Annals of Computer Science and Information Systems
Volume 40

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

September 8–11, 2024. Belgrade, Serbia

Marek Bolanowski, Maria Ganzha, Leszek Maciaszek,
Marcin Paprzycki, Dominik Ślęzak (eds.)

Annals of Computer Science and Information Systems, Volume 40

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

Marek Bolanowski, Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Dominik Ślęzak (eds.)

POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

2024, Warszawa, Polskie
Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 40

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)

USB: ISBN 978-83-969601-0-8
WEB: ISBN 978-83-969601-9-1

ISSN 2300-5963
DOI: 10.15439/978-83-969601-9-1

Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
`http://annals-csis.org/`

Cover art:
Adrianna Emilia Iwanowska,
Elbląg, Poland

Also in this series:

Volume 41: Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-973291-0-2, ISBN USB: 978-83-973291-1-9**

Volume 39: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-6-0, ISBN USB: 978-83-969601-7-7, ISBN ART 978-83-969601-8-4**

Volume 38: Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering, **ISBN WEB: 978-83-969601-5-3**

Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**

Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2**

Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**

Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**

Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4**

Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0**

Volume 31: Position Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-2-9, ISBN USB: 978-83-965897-3-6**

Volume 30: Proceedings of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-962423-9-6, ISBN USB: 978-83-965897-0-5**

Volume 29: Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics**ISBN WEB: 978-83-962423-7-2, ISBN USB: 978-83-962423-6-5**

DEAR Reader, it is our pleasure to present to you Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS 2024), which took place on September 8-11, 2024, in Belgrade, Serbia.

Position papers comprise two categories of contributions – challenge papers and emerging research papers. *Challenge papers* propose and describe research challenges in theory, or practice, of computer science and intelligence systems. Papers in this category are based on deep understanding of existing research or industrial problems. Based on such understanding and experience, they define new exciting research directions and show why these directions are crucial to the society at large. *Emerging research papers* present preliminary research results from work-in-progress, based on sound scientific approach but presenting work not completely validated as yet. They describe precisely the research problem and its rationale. They also define the intended future work including the expected benefits from solution to the tackled problem. Subsequently, they may be more conceptual than experimental.

FedCSIS 2024 was chaired by Ivan Lukovic, while Dragana Makajić-Nikolić was the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute of Polish Academy of Sciences, The Faculty of Mathematics and Information Science Warsaw University of Technology, The Faculty of Electrical and Computer Engineering of the Rzeszów University of Technology, and The Faculty of Organizational Science of the University of Belgrade.

FedCSIS 2024 was technically co-sponsored by IEEE Poland Section, IEEE Serbia and Montenegro Section, Poland Section of IEEE Computer Society Chapter, Czechoslovakia Section of IEEE Computer Society Chapter, Serbia and Montenegro Section of IEEE Computer Society Chapter, Poland Section of IEEE Systems, Man, and Cybernetics Society Chapter, Poland Section of IEEE Computational Intelligence Society Chapter, Serbia and Montenegro Section of IEEE Computational Intelligence Society Chapter, Serbia and Montenegro Section of IEEE Education Society Chapter, Serbia and Montenegro Section of IEEE Young Professionals Affinity Group, Committee of Computer Science of Polish Academy of Sciences, Informatics Association of Serbia, and Mazovia Cluster ICT.

FedCSIS 2024 was organized in collaboration with the Strategic Partner: QED Software, and sponsored by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, Banca Intesa, Nelt Group, Netconomy, Elsevier, Journal of Computer Languages, OnlyOffice, Ascensio Systems d.o.o., Beograd, MDPI and Yettel Bank.

During FedCSIS 2024, the following Keynote and Invited lectures were delivered:

- Frank, Ulrich, University of Duisburg-Essen, Germany, *Multi-Level Language Architectures: Fostering Reuse, Integration and User Empowerment by Allowing for Additional Abstraction*
- Jovanović, Jelena, University of Belgrade, Serbia, *Learning analytics: Challenges and opportunities opened by AI*
- Kutyniok, Gitta, Ludwig-Maximilians-Universität München, Germany, *Reliable AI: Successes, Challenges, and Limitations*
- Tolvanen, Juha-Pekka, Metacase, Finland, *Languages for non-developers: what, how, where?*
- Dujmović, Jozo, San Francisco State University, USA, *Graded Logic and Professional Decision Making*

FedCSIS 2024 consisted of Main Track, with five Topical Areas and Thematic Sessions. Some of Thematic Sessions have been associated with the FedCSIS conference series for many years, while some of them are relatively new. The role of the Thematic Sessions is to focus and enrich discussions on selected areas, pertinent to the general scope of the conference, i.e. intelligence systems.

Each contribution, found in this volume, was refereed by at least two referees. They are presented in alphabetic order, according to the last name of the first author. The specific Topical Area or Thematic Session that given contribution was associated with is listed in the article metadata.

Making FedCSIS 2024 happen required a dedicated effort of many people. We would like to express our warmest gratitude to the members of Senior Program Committee, Topical Area Curators, Thematic Session Organizers and to the members of FedCSIS 2024 Program Committee. In particular, we would like to thank those colleagues who have refereed the 184 submissions.

We thank the authors of the papers for their great contributions to the theory and practice of computer science and intelligence systems. We are grateful to the keynote and invited speakers, for sharing their knowledge and wisdom with the participants.

Last, but not least, we thank Ivan Lukovic and Dragana Makajić-Nikolić and the FON Team. We are very grateful for all your efforts!

We hope that you had an inspiring conference. We also hope to meet you again for the 20th Conference on Computer Science and Intelligence Systems (FedCSIS 2025) which will take place in Kraków, Poland, on September 14-17, 2025. We also hope that you will approve the evolution of the FedCSIS Conference concept, in the direction that properly addresses the current needs of research and applications. We want to continue looking at Computer Science from different angles but, at the same time, acknowledging the topic Intelligence Systems as the central point of everything that has to be considered.

Co-Chairs of the FedCSIS Conference Series:

Marek Bolanowski, *Rzeszów University of Technology, Poland*

Maria Ganzha, *Warsaw University of Technology, and Systems Research Institute Polish Academy of Sciences, Poland*

Leszek Maciaszek (Honorary Chair)*, Macquarie University, Australia and Wrocław University of Economics, Poland*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, and Warsaw University of Management, Poland*

Dominik Ślęzak, *University of Warsaw, Poland and QED Software, Poland and DeepSeas, USA*

Position Papers of the 19th Conference on Computer Science and Intelligence Systems

September 8–11, 2024. Belgrade, Serbia

TABLE OF CONTENTS

POSITION PAPERS

DOI: 10.15439/2024F3747
ISSN 2300-5963 ACSIS, Vol. 40

Analysis of end-to-end test automation tools based on the examples of Selenium WebDriver and Playwright

Agnieszka Antonczak, Beata Bylina
Institute of Computer Science,
Marie Curie-Sklodowska University
Pl. M. Curie-Skłodowskiej 5, 20-031 Lublin, Poland
Email: agnieszka.w.antonczak@gmail.com, beata.bylina@mail.umcs.pl

***Abstract*—In the digital era, ensuring software reliability and effectiveness is crucial for business operations and daily life. This article analyzes and compares two prominent end-to-end test automation tools, Selenium WebDriver and Playwright. By examining their architecture, functionality, and browser interaction, the study provides practical guidance on tool selection. It includes theoretical foundations of testing, the importance of manual testing, and a detailed analysis of both tools. The findings suggest that Playwright generally offers faster test execution, particularly in headless mode, and simpler configuration. In contrast, Selenium benefits from a mature community and extensive documentation. The choice of tool should be based on project-specific needs, with Playwright favored for speed and simplicity, and Selenium for community support and integration capabilities.**

I. Introduction

IN THE digital age, where technology is an integral part of everyday life, software quality and reliability are crucial for both business and personal efficiency. In response to these needs, end-to-end (E2E) testing has gained prominence as a method of software verification aimed at ensuring that IT systems perform to users' expectations under real-world conditions. By simulating the end user's interactions with an application, E2E testing allows for comprehensive verification of a system's functionality [6]. Automated software testing tools are critical in performing extensive system tests, reducing errors that could lead to financial losses, and improving reliability and performance [8]. Test automation, using tools such as Selenium WebDriver and Playwright, offers efficiency, speed and accuracy that are difficult to achieve with manual approaches. A comparative analysis [7] reveals that Selenium, a widely adopted open-source tool, offers extensive support for multiple browsers and platforms, which enhances its utility for diverse testing scenarios. The study highlights Selenium's flexibility and robustness, making it a prevalent choice among testers for complex web application environments. Conversely, commercial tools like HP QuickTest Professional (now HP UFT) provide strong integration capabilities with enterprise environments, offering built-in object repositories and comprehensive technical support, which can be crucial for continuous integration and extensive test management. Another study [4] underscores that while no single tool can meet all testing requirements, TestComplete ranked highest in effectiveness among the evaluated programs, suggesting the importance of aligning tool capabilities with team skills and project needs.

Similarly, the article by E. Pelivani and B. Cico [2] compares Selenium with Katalon Studio, highlighting the strengths and weaknesses of both tools in various testing scenarios. This study underscores the necessity of selecting the right tool based on specific project requirements, further validating the approach taken in our current comparative study between Selenium WebDriver and Playwright.

This article focuses on the analysis of Selenium WebDriver and Playwright tools in the context of their use for end-to-end test automation, motivated by the absence of comparative studies between these two tools, especially given the recent introduction of Playwright in 2020. Playwright has been increasingly mentioned as a potent automation tool in the tech community, prompting an evaluation of its capabilities relative to the well-established Selenium WebDriver.

The article begins with an introduction to Selenium, focusing on its architecture and how Selenium WebDriver communicates with browsers. It then moves on to Playwright, describing it as an advanced tool capable of complex web application testing across browsers and platforms. The article continues with a discussion of developing 15 detailed test cases for an online pet store, using the Page Object Model design pattern and tools such as Maven and TestNG. This is followed by a comparative analysis of test execution times, showing that Playwright generally has shorter test execution times. The stability of tests in headless and non-headless modes is investigated. Personal experiences with the ease of use and configuration of both tools are shared. The availability of documentation, community support and educational resources for both tools is compared. Differences in extensibility and integration capabilities are described, focusing on how each tool can be extended and integrated with other systems. The article concludes with a summary of the findings, offering final thoughts on the strengths and weaknesses of Selenium WebDriver and Playwright, and providing recommendations for choosing the right tool based on specific project requirements and team expertise.

II. Selenium

Selenium WebDriver [14] is an advanced browser test automation tool that simulates user interactions with a web application. The development of Selenium, initiated in 2004 by Jason Huggins, was a response to the limitations of manual user interface testing. Originally an internal project of ThoughtWorks, it quickly gained popularity and became open-source, allowing it to evolve rapidly and adapt to new web browser technologies [1]. Selenium's architecture is characterized by its modularity and ability to integrate with various programming languages such as Java, C#, Python, which significantly extending its versatility and accessibility. The core component, Selenium WebDriver, communicates directly with the browser using dedicated drivers, such as ChromeDriver for Google Chrome or GeckoDriver for Firefox. This direct communication allows Selenium to accurately mimic user behavior, from clicks and text entry to advanced interactions with various web elements. Selenium enables tests to be run on a wide range of browsers, which is crucial for cross-browser compatibility verification of web applications. Selenium is one of the prominent automated GUI testing tools, widely used for its capability to test web applications across different browsers [8]. Its modularity also allows tests to be easily scalable and integrated into existing frameworks, significantly improving development and testing processes in dynamically changing production environments. Setting up a test environment with Selenium is relatively straightforward, typically involving the setup of appropriate browser drivers and selecting a development environment for executing test scripts. This aspect makes Selenium an attractive solution for organizations with varying degrees of technical sophistication, allowing even less technical users to effectively design and execute automated tests [5].

