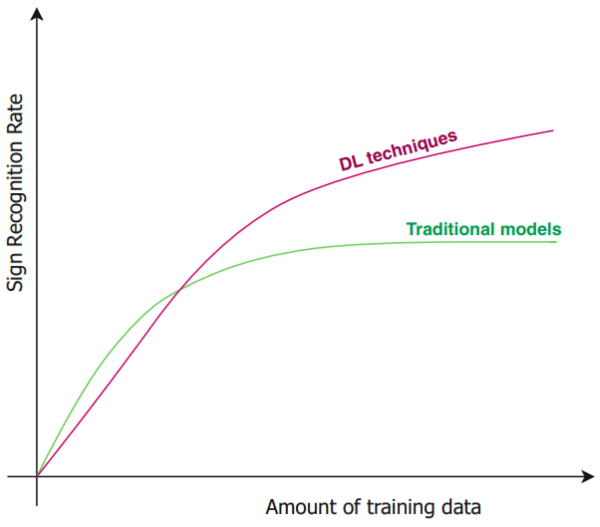


Fig. 11. Comparative performance of the proposed deep model vs. traditional models in sign language recognition as a function of training data volume.

Fig. 11 illustrates a comparative analysis of the performance between traditional machine learning models and deep learning (DL) techniques in the context of sign language recognition as a function of the amount of training data used. The graph plots the sign recognition rate on the vertical axis against the amount of training data on the horizontal axis. As depicted, both curves exhibit an increase in recognition rate with more data, demonstrating the typical behavior that more extensive training datasets generally improve the accuracy of predictive models. However, the curve representing deep learning techniques (colored in pink) is positioned above that of traditional models (colored in green), indicating a consistently higher recognition rate across the spectrum of data volumes. Notably, the deep learning curve shows a steeper initial ascent, suggesting that DL techniques are more effective at leveraging larger datasets to achieve significant improvements in accuracy [48]. This trend highlights the scalability and robustness of deep learning models in handling complex, high-dimensional data typical of sign language video inputs, compared to traditional models which plateau earlier and achieve lower peak recognition rates.

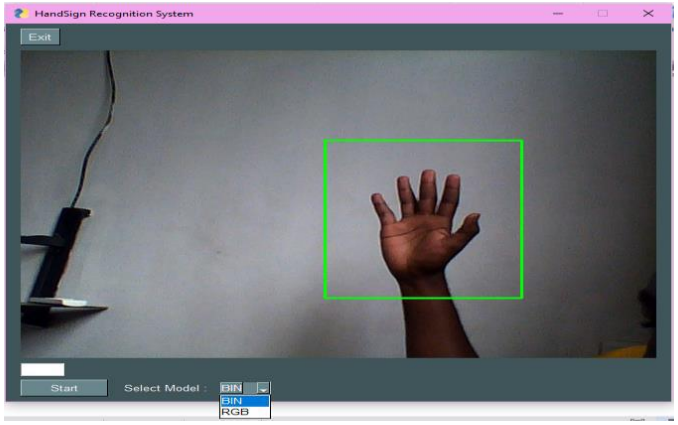


Fig. 12. Comparative performance of the proposed deep model vs. traditional models in sign language recognition as a function of training data volume.

Fig. 12 displays a screenshot of the HandSign Recognition System interface during operation, showcasing the system's ability to detect and highlight a hand gesture within a real-time video feed. The image illustrates a human hand positioned centrally against a plain background, with the hand's open gesture enclosed by a green bounding box, indicating successful detection by the system. This visual feedback is part of the user interface designed to allow users to verify the correct identification and tracking of hand gestures. The interface includes several operational controls such as “Start”, “Select Model”, along with options to switch between binary (“BIN”) and RGB color modes, enhancing the flexibility and usability of the system for different lighting conditions and background scenarios. The inclusion of such features underscores the system's practical application in real-world environments, where variability in input conditions can significantly affect performance.