III. Playwright

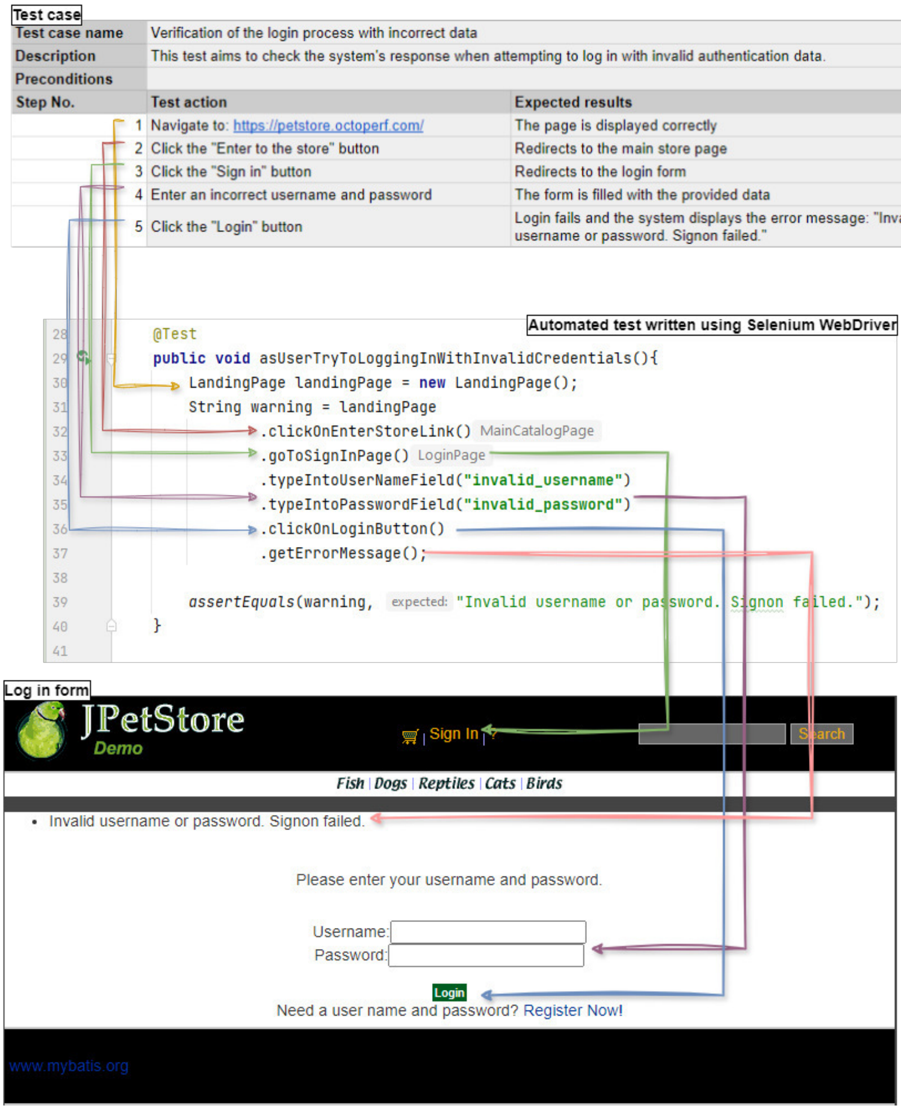
Playwright [12], a state-of-the-art browser test automation tool, has been introduced by Microsoft as a framework that provides comprehensive capabilities for web application testing. The development of Playwright is a response to the growing need for end-to-end testing that can be executed reliably and reproducibly across multiple platforms and browsers simultaneously. Since its debut, Playwright has gained recognition for its ability to work with Chromium, Firefox, and Safari, thus offering extensive cross-browser testing capabilities. Playwright's architecture is built around the idea of a "browser context," which enables the simulation of multiple sessions in a single browser instance, which is particularly useful for testing scenarios that depend on user sessions. In addition, Playwright integrates with a variety of development and testing environments providing developers with the flexibility to choose tools tailored to their specific projects. It is also distinguished by its support for programming languages such as JavaScript, Python, as well as C# and Java, which accounts for its versatility. Setting up a test environment with Playwright is intuitive and mainly involves installing the appropriate npm package or equivalents in other programming languages. Playwright provides custom drivers for browsers that are automatically managed by the framework, eliminating the need to manually configure and update browser drivers. In terms of testing capabilities, Playwright offers features such as screenshot generation, video recording of test sessions, and advanced context and session management. These features make it uniquely suited to the dynamically changing environment of modern web applications, where there are often requirements for testing application state-dependent functionality and user interaction.

IV. Test data

The analysis utilizes the example of the online pet store *https://petstore.octoperf.com* to provide a practical context for this comparison. Fifteen detailed test cases were created for this store, which were used to implement automated tests using both tools. The tests were written in Java, utilizing Maven [11] for dependency management and TestNG [13] for test management, which streamlined the testing process and enhanced the execution framework. The Page Object Model (POM) design pattern was used to implement the tests, making the test scripts more readable and easier to maintain. Figure 1 illustrates the interrelationships between a defined test case titled "Verification of login process with incorrect data," its implementation within an automated test using Selenium WebDriver, and the effect of the test, as observed on the user interface of the pet store. These elements are interconnected by arrows, allowing you to follow the flow from the test specification to the specific system behavior. The test case shown in the figure is for a login process using incorrect credentials. It includes the following steps: accessing the site (orange and red arrows), navigating to the login section (green arrow), entering invalid credentials (purple arrow) and attempting authentication (blue arrow) with the expected result of a login error message (pink arrow).

As part of the article, fifteen test cases were developed. Each test case includes a detailed description of the purpose of the test, the prerequisites necessary for its execution, the detailed specified test activities and expected results, which provides a comprehensive approach to verify the functionality of the system. As noted by Umar and Chen, the creation of detailed and comprehensive test cases is fundamental to the effectiveness of automated testing frameworks, which we have applied in our analysis of the online pet store [3]. In addition, each test case is appropriately titled:

1) Verification of the login process with correct data
2) Verification of the login process with incorrect data
3) Successful creation of a user account
4) Logging out
5) Adding a product to the shopping cart without logging in
6) Removing a product from the shopping cart without logging in
7) Searching for a product from the search bar



Test case

Test case name	Verification of the login process with incorrect data	
Description	This test aims to check the system's response when attempting to log in with invalid authentication data.	
Preconditions		
Step No.	Test action	Expected results
1	Navigate to: https://petstore.octoperf.com/	The page is displayed correctly
2	Click the "Enter to the store" button	Redirects to the main store page
3	Click the "Sign in" button	Redirects to the login form
4	Enter an incorrect username and password	The form is filled with the provided data
5	Click the "Login" button	Login fails and the system displays the error message: "Invalid username or password. Signon failed."

Automated test written using Selenium WebDriver

```
@Test
public void asUserTryToLoggingInWithInvalidCredentials(){
    LandingPage landingPage = new LandingPage();
    String warning = landingPage
            .clickOnEnterStoreLink() MainCatalogPage
            .goToSignInPage() LoginPage
            .typeIntoUserNameField("invalid_username")
            .typeIntoPasswordField("invalid_password")
            .clickOnLoginButton()
            .getErrorMessage();

    assertEquals(warning,  expected: "Invalid username or password. Signon failed.");
}
```

Log in form

- Invalid username or password. Signon failed.

Please enter your username and password.

Username:

Password:

Need a user name and password? Register Now!

www.mybatis.org

Fig. 1. Overview of the test case, automated test and login form view

8) Checking the display of the product image
9) Changing the quantity of a product in the shopping cart
10) Attempting to make a purchase without logging in
11) Making a purchase after logging in
12) Making a purchase with indicating a different delivery address
13) Return to the main menu from the category level
14) Check the display of detailed product information
15) Verifying the absence of errors with an empty search field

V. Test Execution Time

The research involved benchmarking the execution time of a set of 15 end-to-end tests using Selenium and Playwright tools. Automated tests were created based on test cases. The execution time of each test was measured directly in the Intellij IDEA environment — using the built-in functionalities of this IDE to monitor the duration of tests. The tests were executed on three major web browsers in the following versions: Google Chrome 120.0.6099.225 (Official Version) (64-bit), Microsoft Edge 121.0.2277.83 (Official Version) (64-bit) and Mozilla Firefox 122.0 (64-bit), in both standard (non-headless) and headless modes. The tests were performed on a computer running Windows 10 Home, equipped with an Intel(R) Core(TM) i5-6300HQ CPU @ 2.30GHz at 2.30GHz, with 16.0GB of RAM installed and a 64-bit, x64-based system. The overall analysis of the data shows shorter test execution times for the Playwright tool compared to Selenium in both modes. Detailed results for the Google Chrome browser in non-headless mode are shown in Figure 2, where we observe a reduction in execution time by Playwright, especially in test cases number 4, 11 and 12. Observations for the Microsoft

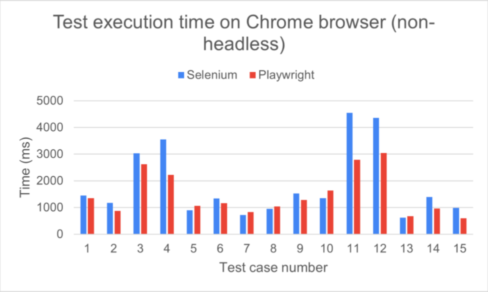


Fig. 2. Execution time of individual tests on Chrome browser (non-headless)

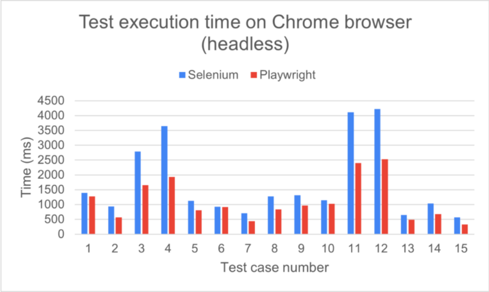


Fig. 5. Execution time of individual tests on Chrome browser (headless)

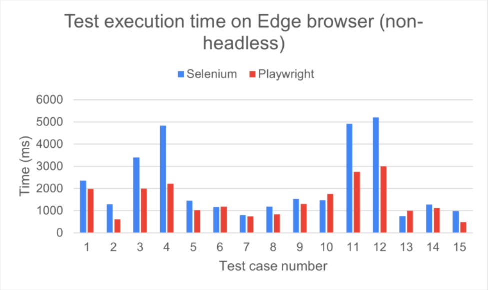


Fig. 3. Execution time of individual tests on Edge browser (non-headless)

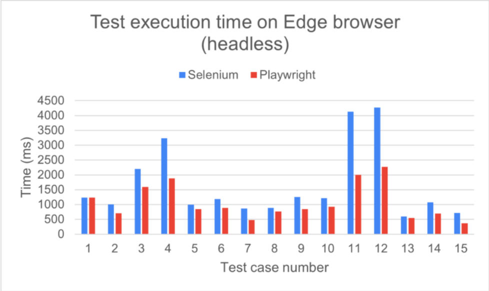


Fig. 6. Execution time of individual tests on Edge browser (headless)

Edge browser, illustrated in Figure 3, also suggest a shorter execution time for Playwright. An analysis of the results for the Mozilla Firefox browser, shown in Figure 4, also indicates an overall time advantage for Playwright over Selenium. However, test cases 1, 5, 6, 8, 9, 10 and 14 note that Selenium performs comparably or slightly better. The differences in test execution results between the tools indicate the importance of considering specific test conditions and scenarios when evaluating the performance of these tools.

In addition, analyzing the performance of tools in headless mode provides additional perspectives on their performance. As shown in Figure 5, for the Google Chrome browser in headless mode, Selenium showed longer test execution times, especially in test cases 3, 4, 11 and 12. Similar trends were observed for the Microsoft Edge browser, where Selenium also took longer to execute tests, as illustrated in Figure 6.

In contrast, Figure 7 illustrates the results for Mozilla Firefox in headless mode, which do not show an equally clear advantage for either tool. Although in some test cases, such as No. 6 and No. 9, it was Playwright that recorded longer execution times, in other cases the results of both tools are comparable. This indicates the need for a deeper analysis of the specific conditions under which the tools were tested, and potential optimizations in the test setup. These observations highlight that while Playwright generally shows better time performance in headless mode, the performance differences are dependent on the specific execution environment and can vary by browser and test scenario. Such variation in results underscores the importance of matching the test tool to the project's specifications, and shows that no tool is a one-size-fits-all solution.

Analyzing the overall execution times of all tests, we observe that Playwright performs better in both headless and standard (non-headless) modes for all tested browsers. The

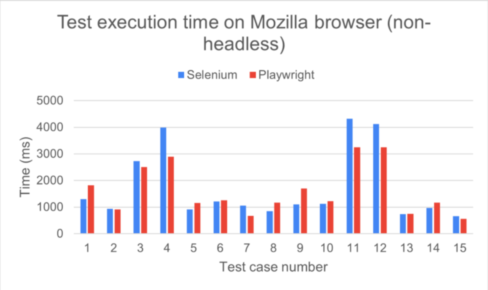


Fig. 4. Execution time of individual tests on Mozilla browser (non-headless)

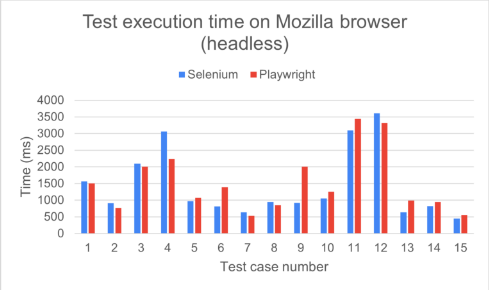


Fig. 7. Execution time of individual tests on Mozilla browser (headless)

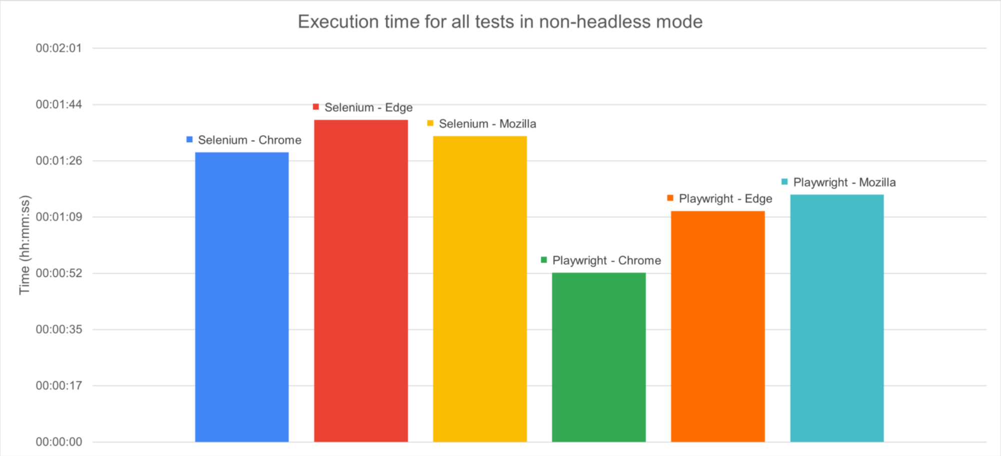


Fig. 8. Execution time for all tests in non-headless mode

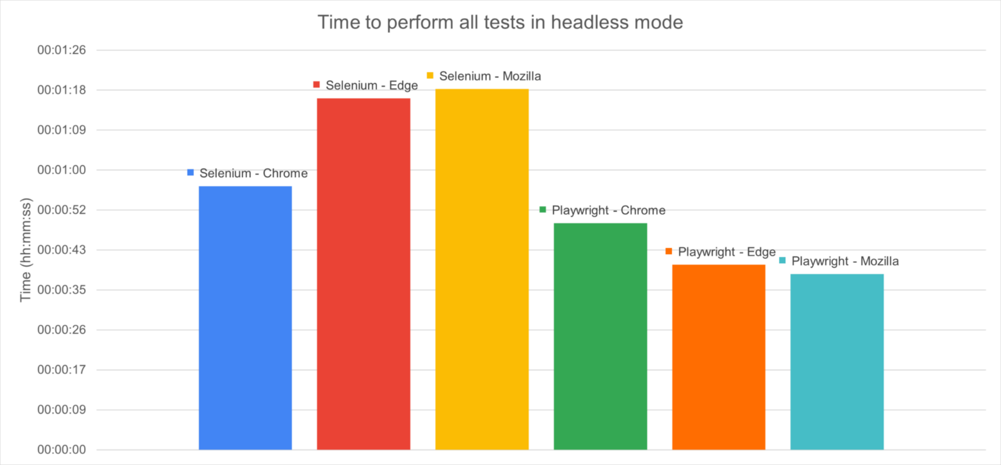


Fig. 9. Execution time for all tests in headless mode

average execution times, calculated from the three test calls, are shown in two graphs, each of which provides additional information about the effectiveness of the test tools.

In Figure 8, showing the results for non-headless mode, it can be seen that Playwright shows a significant advantage over Selenium. In particular, for the Google Chrome browser Playwright was 46% faster, for Microsoft Edge a 28% speedup was achieved, and for Mozilla Firefox a 19% speedup was achieved. These results demonstrate Playwright's superior performance in a typical browser environment, where user interactions are simulated in a full user interface.

Figure 9 illustrates the results for headless mode, where the differences in execution time are even more noticeable. In this scenario, for the Mozilla Firefox browser, Playwright was found to be faster by as much as 51% compared to Selenium, a significant time advantage of 40 seconds. For Microsoft Edge, the difference was 47%, while for Google Chrome Playwright was 14% faster. Such results in headless mode, where tests are performed without a graphical user interface, may indicate that Playwright is better optimized for performance and resource management.

The conclusions of the analysis indicate that the Playwright tool may offer better performance in terms of test execution time compared to Selenium, which may be important when choosing the right test automation tool for projects with limited time resources.

VI. Test stability

The tests were run in two modes: headless and non-headless. In non-headless mode, all tests were successful in each of the ten run cycles. In headless mode, on the other hand, instability was observed in the case of Playwright, where test 8 (concerning the display of the product image) failed an average of four times, regardless of the browser used. It is worth noting that in the case of Selenium this problem did not occur, suggesting differences in how the two tools handle rendering of page elements.

An interesting aspect is that regardless of the tool and mode, there were cases where tests failed due to errors in the application itself, rather than the testing tool. For example, a user registration test using random data revealed a problem in the handling of the ”Country” field, where entering a country name longer than 20 characters prevented registration. This indicates the value of automated testing in identifying application errors, regardless of the tool used.

The results show a negligible difference in test stability between Playwright and Selenium in headless mode. A possible reason for Playwright’s problems could be the way the tool handles element rendering, which requires further research. At the same time, these observations underscore the importance of running tests in different configurations to better understand the limitations and potential problems associated with test automation tools.

VII. Ease of use and configuration

In terms of our experience with Selenium, our first interaction with Playwright resulted in some interesting insights. From a user’s perspective, both tools presented ease of implementation; however, Playwright seemed to offer greater simplicity, especially in the aspect of browser configuration. A differentiating element was the release of the user from having to worry about browser version compatibility and the process of downloading drivers, which was important in the case of Selenium, particularly in its third version.

It should be noted, however, that these perceptions may have been shaped by the sequence of tool usage — with initial use of Selenium, followed by adaptation of existing tests to the Playwright environment. It is possible that the re-implementation of the tests ran with greater efficiency and fluidity. The initial configuration for both Selenium and Playwright is relatively similar, but there are significant differences, especially when considering different versions of Selenium. Selenium 3 requires downloading and configuring the appropriate drivers for each browser, in addition to one added dependency. Example 7.1 shows the configuration code for the Chrome browser. With Selenium 4, on the other hand, the process is somewhat simplified, as there is no need to manually download browser drivers. However, it is required to add an additional dependency — *WebDriverManager*. Nonetheless, the number of lines of code remains similar, as you can see in the example of 7.2.

Unlike Selenium, Playwright does not require adding additional dependencies or manually configuring drivers, which greatly simplifies the configuration process. Implementing the same goal in Playwright requires only one line of code as you can see in the example 7.3.

In summary, both Selenium and Playwright offer satisfying experiences for users with varying levels of experience. Personal observations suggest a preference toward Playwright’s greater intuitiveness, primarily due to its simplified browser configuration. However, it should be stressed that these conclusions are subjective in nature and may be partially determined by the order in which these tools are used. An analysis of the number of lines of code needed to execute tests shows Playwright’s advantage. For Selenium, the number of lines of code is about 2,300 for Selenium 3 and remains similar for Selenium 4, despite the ease of driver management. In comparison, Playwright requires only about 1,700 lines of code.

VIII. Community and support

In the digital age, where software development is happening at a rapid pace, community support and the availability of learning resources are becoming key factors in the selection of test automation tools. An analysis of the tools from the standpoint of documentation availability, size and activity of the community, support for new users, and training resources reveals significant differences between them.

Selenium, which has been around since 2004, has established itself as one of the most mature tools in the test automation field. Its documentation is not only extensive, but also regularly updated, with numerous examples to help users understand various aspects of the tool. This maturity is also reflected in the size and activity of the community, which is evident in forums, newsgroups and social media. Selenium is also a leader in educational resources, with numerous online courses, tutorials and webinars, reflecting its long presence in the market. On the Udemy platform, the number of courses on Selenium is 1,881, which far exceeds those dedicated to Playwright, of which there are 126 [15, 16].

On the other hand, Playwright, despite being a newer player on the market, has gained a rapidly growing community. Its documentation, while less extensive than Selenium’s, is well organized and includes numerous examples. Playwright also stands out for its development activity, as evident in the number of commits on GitHub, surpassing those for Selenium in 2023 [9, 10]. Despite the smaller number of educational resources available, Playwright is rapidly gaining popularity, indicating its future potential.

In terms of online presence, a search for Selenium brings up significantly more results than Playwright, reflecting its long-standing presence and established position in the industry. Similarly, on Stack Overflow, the number of discussions and questions about Selenium far exceeds those devoted to Playwright.

IX. Extensibility and Integration

Selenium and Playwright offer significant integration capabilities with a variety of development tools and environments. Selenium, being a mature tool, is well-established among developer tools, offering integration with popular database management systems, reporting tools and version control systems. Playwright, while newer to the market, also demonstrates impressive integration capabilities, especially with modern development environments and frameworks. Both tools feature support for multiple programming languages. Selenium has traditionally supported languages such as Java, C#, Python, which contributes to its versatility. Playwright, on the other hand, initially focused on Node.js, has also extended its support to other languages, including Java and Python, which increases its appeal in development environments.

Example 7.1 (Java): Driver configuration in Selenium 3

```
System.setProperty("webdriver.chrome.driver", LocalWebDriverProperties.getChromeWebDriverLocation());
ChromeOptions options = new ChromeOptions();
options.addArguments("--remote-allow-origins=*");
```

Example 7.2 (Java): Driver configuration in Selenium 4

```
WebDriverManager.chromedriver().setup();
ChromeOptions options = new ChromeOptions();
options.addArguments("--remote-allow-origins=*");
```

Example 7.3 (Java): Driver configuration in Playwright

```
browser = pw.chromium().launch(new BrowserType.LaunchOptions().setHeadless(false));
```

Example 7.4 (Java): Adding the ability to record tests in Playwright

```
context = browser.newContext(new Browser.NewContextOptions().setRecordVideoDir(Paths.get("videos/")));
```

Playwright is distinguished by the simplicity of using built-in tools, such as screen recording, which is accomplished through short code snippets, which is represented by the example 7.4. This capability significantly simplifies the process of documenting tests and analyzing their progress.

Such solutions built into Playwright minimize the need to look for external tools and speed up the configuration process.

On the other hand, Selenium, although it may require the use of external tools for some advanced features such as screen recording, offers better extensive documentation, making it easier to integrate these tools. The availability of extensive documentation and a user community often makes adding and configuring external tools in Selenium more intuitive and less time-consuming. The choice between Selenium and Playwright should be dictated by the specific needs of the project and the preferences of the development team. Playwright offers simplicity and speed of configuration with its built-in tools, while Selenium provides greater flexibility in integrating external tools through better documentation and community support.

X. Conclusion

The present work aimed to thoroughly analyze and compare two end-to-end test automation tools, Selenium WebDriver and Playwright. In pursuit of this goal, a series of comparative tests were conducted, focusing on aspects such as test execution time, stability, ease of use, configuration and integration capabilities with other tools. Fifteen test scenarios were developed and implemented, which included test automation for an online pet store, available at https://petstore.octoperf.com/, using Java. These scenarios involved key store functionalities, such as logging in, logging out and placing orders, which allowed a deeper evaluation of the performance and effectiveness of Selenium WebDriver and Playwright tools under real-world usage conditions. The execution of these test tasks was an integral part of the research, allowing a direct comparison of the tools in question in terms of their suitability for end-to-end test automation.

According to the study, Playwright generally offers faster test execution times compared to Selenium, which is particularly evident in headless mode. This time advantage can be significant in projects where time constraints are crucial. As shown in Table I, Playwright completed the test suite in 2 minutes and 4 seconds, whereas Selenium took 2 minutes and 38 seconds. In terms of test stability, both tools demonstrated a high level of reliability, although Playwright tended to be unstable in specific scenarios in headless mode, as indicated by the one false negative test out of 15, while Selenium had none.

As for ease of use and configuration, Playwright seemed to offer a simpler approach, especially in terms of browser configuration. This is reflected in the number of lines of code required for the project, with Playwright needing 1700 lines compared to Selenium's 2300 lines, suggesting a more efficient and streamlined setup. However, Selenium, with its maturity and community support, maintains a strong position, offering rich educational resources and better documentation. In terms of integration with other systems, both tools present extensive capabilities, however Selenium stands out with better documentation and support, making it easier to use external libraries.

Based on the analysis, I recommend choosing a test automation tool based on the specific requirements of the project. Playwright may be preferred in projects where short test execution times and simple configuration are a priority, while Selenium will be a better choice in situations where rich community support, high configuration flexibility and the ability to integrate with external tools are key.

There are a number of opportunities for further development of this topic. Future research could focus on an extended comparison of the performance of the two tools in different environments and applications, including more complex testing scenarios. In addition, analysis of the long-term stability and scalability of these tools in large software projects could provide more valuable information. It's also worth exploring how developments in technologies such as artificial intelligence could affect the future of test automation, which could open up new perspectives in the field.

In conclusion, although both tools — Selenium WebDriver

TABLE I
COMPARISON OF SELENIUM WEBDRIVER AND PLAYWRIGHT BASED ON KEY CRITERIA

Criteria	**Selenium WebDriver [14]**	**Playwright [12]**
Browser compatibility	Chrome, Safari, Firefox, Edge	Chromium, Safari, Firefox, Edge
Programming language support	Java, JavaScript, Python, Ruby, C#	Java, Node.js, Python, .NET
Cost	Open source, free to use.	Open source, free to use.
Operating System compatibility	Windows, Linux, and macOS	Windows, Linux, and macOS
Total test execution time	2 Minutes 38 Seconds	2 Minutes 4 Seconds
False negative tests	0/15	1/15
Lines of code in project	2300	1700
Number of courses on Udemy [15, 16]	1881	126
Number of threads on StackOverflow [17, 18]	57696	3213

and Playwright — have their strengths and weaknesses, the choice between them should always be made taking into account the specific needs and limitations of the project. Umar and Chen in their study conclude that the success of automated testing projects heavily relies on the appropriate selection of testing tools and frameworks, a perspective that our comparative study of Selenium WebDriver and Playwright supports [3]. The research and analysis carried out provides valuable input to the evaluation of these tools, and the conclusions drawn from this work can be helpful in deciding on the appropriate test automation tool.

REFERENCES

[1] A. Holmes, M. Kellogg, *Automating functional tests using Selenium*, 2006, doi: 10.1109/AGILE.2006.19

[2] E. Pelivani and B, Cico *A comparative study of automation testing tools for web applications*, 2021 doi: 10.1109/MECO52532.2021.9460242.

[3] M. A. Umar, Z. Chen, *A study of automated software testing: Automation tools and frameworks*, International Journal of Computer Science Engineering 2019, doi: 10.5281/zenodo.3924795.

[4] M. Psujek, A. Radzik, G. Kozieł, *Comparative analysis of solutions used in automated testing of Internet applications*, Department of Computer Science, Lublin University of Technology, Lublin, Poland, 2021, doi: 10.35784/jcsi.2373.

[5] P. Ramya, V. Sindhura, P. V. Sagar, *Testing using selenium web driver*, 2017, doi: 10.1109/ICECCT.2017.8117878.

[6] R. Dahiya, A. Shahid, *Importance of manual and automation testing* Department of Information Technology, AGI Institute, Auckland, New Zealand, 2019, doi: 10.5121/csit.2019.91719.