V. Discussion

This research paper has examined the development and application of a deep learning-based sign language recognition system, which is instrumental in bridging communication gaps between the deaf and the hearing. The discussion elaborates on the implications of the findings, explores the challenges encountered, and suggests future research directions.

A. Efficacy of Deep Learning Models

The results demonstrated that deep learning techniques outperform traditional models in sign language recognition, particularly as the volume of training data increases. As depicted in Fig. 11, deep learning models exhibit a steeper improvement curve in recognition accuracy with the addition of training data. This can be attributed to the ability of deep learning models to extract complex features from high-dimensional data, a capability that traditional machine learning models lack. The advanced feature extraction allows for a more nuanced understanding of sign language gestures, which are inherently complex due to the variety of hand shapes, orientations, and movements involved [49].

B. System Performance in Real-Time Applications

Fig. 12 illustrates the practical application of the system in real-time environments. The system's capability to accurately detect and track hand gestures in real time underscores its potential for use in dynamic scenarios, such as live sign language translation or interactive educational tools. However, the performance in real-world conditions can be affected by factors such as lighting variations, background noise, and rapid movements. Although the current system handles these challenges to a certain extent, further refinement is necessary to improve robustness and ensure consistent performance regardless of external conditions.

C. Integration and Usability Challenges

The user interface, as shown in Fig. 12, is designed to be intuitive and user-friendly, offering essential controls like model selection and color mode adjustments. This is critical for ensuring that the system is accessible not only to researchers and professionals but also to lay users, including educators and individuals within the deaf community. However, the integration of such technology into everyday applications presents challenges, including the need for compatible hardware, user training, and ongoing support. Furthermore, there remains a significant need to develop standardized protocols for evaluating the usability of such systems in diverse settings [50].

D. Limitations and Scope for Improvement

While the system shows promising results, there are several limitations that need to be addressed. The current dataset, although extensive, may not fully represent the diversity within the global deaf community [51]. Sign language varies significantly not just internationally but also regionally; therefore, the system’s training on a more culturally and linguistically diverse dataset could enhance its applicability. Moreover, the confusion matrix in Fig. 10 reveals specific areas where the model confuses similar gestures. This could be mitigated by introducing more granular features and perhaps a temporal component to better differentiate between dynamically similar signs.

E. Future Directions

Looking forward, the research should focus on several key areas:

1) Data Diversification: Collecting and incorporating more diverse training data that cover a broader spectrum of sign languages and include more varied environments and lighting conditions.

2) Algorithm Optimization: Enhancing the model's architecture to improve its ability to learn from fewer data points, which is crucial for rare gestures or signs.

3) Real-Time processing improvements: Reducing latency further and increasing the processing speed to handle rapid sequences of gestures without delay.

4) User-Centric Design: Engaging with the deaf community to tailor the system's development to their needs and preferences, ensuring that the technology is both accessible and practical.

5) Cross-Platform compatibility: Ensuring the system is adaptable to various devices and platforms, enhancing its accessibility and practical utility.

The development of a sign language recognition system using deep learning techniques represents a significant technological advancement with the potential to impact real-world interactions profoundly. By continuously refining the system and addressing the outlined challenges, future iterations can provide even more reliable and inclusive communication tools for the deaf and hard-of-hearing communities.

VI. Conclusion

The research undertaken in this study has culminated in the development of an advanced sign language recognition system powered by deep learning techniques, showcasing significant potential to enhance communication between the deaf and hearing communities. The application of deep learning has been demonstrated to markedly outperform traditional models, especially as the volume of training data increases. This is evident in the system's enhanced ability to interpret complex hand gestures with high accuracy, addressing the dynamic and diverse nature of sign language. Our findings indicate that with sufficient training data, deep learning models can effectively capture the subtleties of sign language, which are often missed by more conventional approaches. The real-time operational capability of the system, as demonstrated, further underscores its practical utility in everyday applications, from educational settings to public services. However, challenges related to system integration, environmental variability, and data diversity call for ongoing improvements. Future research should aim to diversify the training datasets to include a broader array of sign languages and refine the system's robustness against external changes such as lighting and background variations. Engaging with the deaf community to tailor the technology to their needs will ensure that the advancements in sign language recognition technology are both practical and impactful. Ultimately, this research paves the way for creating more accessible and effective communication tools, fostering inclusivity and understanding across different sections of society.

References

[1] Shin, J., Miah, A. S. M., Akiba, Y., Hirooka, K., Hassan, N., & Hwang, Y. S. (2024). Korean Sign Language Alphabet Recognition through the Integration of Handcrafted and Deep Learning-Based Two-Stream Feature Extraction Approach. IEEE Access.

[2] Gao, Q., Ogenyi, U. E., Liu, J., Ju, Z., & Liu, H. (2020). A two-stream CNN framework for American sign language recognition based on multimodal data fusion. In Advances in Computational Intelligence Systems: Contributions Presented at the 19th UK Workshop on Computational Intelligence, September 4-6, 2019, Portsmouth, UK 19 (pp. 107-118). Springer International Publishing.

[3] Omarov, B., Batyrbekov, A., Suliman, A., Omarov, B., Sabdenbekov, Y., & Aknazarov, S. (2020, November). Electronic stethoscope for detecting heart abnormalities in athletes. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-5). IEEE.

[4] Luqman, H. (2022). An efficient two-stream network for isolated sign language recognition using accumulative video motion. IEEE Access, 10, 93785-93798.

[5] Yin, L., Ying, H., & Meng-hao, Y. (2023). Chinese sign language recognition based on two-stream CNN and LSTM network. International Journal of Advanced Networking and Applications, 14(6), 5666-5671.

[6] Rastgoo, R., Kiani, K., & Escalera, S. (2022). Real-time isolated hand sign language recognition using deep networks and SVD. Journal of Ambient Intelligence and Humanized Computing, 13(1), 591-611.

[7] Omarov, B., Narynov, S., & Zhumanov, Z. (2023). Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review. Computers, Materials & Continua, 74(3).

[8] Kumar, E. K., Kishore, P. V. V., Kumar, M. T. K., & Kumar, D. A. (2020). 3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2–stream CNN. Neurocomputing, 372, 40-54.

[9] Singla, N., Taneja, M., Goyal, N., & Jindal, R. (2023, March). Feature Fusion and Multi-Stream CNNs for ScaleAdaptive Multimodal Sign Language Recognition. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1266-1273). IEEE.

[10] Omarov, B., Altayeva, A., Demeuov, A., Tastanov, A., Kassymbekov, Z., & Koishybayev, A. (2020, December). Fuzzy controller for indoor air quality control: a sport complex case study. In International Conference on Advanced Informatics for Computing Research (pp. 53-61). Singapore: Springer Singapore.

[11] Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. IEEE Access, 9, 126917-126951.

[12] Ghadami, A., Taheri, A., & Meghdari, A. (2024). A Transformer-Based Multi-Stream Approach for Isolated Iranian Sign Language Recognition. arXiv preprint arXiv:2407.09544.

[13] Subburaj, S., & Murugavalli, S. (2022). Survey on sign language recognition in context of vision-based and deep learning. Measurement: Sensors, 23, 100385.

[14] da Silva, D. R., de Araújo, T. M. U., do Rêgo, T. G., Brandão, M. A. C., & Gonçalves, L. M. G. (2024). A multiple stream architecture for the recognition of signs in Brazilian sign language in the context of health. Multimedia Tools and Applications, 83(7), 19767-19785.

[15] Miah, A. S. M., Hasan, M. A. M., Nishimura, S., & Shin, J. (2024). Sign language recognition using graph and general deep neural network based on large scale dataset. IEEE Access.

[16] Robert, E. J., & Duraisamy, H. J. (2023). A review on computational methods based automated sign language recognition system for hearing and speech impaired community. Concurrency and Computation: Practice and Experience, 35(9), e7653.

[17] Tao, T., Zhao, Y., Liu, T., & Zhu, J. (2024). Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges. IEEE Access.