[7] R. K. Lenka, U. Satapathy, M. Dey, *Comparative Analysis on Automated Testing of Web-based Application*, Department of CSE, 2018, doi: 10.1109/ICACCCN.2018.8748374.

[8] S. K. Alferidah, S. Ahmed *Automated Software Testing Tools*, International Conference on Computing and Information Technology, ICCIT 2020, doi: 10.1109/ICCIT-144147971.2020.9213735.

[9] https://github.com/microsoft/playwright/graphs/commit-activity (06.01.2024).

[10] https://github.com/SeleniumHQ/selenium/graphs/commit-activity (06.01.2024).

[11] https://maven.apache.org/

[12] https://playwright.dev/docs/intro

[13] https://testng.org/

[14] https://www.selenium.dev/documentation/

[15] https://www.udemy.com/courses/search/?src=ukw&q=Playwright (06.01.2024).

[16] https://www.udemy.com/courses/search/?src=ukw&q=Selenium (06.01.2024).

[17] https://stackoverflow.com/questions/tagged/selenium-webdriver?tab=Newest (10.07.2024).

[18] https://stackoverflow.com/questions/tagged/playwright (10.07.2024).

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 9–16

DOI: 10.15439/2024F9098
ISSN 2300-5963 ACSIS, Vol. 40

No Train, No Pain? Assessing the Ability of LLMs for Text Classification with no Finetuning

Richard Fechner*† , Jens Dörpinghaus*‡
* Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Correspondence: richard.fechner@bibb.de,
jens.doerpinghaus@bibb.de, https://orcid.org/0000-0003-0245-7752
† University of Tübingen, Germany
‡ University Koblenz, Koblenz, Germany

***Abstract*—Modern SotA Text Classification algorithms depend heavily on well annotated and diverse data capturing the intricacies of the unknown data distribution. What options do we have when labeled data is sparse or annotation is expensive and time consuming? With the advent of strong LLM backbones, we have another option at our disposal: Text Classification by making use of the reasoning ability and the strong general prior of contemporary foundation models. In this work we assess the ability of cutting edge LLMs for Text Classification and find that for the right combination of backbone and prompt strategy we're able to near-rival trained baselines for the advanced task of mapping job-postings to a taxonomy of industrial sectors without any finetuning. All our code is made publicly available at our github repository[1].**

I. Introduction

TEXT classification is a widely used technology with a broad range of applications. However, it is rare for a one-size-fits-all solution to exist, and the situation becomes even more complex when the availability of training data and the complexity of clustering questions are taken into account. In previous research, we have worked on the domain of industrial sectors in labor market research data, see [1]. The classification of industrial sectors is of great importance, yet the range of available textual data is vast and only a limited amount of annotated training data exists. It was demonstrated that a categorization is possible, yet the quality of the categorization depends on both the training and evaluation data. Consequently, the specific clustering is dependent on the application and the research question.

For instance, all approaches failed for job advertisements, which are often used in labor market research. Conversely, a satisfactory recall was achieved on Wikipedia data. The proposed method was not yet ready for productive use, but it demonstrated that the initial research question was challenging due to the diversity of data and expected outcomes, as well as the interdisciplinary nature of the research. In this specific area, educational research and the social sciences had a different perspective on industrial sectors than, for example, economics. Therefore, understanding the correct classification depends not only on the research questions but also on the perspective of different scientific domains.

This paper will build upon the existing body of research on the classification of online job advertisements (OJAs) in industrial sectors. Our primary research questions are as follows:

1) *How may we harness the strong prior and reasoning ability of LLMs for knowledge intensive Text Classification directly when we have little to no labeled data?*
2) *How do different prompting strategies and models perform?*

When classifying a job-posting, we may classify the industrial section (IS) of the advertised job or the IS of the company posting the inquiry. However, the IS of the company and the jobad musn't match, i.e. a bakery might post a jobad for a roofer or IT-specialist. *In the following work we're concerned with classifying the IS of the company.* By nature, job-postings contain more information about the job and information about the company besides the name is sparse. The process of annotation is hence very time intensive and tedious as missing information about the company has to be searched on the web, analysed and finally combined with prior knowledge to obtain a good classification. Additionally, text data is often noisy, containing web-scraping boilerplate-text. The human universal prior allows for an easy differentiation of all these effects to the point, where we may perceive a problem at hand to be solvable "out of the box" for machines, when in fact some problems are impossible to solve when not equipped with prior knowledge and reasoning ability. A classical example is an agent for a self-driving car. Both the human driver and the algorithm perceive the same information (stereo RGB images), yet we have no agent driver which can reliably navigate in changing environments. In the same sense, text-classification is perceived as an easy task, where in reality it may be very hard. For specific data a certain amount of reasoning capability is needed in order to perform classification. Hypotheses have to be weighted against each other, even among annotators it is often not clear what class a jobposting may belong to. All the more important is the ability of a classification to be interpretable in the sense that a second annotator might judge the reasoning steps taken that led to the final classification.

[1] https://github.com/rfechner/fedcsis24-llm-textcat

Topical area: Advanced Artificial Intelligence in Applications

II. Related Work

We split this section into two subsections, the first giving a brief (and incomplete) summary of the recent history of text-classification, putting emphasis on neural methods, in particular the dominant Transformer architectures like BERT and the alignment of modern LLMs. The second subsection discusses the related work in the general domain of labor market analysis.

A. On the evolution of Neural Text Classification

Neural approaches to Text-Classification have gained attention since the introduction to architectures like the LSTM [2] and later the Transformer [3] from which the latter emerged as the popular architectural choice for working with text data. The BERT architecture [4] and its derivatives [5], [6], [7] are widely used and are still a common choice for sentence, text or document classification [8]. On a meta-level, we'd like to narrow down the choices for the next token by injecting more information either during inference or training time. Works like TransformerXL [9] or Longformer [10] try to trade off performance and context size which is imperative to capturing semantics in a text. In the domain of job-advertisements the authos of [11] have used continous pre-training on in-domain data (i.e. job-postings) to reduce the confusion of the language model on downstream tasks. For recent large models like ones of the LLaMA family [12] this approach becomes unattractive as the amount of data and compute needed to pre-train a model is likely very large. Instead we'd like to adapt the output distribution of the LLM by exterior methods and rely more heavily on the models reasoning capability. Aligning LLMs to conform to desired behaviour and alleviating the common mishaps of LLMs such as hallucinations or the lack of structure in the models answer is an area of active research. Techniques such as prompt engineering are among the most natural ways of alignment. On the other hand, prompts may mislead a LLM and throw it off in such a way that makes it ignore all previous safety instructions, leading to possible misuse [13]. A prompting strategy like Chain-of-Thought [14] or Tree-of-Thought [15] has been shown to substantially improve model performance. Addressing the issue of hallucination and lack of up-to-date in-domain knowledge is Retrieval Augmented Generation (RAG) which augments the token generation of an LLM with context provided by a so-called retriever [16]. Most recent work by the open source community was focused on creating so called chains of Language Models or "LangChains" [17] for short, structuring the token-generation process and unifying the previously mentioned exterior alignment methods. These methods in turn suffer from error accumulation over the multiple prediction steps.

B. On Related work in labor market analysis

Very little work has been done in this area. There are several applications for the given research question: For example, Pejic et al. state the need to analyse Industry 4.0 skills, but do not present a generic categorization approach, but rather pre-select job advertisements according to their needs [18]. Chaisricharoen et al. noted the importance of industrial sectors for legal categories. However, their work is limited to industry-standard keywords [19]. For the generic categorization of English texts, some work has been done by McCallum [20] and Kibriya et al. [21]. However, the data and industrial sectors are mainly for marketing purposes and cannot be used in economic and sociological research. Several other works rely on these data-sets, see for example [22], [23], which underlines the general need for publicly available training and evaluation data.

Text mining on labor market data is a widely considered topic. For an automated analysis of labor-market related texts, the situation in German-speaking countries like Germany, Austria and Switzerland is not much different to English-speaking countries: "Catalogs play a valuable role in providing a standardized language for the activities that people perform in the labor market" [24]. However, while these catalogs are widely used for creating and computing statical values, for managing labor market and educational needs or for recommending trainings and jobs, there is no single ground truth. According to Rodrigues et al., one reason for this could be the fact that labor market concepts are modeled by multiple disciplines, each with a different perspective on the labor market [25]. For German texts, in particular job advertisements, Gnehm et al.[26] introduced transfer learning and domain adaptation approaches with jobBERT-de and jobGBERT. This model was also used for the detection of skill requirements in German job advertisements [27], [28].

For regional data, especially in German-speaking countries, industrial sectors are widely used as a basis for economic and labour market research, see for example [29], [30], [31], they are particularly important for future skills and qualifications [32]. Although classification is a key issue for industrial sectors, see [33], little research has been carried out using computational methods. Examples are mainly limited to regional industries [34] or agriculture and green economy [35].

To our knowledge, no work has been done on German texts. Company data are usually collected and sold by commercial providers such as statista. There is also an online guide from the Federal Office of Economic Affairs and Export Control (BAFA) ("Merkblatt Kurzanleitung Wirtschaftszweigklassifikation"[2]), but this is only a short version of the data available from the Federal Statistical Office. Therefore, we will now discuss the available data.

III. Data

As discussed in the Appendix section on general Information about the German Industrial Sector Taxonomy WZ-2008 (See Appendix: A), several classifications of industrial sectors exist. We will continue by giving insight into the dataset construction and annotation process.

[2] https://www.bafa.de/SharedDocs/Downloads/DE/Wirtschaft/unb_kurzanleitung_wirtschaftszweigklassifikation.pdf.

A. Dataset construction and Annotation Process

Out of a large database of unlabeled job-postings, we drew a sample of about 2000 Jobpostings (1976. Assuming a general "distribution of jobads" $p(x)$, it should be noted that the drawn samples came from another (conditional) distribution $p(x|y)$ where y is the actual advertised job. More precisely we drew from a distribution which advertised open positions for roofers overproportionally. Hence, there exists a strong bias in the evaluation dataset, which is further discussed in the section on bias.

A small team of annotators took extensive time and co-ordinantion to annotate these jobpostings. The process of annotation includes reading through the sample, gathering information about the industrial section of the employer and finally coming up with hypotheses, which are then checked for validity. In the end, one has to verify that the hypothesized industrial section matches. Later, we will discuss how we modelled the prompting strategy after the annotation process. Invalid data, e.g. datapoints which were not actually jobpostings but advertisements or simply degenerate, were filtered out during the annotation process. We split the dataset into a 0.8-train, 0.1-test and 0.1-validation sets, to train and evaluate baseline models. The sample distribution was preserved inside the testset. To gain intuition on what the sample label frequencies (See Appendix: B) and examples of the data (See Appendix: C), we refer the reader to the Appendix.

B. Online Job Advertisements

In our research, we focus on several large corpora for OJAs. The first dataset was obtained from the German Federal Employment Agency (Bundesagentur für Arbeit – BA). This dataset contains approximately 5.5 million OJAs spanning the years 2013 to 2022. This portal is one of Germany's largest job portals. The OJA records include several metadata, including a job classification (KldB) and industrial sectors according to WZ08. All data is manually curated. The second corpus comprises approximately 4 million OJAs from various data sources, including job portals such as Academics, Monster, and BA. This data contains several metadata, a classification of occupations according to ISCO, and industrial sectors on WZ08. However, it should be noted that these annotations are not manually curated but rather the result of unknown AI approaches which do not have a high level of quality.

IV. Method

First it should be noted that our experiments were strongly influenced by the fact that most LLMs were most likely trained on the WZ1993, WZ2003 and WZ2008 taxonomies. This opens up possibilties for prompts, as we may assume that the model has some form of understanding of the classes it is supposed to map onto. We conducted experiments on a small group of openly avaiable (open weights) contemporary Instruction finetuned LLMs using the python-ollama library [36] running on a local NVIDIA L40 GPU. At this point, we'd like to note that due to data privacy laws we are unable to test API-models as GPT-4o or Claude Sonnet. We tested several prompt strategies . When designing the prompts, we stuck to the general principle for success when engineering a prompt: Inducing a bias towards clear and small, step by step reasoning. The outputs of the model are verified heuristically at each intermediate step, making sure that the output is well behaved. At the end, an output parser makes a final response validation, making sure that the predicted class is valid. For invalid assistant responses, the chat is at most repeated a fixed number of times until the query is failed for the specific datapoint. Failure was most commonly due to an invalid output format or invalid classification (i.e. hallucination) of the model.

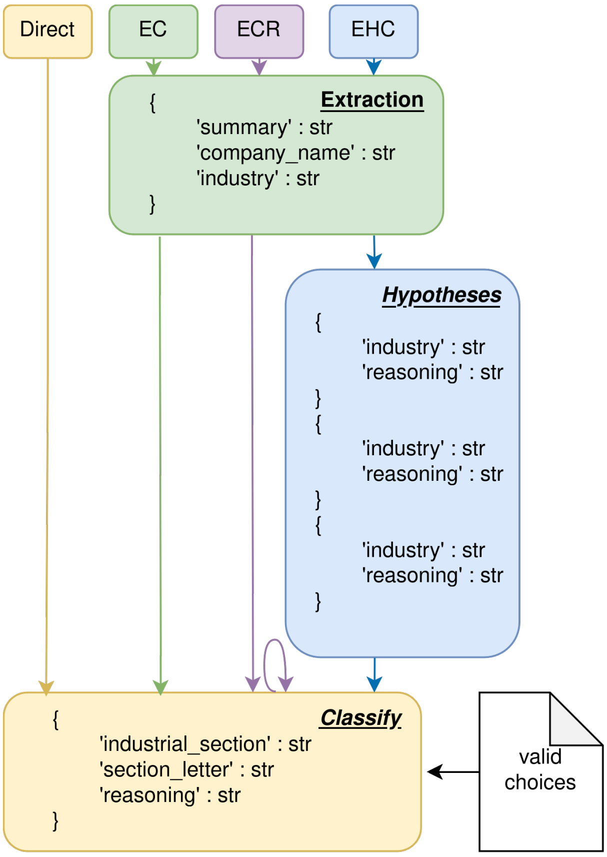


Fig. 1. Different strategies for prompting : "Direct", "Extract-Classify" (EC), "Extract-Classify-Reflect" (ECR) and "Extract-(Generate) Hypotheses-Classify" (EHC). Input is a raw jobposting. All outputs are formatted in JSON. Output is verfied for a valid industrial section to ensure a non-degenerate answer.

For the baseline model, we finetuned a german distilBERT model [37] and a finetuned model [38] on the text classification task using the `transformers` library and Focal Loss [39] to put more emphasis on low-frequency classes. Additionally, we trained a few more standart classifiers from the `sklearn` library [40].

V. Results

Generally, we can see in the results that bigger models outperform smaller ones for Direct prompting. The 70B-parameter version of LLaMA3 (See Table I, ✠) reaches

TABLE I
METRICS FOR DIFFERENT LLMS, PROMPTING STRATEGIES AND PROMPT TYPES. LARGE MODELS (LLAMA3:70B AND COMMAND-R:35B) OUTPERFORM SMALLER ONES.

LLM	Strategy	Prompt Type	Metrics (Macro/Weighted)			Failed Samples
			Precision	Recall	F1 Score	
llama3:8b	Direct	zero_shot	0.01 / 0.00	0.12 / 0.01	0.02 / 0.00	0
		one_shot	0.04 / 0.13	0.03 / 0.01	0.01 / 0.02	0
		few_shot	0.11 / 0.61	0.05 / 0.16	0.04 / 0.19	2
	EC	zero_shot	0.21 / 0.73	0.20 / 0.42	0.14 / 0.39	16
		one_shot	0.04 / 0.28	0.07 / 0.08	0.02 / 0.10	0
		few_shot	0.00 / 0.00	0.06 / 0.01	0.00 / 0.00	2
	ECR	zero_shot	0.15 / 0.62	0.22 / 0.54	0.15 / **0.53**	15
		one_shot	0.03 / 0.23	0.08 / 0.16	0.03 / 0.18	0
		few_shot	0.00 / 0.00	0.06 / 0.01	0.00 / 0.00	6
	EHC	zero_shot★	**0.27 / 0.69**	0.21 / 0.52	**0.19** / 0.49	12
		one_shot⧫	0.04 / 0.20	0.11 / 0.10	0.103 / 0.11	31
		few_shot⧫	0.09 / 0.47	0.14 / 0.04	0.02 / 0.02	84
aya:8b	Direct	zero_shot	0.00 / 0.00	0.03 / 0.01	0.00 / 0.00	0
		one_shot	0.06 / 0.47	0.06 / 0.02	0.00 / 0.00	0
		few_shot	0.03 / 0.25	0.04 / 0.31	0.04 / 0.27	5
	EC	zero_shot	0.12 / 0.54	0.14 / 0.48	0.10 / 0.40	2
		one_shot	0.00 / 0.00	0.06 / 0.01	0.00 / 0.00	0
		few_shot	0.03 / 0.23	0.01 / 0.01	0.00 / 0.02	0
	ECR	zero_shot⧫	0.11 / 0.48	0.23 / 0.50	0.13 / 0.44	28
		one_shot⧫	0.00 / 0.00	0.07 / 0.02	0.00 / 0.00	33
		few_shot	0.00 / 0.00	0.07 / 0.01	0.00 / 0.00	2
	EHC	zero_shot⧫	0.17 / 0.67	0.15 / 0.54	0.11 / 0.41	97
		one_shot	0.07 / 0.47	0.08 / 0.03	0.01 / 0.03	5
		few_shot	0.10 / 0.57	0.10 / 0.02	0.03 / 0.03	3
gemma:7b	Direct	zero_shot	0.07 / 0.39	0.04 / 0.02	0.01 / 0.02	0
		one_shot⧫	0.00 / 0.00	0.02 / 0.01	0.00 / 0.00	25
		few_shot⧫	0.04 / 0.27	0.06 / 0.46	0.05 / 0.34	34
	EC	zero_shot	0.16 / 0.32	0.14 / 0.44	0.15 / 0.37	3
		one_shot⧫	0.00 / 0.00	0.09 / 0.03	0.01 / 0.00	131
		few_shot⧫	0.03 / 0.15	0.07 / 0.38	0.04 / 0.21	76
	ECR	zero_shot⧫	0.14 / 0.36	0.22 / 0.46	0.14 / 0.40	47
		one_shot⧫	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	208
		few_shot⧫	0.04 / 0.14	0.09 / 0.38	0.05 / 0.21	143
	EHC	zero_shot	0.14 / 0.66	0.15 / 0.47	0.10 / 0.37	18
		one_shot⧫	0.00 / 0.00	0.08 / 0.04	0.01 / 0.00	134
		few_shot⧫	0.02 / 0.13	0.07 / 0.03	0.00 / 0.01	196
phi3:3.8b	Direct	zero_shot	0.04 / 0.29	0.04 / 0.06	0.01 / 0.09	0
		one_shot	0.02 / 0.12	0.08 / 0.02	0.01 / 0.02	4
		few_shot	0.04 / 0.25	0.06 / 0.16	0.03 / 0.18	0
	EC	zero_shot⧫	0.12 / 0.23	0.18 / 0.43	0.12 / 0.29	20
		one_shot⧫	0.03 / 0.17	0.08 / 0.13	0.02 / 0.14	22
		few_shot	0.04 / 0.17	0.13 / 0.03	0.02 / 0.03	18
	ECR	zero_shot⧫	0.20 / 0.59	0.21 / 0.48	0.18 / 0.37	24
		one_shot	0.09 / 0.56	0.04 / 0.15	0.02 / 0.16	17
		few_shot⧫	0.06 / 0.41	0.02 / 0.10	0.02 / 0.15	56
	EHC	zero_shot⧫	0.25 / 0.56	0.19 / 0.47	0.19 / 0.37	76
		one_shot	0.04 / 0.21	0.11 / 0.07	0.02 / 0.08	20
		few_shot⧫	0.04 / 0.19	0.09 / 0.04	0.02 / 0.04	23
command-r:35b	Direct	zero_shot⧫	0.16 / 0.54	0.18 / 0.11	0.11 / 0.13	31
		one_shot	0.16 / 0.28	0.14 / 0.17	0.10 / 0.20	0
		few_shot	0.05 / 0.27	0.17 / 0.31	0.06 / 0.28	3
	EC	zero_shot	0.31 / 0.68	0.23 / 0.64	0.24 / 0.60	0
	EHC	zero_shot⧫	0.34 / 0.76	0.31 / 0.75	0.30 / 0.73	21
llama3:70b	Direct	zero_shot✠	0.34 / **0.80**	**0.29 / 0.77**	0.28 / **0.76**	1
		one_shot	0.32 / 0.72	0.32 / 0.65	0.26 / 0.62	1
		few_shot	0.28 / 0.69	0.29 / 0.70	0.27 / 0.67	1
	EC	zero_shot⧫	0.35 / 0.76	0.29 / 0.69	0.29 / 0.68	128
	EHC	zero_shot⧫	0.35 / 0.71	0.30 / 0.66	0.28 / 0.64	137

competetive scores on the benchmark **without any finetuning**. Smaller models (See Table I, ★) are able to archive solid performance with the right prompting strategy. Most configurations however (See Table I, ♦) are disqualified for their high failure rate ($\geq 10\%$) during evaluation. This is attributed to the fact that output produced by the model wasn't conforming to the specifications of the respective prompting strategy and a patience level was reached leading to termination. Smaller models as phi3:3.8B performed evaluation faster (about 1 second/sample for Direct, 5 for EC/ECR and about 15 seconds/sample for EHC), but lack strong performance. The large models, as LLaMA3:70B took longer (about 5 seconds/sample for Direct and 50 seconds/sample for EHC). We've trained a few baseline models II for comparison and evaluated on the same test-set.

TABLE II
METRICS FOR BASELINE MODELS

Model	Metrics (Macro/Weighted)		
	Precision	Recall	F1 Score
distilbert-base-german-cased	0.31/0.84	0.31/0.86	0.31/0.85
agne/jobBERT-de	0.35/0.86	**0.36**/0.87	0.34/0.86
MNNaiveBayes	0.10 / 0.70	0.12/0.81	0.11/0.75
RandomForest◊	**0.49** / 0.86	0.34/**0.89**	**0.39/0.87**
LinearSVC	0.46 / **0.87**	0.34/**0.89**	0.38/**0.87**

We see that SotA models still outperform text generation based methods. All baselines apart from the Naive Bayes Classifier are within the same performance class, with the RandomForest Classifier excelling even compared to Transformer-based language models.

VI. DISCUSSION

We showed that some LLMs are capable of delivering zero-shot competetive performance on the task of Text Classification when compared to contemporary neural and conservative methods. Again, we'd like to note that all tested LLMs were most likely trained on the taxonomies WZ-1993, WZ-2003 and WZ-2008, which raises the question, whether we may be able to see the same performance on newly established taxonomies - most likely not. Other ways of injecting knowledge into the classification process would have to be explored.

Throughout our experiments, we've seen that introducing additional information (as in one-shot and few-shot prompting) may be harmful to performance *in this particular problem setup*. Most of the times a direct prompting strategy seems to be preferred. However, this may also be caused by the choice of prompt. We forced the models to output json in each output step, which may have hindered "flow of thought" of the model and lead to increased failure rates, as in our setup we failed a sample if the model output didn't conform to the specification after a fixed number of tries (the patience hyperparameter was chosen to be 3 in most cases).

Surprisingly, the more complex strategies like EC, ECR and EHC didn't increase performance for LLaMA:70B and command-r:35B, instead it seemed that it worsened performance unlike with smaller models where performance generally increased with this prompting strategy. The added complexity of the prompt seems to conflict with the complexity of the model. For large models a less strict prompt-output verification scheme should probably be explored, as less samples may fail. Another important point is the fact that we were able to "persuade" the model to output an argument (captured in the output value for 'reasoning') for the algorithms decision for the given industrial section. We acknowledge that it isn't a real "decription" of the internal model reasoning, as the model isn't keeping a hidden state over time but it still may prove very useful as it may help to reduce annotation time in real world applications.

A. Bias

Bias was introduced by selecting the sample distribution $p(x|y)$ which over-represents positions for roofers (Dachdecker) and the human annotation process. Inherently most companies allow for a multi-class classification, which introduces annotation conflicts and hinders learning. The training and test data distributions put about 80% of the probability mass on samples with sections F (Baugewerbe - Construction) and N78 (Vermittlung und Überlassung von Arbeitskräften - Placement and leasing of workers). Furthermore the pre-trained models `distilbert-base-german-cased` aswell as `agne/jobBERT-de` contain an unknown form of bias, as the pretraining datasets are not available.

In the conducted experiments, the choice of the wording of the prompt, the prompting strategy and the examples and solutions to the examples are chosen randomly from the training data distribution. It may be that choice of examples strongly influences the models performance on the test set.

VII. SUMMARY AND OUTLOOK

A. Summary

In this work, we've explored the ability of LLMs to perform Text Classification without any further finetuning. We've empirically shown that for some models the right prompting strategy yields comparable performance to methods which require extensive data (See Table II, ◊). Prompting LLaMA3:8b with the proposed Extract-Hypothesize-Classify Strategy (See Table I, ★) or LLaMA3:70b with a Direct Classification Strategy (See Table I, ✠) shows promise. Challenges remain, as for some LLMs failure rates are high (See Table I, ♦) and reasoning capability is limited. Now, we'd like to answer the two questions posed in the introduction:

1) *How may we harness the strong prior, search and reasoning ability of LLMs for knowledge intensive Text Classification directly when we have little to no labeled data?*
   We may choose a good combination of LLM (one that is fit for our use-case) and prompting strategy. Additionally, we may have to inject domain knowledge to ensure valid output.
2) *How do different prompting strategies and models perform?*

> We have empirically shown that models are very sensitive to the a surplus of information. Generally a simple strategy, even a zero-shot strategy is the best for our use-case. This may of course be different for other taxonomies or text classification problems.

B. Outlook

Other ways of injecting knowledge into the classification process would have to be explored. To this end we hypothesize a sort of "Generation Augmented Retrieval" might be an interesting direrction of research, where unlike in RAG we augment the retriever with processed information about the data to be able to map more reliably into a latent space - and make a good prediction. An LLM may extract important information about the employer before mapping the extracted information into a latent space. This would indeed require labeled data to train the retriever - the question remains whether this approach yields a performance improvement.

A central take-away of our work is that for annotation-sparse problem setups LLM-based Text Classification will play an important role in the future - especially considering the growth in reasoning ability and capacity of modern models. Further work should be done in this direction. When context size grows and models are isntruction fine-tuned to make use of tools such as web-search and document-search we conjecture that LLM generation based classification will be able to outperform SoTA.

LLM Disclaimer

For the generation of LaTeX code (only Table structure) and for the purpose of prototyping experiment code we used Large Language Models.

Acknowledgments

We would like to thank our colleagues Christoph Holzmüller and Katerina Kostadinovska for helping us out with data annotation.

References

[1] R. Fechner, J. Dörpinghaus, and A. Firll, "Classifying industrial sectors from german textual data with a domain adapted transformer," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023, pp. 463–470.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] 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.