[18] Miah, A. S. M., Hasan, M. A. M., Jang, S. W., Lee, H. S., & Shin, J. (2023). Multi-stream graph-based deep neural networks for skeleton-based sign language recognition.

[19] Hamza, H. M., & Wali, A. (2023). Pakistan sign language recognition: leveraging deep learning models with limited dataset. Machine Vision and Applications, 34(5), 71.

[20] Patel, D. U., & Joshi, J. M. (2022). Deep leaning based static Indian-Gujarati Sign language gesture recognition. SN Computer Science, 3(5), 380.

[21] Bahia, N. K., & Rani, R. (2023). Multi-level taxonomy review for sign language recognition: Emphasis on indian sign language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(1), 1-39.

[22] Liu, T., Tao, T., Zhao, Y., Li, M., & Zhu, J. (2024). A signer-independent sign language recognition method for the single-frequency dataset. Neurocomputing, 582, 127479.

[23] Nihalani, R., Chouhan, S. S., Mittal, D., Vadula, J., Thakur, S., Chakraborty, S., ... & Saxena, A. (2024). Long Short-Term Memory (LSTM) model for Indian sign language recognition. Journal of Intelligent & Fuzzy Systems, (Preprint), 1-19.

[24] Chroni, E. T. (2024). Skeleton based approaches for isolated sign language recognition (Doctoral dissertation, Rutgers University-School of Graduate Studies).

[25] Bhaumik, G., Verma, M., Govil, M. C., & Vipparthi, S. K. (2022). ExtriDeNet: an intensive feature extrication deep network for hand gesture recognition. The Visual Computer, 38(11), 3853-3866.

[26] Doskarayev, B., Omarov, N., Omarov, B., Ismagulova, Z., Kozhamkulova, Z., Nurlybaeva, E., & Kasimova, G. (2023). Development of Computer Vision-enabled Augmented Reality Games to Increase Motivation for Sports. International Journal of Advanced Computer Science and Applications, 14(4).

[27] Xu, F., Chaudhary, L., Dong, L., Setlur, S., Govindaraju, V., & Nwogu, I. (2024, May). A Comparative Study of Video-Based Human Representations for American Sign Language Alphabet Generation. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) (pp. 1-6). IEEE.

[28] Sreemathy, R., Jagdale, J., Sayed, A. A., Ramteke, S. H., Naqvi, S. F., & Kangune, A. (2023, December). Recent works in Sign Language Recognition using deep learning approach-A Survey. In 2023 OITS International Conference on Information Technology (OCIT) (pp. 502-507). IEEE.

[29] Arslan, N. N., Şahin, E., & Akçay, M. (2023). Deep learning-based isolated sign language recognition: a novel approach to tackling communication barriers for individuals with hearing impairments. Journal of Scientific Reports-A, (055), 50-59.

[30] Hüseyinoğlu, A., Bilge, F. A., Bilge, Y. C., & Ikizler-Cinbis, N. (2024). Tinysign: sign language recognition in low resolution settings. Signal, Image and Video Processing, 1-10.

[31] Shin, J., Miah, A. S. M., Kabir, M. H., Rahim, M. A., & Shiam, A. A. (2024). A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities. arXiv preprint arXiv:2408.05436.

[32] Deng, Z., Leng, Y., Hu, J., Lin, Z., Li, X., & Gao, Q. (2024). SML: A Skeleton-based multi-feature learning method for sign language recognition. Knowledge-Based Systems, 112288.

[33] Núñez-Marcos, A., Perez-de-Viñaspre, O., & Labaka, G. (2023). A survey on Sign Language machine translation. Expert Systems with Applications, 213, 118993.

[34] Shah, S., Vaidya, J., Pipariya, K., & Shah, M. (2024). A Comprehensive Study on Relative Distances of Hand Landmarks Approach for American Sign Language Gesture. Augmented Human Research, 9(1), 1.

[35] Ilham, A. A., & Nurtanio, I. (2023). Dynamic Sign Language Recognition Using Mediapipe Library and Modified LSTM Method. International Journal on Advanced Science, Engineering & Information Technology, 13(6).

[36] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. Open Access.

[37] GuruAkshya, C. (2024, April). Deep Learning Framework for Sign Language Recognition Using Inception V3 with Transfer Learning. In 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-6). IEEE.

[38] Hashi, A. O., Hashim, S. Z. M., & Asamah, A. B. (2024). A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024. IEEE Access.

[39] Mohammadi, Z., Akhavanpour, A., Rastgoo, R., & Sabokrou, M. (2024). Diverse hand gesture recognition dataset. Multimedia Tools and Applications, 83(17), 50245-50267.

[40] Joy, T. S., Efat, A. H., Hasan, S. M., Jannat, N., Oishe, M., Mitu, M., & Fahim, A. M. (2023, December). Attention Trinity Net and DenseNet Fusion: Revolutionizing American Sign Language Recognition for Inclusive Communication. In 2023 26th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[41] Zhou, Y., Xia, Z., Chen, Y., Neidle, C., & Metaxas, D. (2024, May). A multimodal spatio-temporal GCN model with enhancements for isolated sign recognition. In Proceedings of the {LREC-COLING} 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

[42] Wang, L., Ni, J., Gao, H., Li, J., Chang, K. C., Fan, X., ... & Yoo, C. (2023, July). Listen, Decipher and Sign: Toward Unsupervised Speech-to-Sign Language Recognition. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 6785-6800).

[43] Altayeva, A., Omarov, B., & Im Cho, Y. (2018, January). Towards smart city platform intelligence: PI decoupling math model for temperature and humidity control. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 693-696). IEEE.

[44] Liu, Y., Zhang, S., & Gowda, M. (2022). A practical system for 3-D hand pose tracking using EMG wearables with applications to prosthetics and user interfaces. IEEE Internet of Things Journal, 10(4), 3407-3427.

[45] Tursynova, A., Omarov, B., Sakhipov, A., & Tukenova, N. (2022). Brain Stroke Lesion Segmentation Using Computed Tomography Images based on Modified U-Net Model with ResNet Blocks. International Journal of Online & Biomedical Engineering, 18(13).

[46] Robinson, N., Tidd, B., Campbell, D., Kulić, D., & Corke, P. (2023). Robotic vision for human-robot interaction and collaboration: A survey and systematic review. ACM Transactions on Human-Robot Interaction, 12(1), 1-66.

[47] Tursynova, A., Omarov, B., Tukenova, N., Salgozha, I., Khaaval, O., Ramazanov, R., & Ospanov, B. (2023). Deep learning-enabled brain stroke classification on computed tomography images. Comput. Mater. Contin, 75(1), 1431-1446.

[48] Fragkiadakis, M. (2024). LOT, msterdam. from https://hdl. handle. net/1887/3734159 Version: Publisher's Version License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repositor of the Uni ersit of Leiden Downloaded from: https://hdl. handle. net/1887/3734159.

[49] Onalbek, Z. K., Omarov, B. S., Berkimbayev, K. M., Mukhamedzhanov, B. K., Usenbek, R. R., Kendzhaeva, B. B., & Mukhamedzhanova, M. Z. (2013). Forming of professional competence of future tyeacher-trainers as a factor of increasing the quality. Middle East Journal of Scientific Research, 15(9), 1272-1276.

[50] Omarov, B., Orazbaev, E., Baimukhanbetov, B., Abusseitov, B., Khudiyarov, G., & Anarbayev, A. (2017). Test battery for comprehensive control in the training system of highly Skilled Wrestlers of Kazakhstan on national wrestling" Kazaksha Kuresi". Man In India, 97(11), 453-462.

[51] Jiang, X., Zhang, Y., Lei, J., & Zhang, Y. (2024). A Survey on Chinese Sign Language Recognition: From Traditional Methods to Artificial Intelligence. CMES-Computer Modeling in Engineering & Sciences, 140(1).