[5] 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.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[8] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Systems with Applications*, vol. 106, pp. 36–54, 2018.

[9] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[10] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[11] J.-J. Decorte, J. Van Hautte, T. Demeester, and C. Develder, "Jobbert: Understanding job titles through skills," *arXiv preprint arXiv:2109.09605*, 2021.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against llm-integrated applications," *arXiv preprint arXiv:2306.05499*, 2023.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[15] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

[16] 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," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[17] H. Chase, "LangChain," Oct. 2022. [Online]. Available: https://github.com/langchain-ai/langchain

[18] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements," *International journal of information management*, vol. 50, pp. 416–431, 2020.

[19] R. Chaisricharoen, W. Srimaharaj, S. Chaising, and K. Pamanee, "Classification approach for industry standards categorization," in *2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*. IEEE, 2022, pp. 308–313.

[20] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48.

[21] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*. Springer, 2005, pp. 488–499.

[22] H. Hayashi and Q. Zhao, "Quick induction of nntrees for text categorization based on discriminative multiple centroid approach," in *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2010, pp. 705–712.

[23] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[24] C. Ospino, "Occupations: Labor market classifications, taxonomies, and ontologies in the 21st century," *Inter-American Development Bank*, 2018.

[25] M. Rodrigues, Fernández-Macías, and Enrique, Sostero, Matteo, "A unified conceptual framework of tasks, skills and competences," Seville, 2021. [Online]. Available: https://joint-research-centre.ec.europa.eu/publications/unified-conceptual-framework-tasks-skills-and-competences_en

[26] A.-S. Gnehm, E. Bühlmann, and S. Clematide, "Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3892–3901.

[27] A.-S. Gnehm, E. Bühlmann, H. Buchs, and S. Clematide, "Fine-grained extraction and classification of skill requirements in german-speaking job ads." Association for Computational Linguistics, 2022.

[28] J. Büchel, J. Engler, and A. Mertens, "The demand for data skills in german companies: Evidence from online job advertisements," *How to Reconstruct Ukraine? Challenges, Plans and the Role of the EU*, p. 56, 2023.

[29] B. Gehrke, H. Legler, M. Leidmann, and K. Hippe, "Forschungs- und wissensintensive wirtschaftszweige: Produktion, wertschöpfung und beschäftigung in deutschland sowie qualifikationserfordernisse im europäischen vergleich," Studien zum deutschen Innovationssystem, Tech. Rep., 2009.

[30] N. Gillmann and V. Hassler, "Coronabetroffenheit der wirtschaftszweige in gesamt-und ostdeutschland," *ifo Dresden berichtet*, vol. 27, no. 04, pp. 03–05, 2020.

[31] U. Kies, D. Klein, and A. Schulte, "Cluster wald und holz deutschland: Makroökonomische bedeutung, regionale zentren und strukturwandel der beschäftigung in holzbasierten wirtschaftszweigen," *Cluster in Mitteldeutschland–Strukturen, Potenziale, Förderung*, p. 103, 2012.

[32] V.-P. Niitamo, "Berufs-und qualifikationsanforderungen im ikt-bereich in europa erkennen und messen," *Schmidt, SL; Strietska-Ilina, O.; Dworschak, B*, pp. 194–201, 2005.

[33] J. Hartmann and G. Schütz, "Die klassifizierung der berufe und der wirtschaftszweige im sozio-oekonomischen panel-neuvercodung der daten 1984-2001," SOEP Survey Papers, Tech. Rep., 2017.

[34] M. Titze, M. Brachert, and A. Kubis, "The identification of regional industrial clusters using qualitative input–output analysis (qioa)," *Regional Studies*, vol. 45, no. 1, pp. 89–102, 2011.

[35] U. Kies, T. Mrosek, and A. Schulte, "Spatial analysis of regional industrial clusters in the german forest sector," *International Forestry Review*, vol. 11, no. 1, pp. 38–51, 2009.

[36] Ollama, "Ollama software repository," 2024. [Online]. Available: https://github.com/ollama/ollama

[37] distilbert, "distilbert-base-german-cased software repository," 2024. [Online]. Available: https://huggingface.co/distilbert/distilbert-base-german-cased

[38] A.-S. "Gnehm, E. Bühlmann, and S. Clematide, ""evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements"," in *"Proceedings of the 13th Language Resources and Evaluation Conference"*. "Marseille, France": "European Language Resources Association", june "2022".

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] Statistisches Bundesamt, "Klassifikation der Wirtschaftszweige," Wiesbaden, 2008. [Online]. Available: https://www.destatis.de/static/DE/dokumente/klassifikation-wz-2008-3100100089004.pdf

APPENDIX

(I) INFO ABOUT WZ08

A. Classification of industrial sectors

Classification of industrial sectors include classifications for international comparative research (e.g., NACE) and administrative subdivisions (such as the Eurostat definition for knowledge- and technology-intensive sectors based on NACE), see our discussion in [1]. These classifications are usually interrelated. For example, the Classification of Economic Activities (WZ) is developed by the Federal Statistical Office of Germany and has been refined since 1950, with WZ 2008 being the latest edition which we will discuss later. Its objective is to ensure uniformity in the classification of economic activities across all official statistics in Germany. The classification is hierarchically structured.

The "Klassifikation der Wirtschaftszweige" (Classification of Branches of Industry, abbreviated as WZ) is utilized in Germany, particularly by the Statistische Bundesamt (Federal Statistical Office), for the classification of employers' economic activities in official statistics. The latest version is WZ 2008[3], which renders WZ 2003 and 1993 obsolete. This classification is compatible with the European "Nomenclature statistique des activités économiques dans la Communauté européenne" (NACE), but it includes more detailed data. NACE is the European classification system developed by Eurostat in the 1970s and updated regularly, with NACE Rev. 2 being the latest version from 2008. It provides a framework for collecting and presenting statistical data by economic activity and is hierarchically structured. For further information, please see [41]. Similar to NACE, WZ 2008 provides several hierarchical levels. A first level describes 21 sections (letters A-U), a second divisions, a third groups, a fourth classes. In contrast to NACE, WZ 2008 adds subgroups as fifth level, which is, however, only added to particular classes. WZ08 includes Sections (21), A-U, Divisions (88), 01-99, Groups (272), 01.1-99.0, Classes (615), 01.11-99.00, and Sub-classes (839), 01.11.0-99.00.0, see Figure 2.

While sectors are broad and specific, for example A (Agriculture, Forestry and Fishing) and B (Mining and Quarrying), others lack clear definition at this level, for example S (Other Service Activities). Conversely, classes and groups often exhibit indistinguishable characteristics, and the designation of divisions and groups typically provides minimal additional insight (e.g., 77 "Rental and leasing activities" versus 77.1 "Renting and leasing of motor vehicles". Furthermore, a company may belong to multiple industrial sectors, such as several manufacturing divisions. Nevertheless, the official guidelines recommend labeling the most dominant sector. Consequently, while the taxonomy of industrial sectors is well-defined by WZ08, we rely on external data to train and evaluate our approaches.

The International Standard Industrial Classification of All Economic Activities (ISIC) is a United Nations system for classifying economic activities, updated periodically since 1948, with ISIC Rev. 4 being the latest version. ISIC underpins both NACE and the German WZ. Its objective is to provide categories for the collection and reporting of statistics, and it is hierarchically structured. Other approaches are based on some of these taxonomies. For instance, the Eurostat classification of NACE sectors is based on technology intensity (manufacturing) and knowledge intensity (services). Manufacturing industries are classified as high, medium-high, medium-low, or low technology, while services are divided into knowledge-intensive and less-knowledge-intensive categories. This classification is available for both NACE Rev. 1.1 and NACE Rev. 2 (Eurostat 2009).

(II) DATA DISTRIBUTION AND EXAMPLES

B. Frequencies

C. Some samples of the dataset:

Given is a jobposting where a job for a roofer (Section F) is advertised, but the company posting the ad belongs to

[3] All data is accessible in both English and German at https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-englisch.html. In this text, we generally use the official English translation for examples.

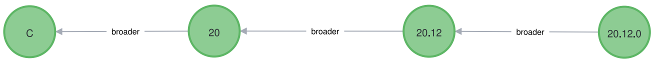


Fig. 2. An example subset of WZ08: Sector C (Manufacturing), division 20 (Manufacture of chemicals and chemical products), group 20.12 and class 20.12.0 (Manufacture of dyes and pigments).

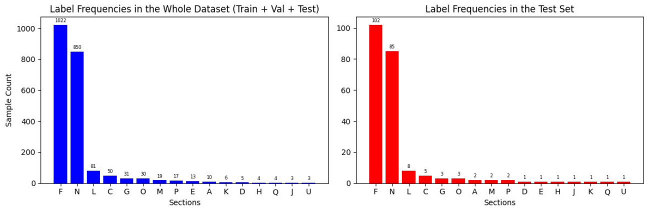


Fig. 3. Data frequencies of labels for whole dataset (left) and test set (right)

section N. Please note that we removed names of companies and any other sensetive information and replaced it with `[REMOVED]` in the following texts.

```
{'section_letter' : 'N',
 'text' : "Stellenangebot - Dachdecker/ Bauklempner zur Festeinstellung
gesucht (m/w) (Dachdecker/in und Bauklempner/in)

Überblick über das Stellenangebot

Referenznummer

[REMOVED]
Titel des Stellenangebots

Dachdecker/ Bauklempner zur Festeinstellung gesucht (m/w) (Dachdecker/in und Bauklempner/in)
Alternativberufe
Helfer/in - Hochbau
Konstruktionsmechaniker/in - Feinblechbautechnik
Stellenangebotsart

Arbeitsplatz (sozialversicherungspflichtig)
Arbeitgeber
[REMOVED]

Branche: Vermittlung von Arbeitskraeften, Betriebsgroeße: zwischen 51 und 500
Stellenbeschreibung

Im Auftrag unserer Partnerunternehmen suchen wir zur sofortigen Festanstellung mehrere
gelernte Dachdecker, Spengler, Dachdeckerhelfer (m/w) mit Berufserfahrung in Vollzeit.

Der Einsatz erfolgt Überwiegend im Süddeutschen Raum.
Ein Führerschein der Klasse3 (C1) ist von Vorteil, jedoch keine Bedingung.

Bei Interesse an diesem Angebot senden Sie uns bitte Ihre Bewerbungsunterlagen zu
(vorzugsweise per Email).
Ihre Aufgabe ist die Mithilfe bei:
- Bedachungen
- Dachum- und Ausbauarbeiten
- Modernisierungen
- Einbau von Dachfenstern und Gauben
- Wartung- und Reparaturarbeiten

Unsere Anforderungen an Sie:
- körperliche Fitness, Teamfähigkeit und selbständiges Arbeiten

Haben wir Ihr Interesse geweckt? Dann freuen wir uns auf Ihre Kontaktaufnahme."
}
```

Another example: A jobposting for a Product Manager posted by a company working in the industrial section G.

```
{'section_letter' : 'G',
'text' : "Produktmanager (m/w/d) Dachflächenfenster & Stahlelemente
DE, Germany
[REMOVED]
[REMOVED]

nach Vereinbarung / Qualifikation

Die [REMOVED] für Baustoffe, [REMOVED], ist mit einem Gruppenumsatz von
6,9 Mrd. Euro (2021) und über 1.500 Standorten in [REMOVED] eine der marktführenden
 Kooperationen im Baustoff-, Holz- und Fliesenhandel.
Auch in der Do-it-yourself-Branche nimmt das Unternehmen mit den [REMOVED] eine
führende Position ein. Zur Dienstleistungszentrale gehören als Tochterunternehmen
[REMOVED], die [REMOVED],
der [REMOVED] Versicherungsdienst, die [REMOVED] Logistik sowie die [REMOVED]
Beratungs- und
Beteiligungsgesellschaft und hagedoo mit insgesamt ca. 1.400 Mitarbeitern. Sie unterstützen
die Gesellschafter der [REMOVED]-Kooperation flächendeckend in sämtlichen
Bereichen deren unternehmerischen Handelns.

Für die Abteilung Einkaufssteuerung Logistikfachhandel unserer Zentrale in
[REMOVED] suchen wir ab
sofort in Vollzeit einen
Produktmanager (m/w/d) Dachflächenfenster & Stahlelemente
* Kontinuierliche Analyse und Weiterentwicklung aller Sortimente von [REMOVED] Logistik
* Sicherstellung der Einhaltung aller Prozesse des Warenflusses, insbesondere sämtlicher
 Listungstätigkeiten wie Stammdaten, Ein- und Verkaufspreise sowie Logistikparameter
* Abstimmung mit Lieferanten, dem Einkauf und der Logistik zur Aufrechterhaltung
der Lieferfähigkeit
* Erstellung von Vorgaben für die Abteilungen Customer Service und Beschaffung
* Projektarbeit zur Optimierung von Sortimenten und Prozessen
* Monitoring warenwirtschaftlicher Kennziffern

* Abgeschlossene kaufmännische Ausbildung oder Studienabschluss in einem für diese Position
relevanten Fachgebiet
* Berufserfahrung im Baustoffhandel oder in der Industrie im Ein- bzw. Verkauf ist von Vorteil
* Gute MS Office-Kenntnisse werden vorausgesetzt
* Grundlegende Sortiments- und Lieferantenkenntnisse im Bereich Dachflächenfenster oder
Stahlelemente sind wünschenswert
* Analytische Denk- und strukturierte Vorgehensweise sowie Innovationskraft und Vorwärtsdrang
* Selbstständigkeit, ausgeprägte Kommunikationsstärke, Überzeugungkraft und Belastbarkeit
* 30 Tage Urlaub
* Weihnachts- und Urlaubsgeld
* Mobiles Arbeiten und flexible Arbeitszeitkonten inklusive Gleitzeittagen
* Spannendes Aufgabengebiet und eigenverantwortliches Arbeiten
* Weiterentwicklungsmöglichkeiten durch Projektarbeit sowie weitergehende
Qualifizierungsmöglichkeiten
* Angenehme und offene Arbeitsatmosphäre in einem hochmotivierten und sympathischen
Team mit Patenmodell
* [REMOVED] mit diversen Vergünstigungen
* Rabatte bei regionalen Partnern (u. a. Heide Park, Fitnessstudio)
* Eigener Versicherungsdienst und vermögenswirksame Leistungen
* Betriebliche Benefits wie u.a. eine Kantine, Firmenfeiern, Gesundheitswochen und
freies WLAN für Mitarbeiter

Auf einen Blick:
* Bereich: Produktmanagement
* Einsatzort: [REMOVED]
* Arbeitszeit: 38,5 Stunden/Woche
* Eintrittstermin: ab sofort
* Arbeitsverhältnis: unbefristet"
}
```

DOI: 10.15439/2024F3018
ISSN 2300-5963 ACSIS, Vol. 40

Transforming Attribute-Based Encryption schemes into Asymmetric Searchable Encryption schemes

Elisa Giurgea
0009-0004-3610-3666
Alexandru Ioan Cuza University,
Department of Computer Science,
Iasi, Romania
Email: elisa.giurgea@gmail.com

***Abstract*—Attribute-Based Encryption (ABE) and Asymmetric Searchable Encryption (ASE) are two highly useful Public-Key Encryption (PKE) technologies in today's cloud computing landscape.**

By leveraging the idea that the attributes from ABE can serve as keywords for ASE, we propose an efficient technique to translate any ABE schemes into ASE schemes. We address both the case of Ciphertext-Policy Attribute-Based Searchable Encryption (CP-ABSE) and Key-Policy Attribute-Based Searchable Encryption (KP-ABSE) schemes.

Our main goal with these schemes is to maintain the security properties of ABE while introducing efficient search capabilities, thereby facilitating further advancements in ASE development. To validate our theoretical proposals, we have analyzed their practical applicability using existing ABE implementations.

I. Introduction

SEARCHABLE Encryption (SE) is a cryptographic technique that allows users to search over encrypted data without decrypting it, preserving data confidentiality while enabling search functionality. SE can be broadly categorized into two types: Symmetric Searchable Encryption (SSE) [1] and Asymmetric Searchable Encryption (ASE) [2].

Symmetric Searchable Encryption schemes utilize a single secret key that is shared between the data owner and the authorized users. This approach offers several advantages: efficiency in terms of computational overhead and search speed, making them suitable for scenarios where performance is critical; and simplicity, as the key management in SSE is straightforward since it involves only one key for both encryption and decryption. However, SSE also has notable drawbacks, such as limited access control with all the users with the secret key being able to access all the data, which may not be desirable in many applications; and scalability issues, as securely distributing and managing the single secret key becomes challenging with the number of system users increasing.

In contrast, Asymmetric Searchable Encryption employs a pair of cryptographic keys: a public key for encryption and a private key for decryption. This approach addresses some of the limitations of SSE by enabling more granular access control, allowing the data owner to specify which users can access which data. This is particularly useful in multi-user environments with varying access privileges. Moreover, ASE offers better scalability, each user having their own key pair, simplifying key distribution and management as the system scales up. Despite these advantages, ASE comes with its own set of challenges, as the use of public and private keys introduces additional complexity in terms of key management and the overall cryptographic operations. ASE schemes often incur higher computational costs and longer search times compared to SSE, which can be a major drawback in resource-constrained environments.

Given the trade-offs between SSE and ASE, there is a significant interest in developing more efficient ASE schemes that can leverage the strengths of existing cryptographic methods, as seen with the recently proposed [7], [3], [4], [5] and [6]. One promising approach is to derive ASE schemes from Attribute-Based Encryption (ABE). By treating attributes in ABE as keywords in ASE, it is possible to create systems that offer both efficient search capabilities and robust access control. Recently, the idea of treating attributes as keywords was also used by Long Meng, Liqun Chen and Yangguang Tian as a trivial assumption behind a new proposal of an efficient ASE scheme extended from an A-KP-ABE scheme [7].

This paper aims to propose a new class of efficient ASE schemes derived from ABE. Specifically, we present theoretical formalizations for Ciphertext-Policy Attribute-Based Searchable Encryption (CP-ABSE) and Key-Policy Attribute-Based Searchable Encryption (KP-ABSE), showing that any ABE scheme can become an ASE scheme. Our objective is to utilize the inherent advantages of ABE, such as fine-grained access control and scalability, to develop ASE schemes that offer efficient, secure, and flexible search functionality.

II. General transformation of ABE in ASE

In ABE, attributes can be viewed as keywords. By treating these attributes as keywords in ASE, we can introduce search functionalities without compromising the existing access control mechanisms. This allows for the creation of searchable indexes based on attributes or access policies, enabling efficient retrieval of encrypted data based on specified search criteria.

Building on this concept of treating ABE attributes as keywords for ASE, having as basis the general schemes from [8], we will be further presenting the formalization for the extension of both ABE forms to ASE. The notations of general functions needed in the formalization are subsequently explained in “Table 1”.

A. CP-ABSE formalization

Algorithm 1 CP-ABSE GlobalInit

```
    $index \leftarrow [\,][\,]$; $H \leftarrow HashFunction()$;
```

Algorithm 2 CP-ABSE Setup

```
Require: Security parameter $1^\lambda$;
Ensure: Public parameters $PK$, master key $msk$;
    $(PK, msk) \leftarrow Setup(1^\lambda)$;
    return $(PK, msk)$;
```

For Ciphertext Attribute-Based Encryption schemes, the setup algorithm of the system, $Setup(1^\lambda)$, takes as input a security parameter, 1^λ and outputs the public parameters, PK, and a master key, msk, which is known only to the private key generator (PKG). Aside from the CP-ABE setup related initialisation, we are also adding to the system 2 new global parameters: $index$, which will store the encrypted documents/messages at a given index, and H, a hash function which will be further used in the index creation and query generation steps.

Algorithm 3 CP-ABSE KeyGen

```
Require: Public parameters $PK$, master key $msk$, set of
    attributes $\gamma = \{a_1, a_2, \ldots, a_n\}$;
Ensure: Private key $D_\gamma$;
    $D_\gamma \leftarrow KeyGen(PK, msk, \gamma)$;
    return $D_\gamma$;
```

Next, onto the key generation algorithm, there are no changes from the regular behaviour expected of a CP-ABE scheme: $KeyGen(PK, msk, \gamma)$ algorithm takes as input the public parameters, PK, the master key, msk, both generated during the setup phase, and a set of attributes, $\gamma = \{a_1, a_2, \ldots, a_n\}$, for which the secret key is to be generated. It outputs the private key D_γ which encapsulates γ.

Algorithm 4 CP-ABSE Encrypt

```
Require: Message $m$, public parameters $PK$, access policy
    $A = BooleanExpr(\{a_1, a_2, \ldots, a_n\})$;
Ensure: Ciphertext $ct$;
    $ct \leftarrow Encrypt(m, PK, A)$;
    return $ct$;
```

The encryption algorithm also remains the same as with regular CP-ABE: $Enc(m, PK, A)$ takes as input parameters a message, m, an access policy, A (which is or can be trivially transformed into a logical boolean expression $A = BooleanExpr(\{a_1, a_2, \ldots, a_n\})$), and the public parameters, PK. It outputs the ciphertext, ct.

Algorithm 5 CP-ABSE CreateIndex

```
Require: Ciphertext $ct$, same access policy
    $A = BooleanExpr(\{a_1, a_2, \ldots, a_n\})$;
    $A' \leftarrow A$;
    $\gamma' \leftarrow ParsePolicy(A)$;
    for $a'_i$ in $\gamma'$ do
        $A'.replace(a'_i, H(a'_i))$;
    end for
    $id \leftarrow ct$;
    if $\exists index[A']$ in $index$ then
        $index[A'].append(id)$;
    else
        $index[A'] \leftarrow [id]$;
    end if
```

To ensure that the documents encrypted by the system are stored and searchable, we are executing $CreateIndex(ct, A)$ right after a document/message is encrypted based on a given policy. The algorithm is simply storing the ciphertexts at an index obtained from the same policy A based on which the message was encrypted in the previous step.

To further hide the index values, the attributes from the given access policy A (obtained via parsing the policy) will be further hashed with the system defined hash function H, and their clear-text values will be subsequently replaced in the policy with the hashed ones. The initial logical relations between the attributes in the policy will be kept for the newly obtained hashed policy, $A' = BooleanExpr(\{H(a_1), H(a_2), ..., H(a_n)\})$.

After the execution of the aforementioned operations, we will now append the document id for the ciphertext (in this demonstration case, the document id is the ciphertext itself) to the index of the hashed policy to store it.

These steps are to be repeated as needed for the multiple users which are to use the system and the multiple protected-access documents which need to be stored.

Now, how can a system user with a secret key, D_γ, get their list of available documents given the user's attached attributes?

Algorithm 6 CP-ABSE GenerateQuery

```
Require: Private key $D_\gamma$;
Ensure: Query $Q = \{H(a_1), H(a_2), ..., H(a_n)\}$;
    $\gamma \leftarrow ExtractAttributes(D_\gamma)$;
    $Q \leftarrow \{\}$;
    for $a_i$ in $\gamma$ do
        $Q.append(H(a_i))$;
    end for
    return $Q$;
```

First, the $GenerateQuery(D_\gamma)$ algorithm will be executed with the user's secret key as an input parameter. The generation of the query itself is based on the set of attributes, γ, encapsulated in the secret key D_γ.

The set of attributes, γ, is extracted from D_γ (the extraction method itself is specific to the proposed CP-ABSE scheme)

TABLE I
FUNCTIONS USED IN THE FORMALIZATIONS AND THEIR MEANING

Notation	First operation
$HashFunction()$	Any trapdoor or one-way hash function desired to hide the values
$BooleanExpr(set\ of\ attributes)$	An access control policy obtained as a boolean expression over a given set of attributes
$EvaluatePolicy(access\ policy,\ set\ of\ attributes)$	Function to evaluate whether a given set of attributes satisfies a given access policy
$ParsePolicy(access\ policy)$	Function to extract the attributes used in a private policy in a set of attributes
$ExtractAttributes(CP-ABE\ private\ key)$	Function to extract the attributes encapsulated in a given CP-ABE specific private key
$ExtractPolicy(KP-ABE\ private\ key)$	Function to extract the access policy encapsulated in a given KP-ABE specific private key

and each of them are hashed with the same one-way hash function H used for the attributes in the access policy during the index creation step to ensure unitary information. The hashed extracted attributes are next compacted into a set which will be returned as the search query, $Q = \{H(a_1), H(a_2), ..., H(a_n)\}$.

Algorithm 7 CP-ABSE Search

```
Require: Query $Q$;
Ensure: List of document ids $results = \{id_1, id_2, ..., id_n\}$;
  $results \leftarrow \{\}$
  for $(A', ids)$ in $index$ do
    if $EvaluatePolicy(A', Q) = TRUE$ then
      $results.extend(ids)$;
    end if
  end for
  return $results$;
```

After successfully generating a query, Q, we can now search for the encrypted documents which are satisfying Q by executing the $Search(Q)$ algorithm, which is to return the document ids (in this demonstration, the ciphertexts) from the $index$. The search itself is based on evaluating the hashed policies indices against query Q, which contains a list with the user's hashed attributes.

How would this evaluation work?

The evaluation step is quite trivial, as, having the policy, $A' = BooleanExpr(\{H(a_1), H(a_2), ..., H(a_n)\})$, in order for it to be satisfied the logical expression needs to be $TRUE$. And how do we check that having the set of hashed attributes, $Q = \{H(a_1), H(a_2), ..., H(a_n)\}$? For each hashed attribute in the set, we are replacing its appearances in the hashed policy with the $TRUE$ value. For the rest of the hashed attributes in the hashed policy which were not found in the set of hashed attributes given as an input parameter, we are replacing their appearances with the $FALSE$ value, as them not being in the set implies that the user does not have them attached to its secret key, D_γ. After obtaining the policy to be evaluated under a format such as $(TRUE\ and\ FALSE)\ or\ TRUE$, we are evaluating it using a built-in default boolean $eval$ function.

In case of the evaluation step for a certain indexed hashed policy returning $TRUE$, the document ids indexed under the aforementioned policy will be collected in a set of $results$. We will be repeating the step for all the indexes to retrieve all the document ids which would be viable results for the search query Q.

Algorithm 8 CP-ABSE Decrypt

```
Require: Ciphertext $ct$, public parameters $PK$, private key $D_\gamma$;
Ensure: Message $m'$;
  $m' \leftarrow Decrypt(ct, PK, D_\gamma)$;
  return $m'$;
```

Now, after obtaining the set of $results$, for each $result$ (the document id being the actual ciphertext in this case) and we will be proceeding with the CP-ABE specific decryption algorithm, $Decrypt(result, PK, D_\gamma)$. The decryption algorithm takes as input the ciphertext, $result$, which was encrypted with an access policy, A, the public parameters, PK, and the private key of the user, D_γ. It outputs the initially encrypted document/message, if the set of attributes, γ, encapsulated in the private key, D_γ, satisfies the access policy A.

Algorithm 9 CP-ABSE Usage Example

```
$GlobalInit()$;
$(PK, msk) \leftarrow Setup()$ {Global system setup}
$\gamma \leftarrow \{a_1, a_2, ..., a_n\}$;
... {Attribute initialization for system users}
$D_\gamma \leftarrow KeyGen(PK, msk, \gamma)$;
... {System registration / private key generation for system users}
Get messages / documents to be encrypted and stored;
Define necessary access policies;
$ct \leftarrow Encrypt(m, PK, A)$; $CreateIndex(ct, A)$;
$Q \leftarrow GenerateQuery(D_\gamma)$;
$results \leftarrow Search(Q)$;
for $result$ in $results$ do
  $m' \leftarrow Decrypt(result, PK, D_\gamma)$;
end for
```

B. KP-ABSE formalization

Algorithm 10 KP-ABSE GlobalInit

```
$index \leftarrow [\ ][\ ]$; $H \leftarrow HashFunction()$;
```

Algorithm 11 KP-ABSE Setup

```
Require: Security parameter $1^\lambda$;
Ensure: Public parameters $PK$, master key $msk$;
    $(PK, msk) \leftarrow Setup(1^\lambda)$;
    return $(PK, msk)$;
```

For Key Policy Attribute-Based Encryption schemes, the setup algorithm of the system, $Setup(1^\lambda)$, is formalized similarly to the CP-ABE one, with it taking a security parameter, 1^λ and outputing the public parameters, PK, and a master key, msk, known only to the PKG. Likewise to previously formalized CP-ABSE scheme, we are adding the 2 global parameters: $index$ and H.

Algorithm 12 KP-ABSE KeyGen

```
Require: Public parameters $PK$, master key $msk$, access
    policy $A = BooleanExpr(\{a_1, a_2, \ldots, a_n\})$;
Ensure: Private key $D_A$;
    $D_A \leftarrow KeyGen(PK, msk, A)$;
    return $D_A$;
```

Now, for the key generation algorithm KeyGen (PK, msk, A) we are keeping the same on as it was for KP-ABE schemes in general: by taking as input the previously generated PK and msk and an access policy, $A = BooleanExpr(\{a_1, a_2, \ldots, a_n\})$, the algorithm outputs the private key, D_A, which encapsulates the access policy based on which it was generated.

Algorithm 13 KP-ABSE Encrypt

```
Require: Message $m$, public parameters $PK$, set of attributes
    $\gamma = \{a_1, a_2, \ldots, a_n\}$;
Ensure: Ciphertext $ct$;
    $ct \leftarrow Encrypt(m, PK, \gamma)$;
    return $ct$;
```

For KP-ABE based KP-ABSE, the encryption step also remains the same to the original, with $Enc(m, PK, \gamma)$ getting as input a document/message, m, a set of attributes, $\gamma = \{a_1, a_2, \ldots, a_n\}$, and the public parameters, PK and returning the ciphertext, ct, generated based on the attributes given.

Algorithm 14 KP-ABSE CreateIndex

```
Require: Ciphertext $ct$, same set of attributes $\gamma =
    \{a_1, a_2, \ldots, a_n\}$;
    $\gamma' \leftarrow \{\}$;
    for $a'_i$ in $\gamma'$ do
        $\gamma'.add(H(a'_i))$;
    end for
    $id \leftarrow ct$;
    if $\exists index[\gamma']$ in $index$ then
        $index[\gamma'].append(id)$;
    else
        $index[\gamma'] \leftarrow [id]$;
    end if
```

After the encryption step, we are indexing the ciphertext in the system by executing $CreateIndex(c, \gamma)$ algorithm, which is storing the ciphertext at an index generated based on the same set of attributes, γ, on which the ciphertext was encrypted. To hide the index values, the attributes are be hashed with the system defined hash function H and compacted in a set. Next, we will append the document id for the ciphertext (again, the document id is the ciphertext itself) to the index of the hashed attributes set object.

The key generation is to be done for the needed number of users, and, for the documents/messages to be stored, the encryption and index creation algorithms are to be executed for whichever number of documents necessary.

When a search is wanted to be performed over the stored encrypted documents, the following algorithms will be executed:

Algorithm 15 KP-ABSE GenerateQuery

```
Require: Private key $D_A$;
Ensure: Query $Q =$
    $BooleanExpr(\{H(a_1), H(a_2), ..., H(a_n)\})$, the hashed
    policy;
    $A' \leftarrow ExtractPolicy(D_A)$;
    $\gamma' \leftarrow ParsePolicy(A')$;
    $Q \leftarrow A'$;
    for $a'_i$ in $\gamma'$ do
        $Q.replace(a'_i, H(a'_i))$;
    end for
    return $Q$;
```

First, we need to generate the search query based on the user's secret key, $GenerateQuery(D_A)$. The KP-ABE secret key encapsulates the access policy attributed to the user and, in order to generate the query, we will first be extracting the policy from the key. From the obtained policy, we will now be extracting the attributes. For each attribute's occurrences in the policy, we will be replacing its appearance with its hashed value, $H(a_i)$. Thus, we will be obtaining a hashed policy, which will be acting as our search query, $Q = BooleanExpr(\{H(a_1), H(a_2), ..., H(a_n)\})$.

Algorithm 16 KP-ABSE Search

```
Require: Query $Q$;
Ensure: List of document ids $results = \{id_1, id_2, ..., id_n\}$;
  $results \leftarrow \{\}$
  for $(\gamma', ids)$ in $index$ do
    if $EvaluatePolicy(Q, \gamma') = TRUE$ then
      $results.extend(ids)$;
    end if
  end for
  return $results$;
```

Having the query, Q, the user can now perform a search for the secret documents which are satisfying Q with the $Search(Q)$ algorithm, which will be returning a list with the accessible document ids/ciphertexts. The search is based on evaluating the hashed policy from the query against all the indices, each of them being a compacted set of hashed attributes.

When the evaluation for a certain indexed attribute set returns $TRUE$, the document ids indexed under it will be collected in a set of $results$.

Algorithm 17 KP-ABSE Decrypt

```
Require: Ciphertext $ct$, public parameters $PK$, private key $D_\gamma$;
Ensure: Message $m'$;
  $m' \leftarrow Decrypt(ct, PK, D_\gamma)$;
  return $m'$;
```

For each $result$ (with the document id as the ciphertext) obtained in the $results$ set, the KP-ABE specific decryption algorithm, $Decrypt(result, PK, D_A)$ will be executed. Having as input parameters $result$, which was encrypted with a set of attributes, γ, the public parameters, PK, and the secret key of the user, D_A, the decryption algorithm outputs the initially encrypted document/message, if the set of attributes, γ, satisfies the access policy A, encapsulated in the private key, D_A.

Algorithm 18 KP-ABSE Usage Example

```
  $GlobalInit()$;
  $(PK, msk) \leftarrow Setup()$ {Global system setup}
  $A \leftarrow BooleanExpr(\{a_1, a_2, ..., a_n\})$;
  ... {Access policy initialization for system users}
  $D_A \leftarrow KeyGen(PK, msk, A)$;
  ... {System registration / private key generation for system users}
  Get messages / documents to be encrypted and stored;
  Define attribute sets;
  $ct \leftarrow Encrypt(m, PK, \gamma)$; $CreateIndex(ct, \gamma)$;
  $Q \leftarrow GenerateQuery(D_A)$;
  $results \leftarrow Search(Q)$;
  for $result$ in $results$ do
    $m' \leftarrow Decrypt(result, PK, D_A)$;
  end for
```

III. Security

A. CP-ABSE

The initial assumption is that the security of the CP-ABSE scheme is similar to the CP-ABE scheme on which it is based. We will further demonstrate that the additional operations introduced for searchability do not compromise the security guarantees provided by base CP-ABE. Specifically, we will sketch the arguments as to why the confidentiality of the encrypted documents and the privacy of the attributes from the access policies are preserved.

1) CP-ABE Security Model: The $Setup$ algorithm security for CP-ABE implies that the algorithm generates the public parameters, PK, and a master secret key, msk. The security requirement is that without msk, it is not feasible for adversary to generate valid private keys for any given attribute set.

The $KeyGen$ algorithm generates a private key D_γ for a set of attributes γ. The security requirement is that an adversary with some private keys $D_{\gamma 1}$, $D_{\gamma 2}$, $D_{\gamma 3}$..., $D_{\gamma k}$ cannot generate a valid private key for a new attribute set γ', unless γ' is a subset of one of the sets γi where $1 \leq i \leq k$.

Encryption and decryption security wise, the $Encrypt$ algorithm ensures that a ciphertext ct encrypted under an access policy A can only be decrypted by a private key D_γ if γ satisfies A. The security requirement is that an adversary should not be able to decrypt ct without possessing such a D_γ.

2) CP-ABSE General Scheme Overview: In addition to the CP-ABE setup, CP-ABSE introduces an $index$ structure and a hash function H as system global parameters. Algorithms wise, it introduces $CreateIndex$, which is executed after encryption and it indexes the ciphertext using a hashed version of the access policy; $GenerateQuery$, which generates queries by hashing the attributes in the user's private key; and $Search$, which evaluates the hashed query against the hashed policies in the index and returns available results.

Having these added algorithms, we need to take the following security proof components into account:

- The confidentiality of the encrypted documents: Since the encryption and decryption processes in CP-ABSE are identical to CP-ABE, the confidentiality of the encrypted documents relies on the security of the underlying CP-ABE scheme. Any attack on the confidentiality of CP-ABSE ciphertexts can be reduced to an attack on CP-ABE ciphertexts; the $index$ stores only hashed versions of the access policies and does not reveal any additional information about the plaintext or the original attributes. The hash function H should be an irreversible collision-free hash function, ensuring that the hashed values do not leak information about the original attributes.
- The privacy of the attributes in the access policies: The security of the hashed attributes in the index access policies and queries depends on the hash function H, which should be chosen to ensure that it is computationally infeasible to reverse-engineer the original attributes from their hashed values. The $GenerateQuery$ process uses

the same hash function H, ensuring that the attributes in the user's private key are protected in the same way as the attributes in the access policies.

- The search operation security: The search operation involves evaluating whether the hashed query satisfies the hashed policies in the index. This process does not reveal any additional information about the original attributes or the plaintext, as it only involves hashed values.

B. KP-ABSE

Similarly to CP-ABSE, the initial assumption is that the security of the KP-ABSE scheme is the same as it is for the base KP-ABE scheme. We will demonstrate that the additional operations introduced to enable searchability do not compromise the security guarantees provided by the underlying KP-ABE scheme; this will be accomplished by showing that both the confidentiality of the encrypted documents and the privacy of the attributes are maintained.

1) KP-ABE Security Model: The security of the $Setup$ algorithm in KP-ABE ensures that it generates public parameters, PK, and a master secret key, msk. The security condition is that, without access to msk, it is infeasible for an adversary to create valid private keys for any access control policy.

For the $KeyGen$ algorithm, it produces a private key D_A corresponding to a specific access policy A. The security condition here is that even if an adversary has access to private keys D_{A1}, D_{A2}, D_{A3}, $\ldots$, D_{Ak}, they cannot derive a valid private key for a new access policy A', unless A' is logically equivalent to one of the known policies A_i where $1 \leq i \leq k$.

In terms of encryption and decryption security, the $Encrypt$ algorithm guarantees that a ciphertext ct encrypted with a set of attributes γ can only be decrypted by a private key D_A if the access policy A is satisfied by γ. The security requirement is that an adversary should not be able to decrypt ct without having an appropriate D_A.

2) KP-ABSE General Scheme Overview: In addition to the standard KP-ABE setup, KP-ABSE includes an $index$ structure and a hash function H as global system parameters. Algorithmically, it introduces several new components: $CreateIndex$, which is run post-encryption to index the ciphertext based on a hashed version of the attribute set used for encryption; $GenerateQuery$, which produces queries by hashing the attributes in the user's access policy contained in their private key; and $Search$, which matches the hashed query against the hashed attribute sets in the index to return relevant results.

Given the additional algorithms, we must consider the following security proof components:

- The confidentiality of the encrypted documents: Since the encryption and decryption processes in KP-ABSE are the same as in KP-ABE, the confidentiality of the encrypted documents depends on the security of the underlying KP-ABE scheme. Any attack on the confidentiality of KP-ABSE ciphertexts reduces to an attack on KP-ABE ciphertexts. The $index$ only stores hashed versions of the attribute sets, ensuring no extra information about the plaintext or original attributes is revealed. The hash function H must be an irreversible, collision-free one-way function to prevent leakage of information about the original attributes.
- The privacy of the access policies: The security of the hashed attributes in the index and queries relies on the hash function H. This function must be chosen such that it is computationally infeasible to derive the original attributes from their hashed values. The $GenerateQuery$ algorithm also uses the hash function H, ensuring that the access policies in the user's private key are protected similarly to the attribute sets in the indices.
- The search operation security: The search operation involves checking whether the hashed query satisfies the hashed attribute sets in the index. This evaluation process only uses hashed values and does not reveal any additional information about the original attributes or the plaintext.

In our following work, we plan to formally prove that any adversary who can break the security of a CP-ABSE or KP-ABSE scheme with non-negligible probability can also break the security of the CP-ABE, respectively the KP-ABE on which it is based, which implies that the security of the ASE scheme built on top of the ABE scheme with the keywords acting as attributes idea is at least as strong as the ABE scheme.

IV. Practical Implementations

The proposed CP-ABSE and KP-ABSE scheme structures are expected to be generally applicable for all existent ABE schemes with the addition of the specific index creation, query generation and search algorithms. In order to test the feasibility and applicability of the transformations proposed, several existent implementations for both CP-ABE and KP-ABE were extended with the aforementioned steps in order for them to become working ASE schemes.

The batch of initial implementations were taken from the open source toolbox library Charm [15], used for prototyping cryptosystems based on a series of provided schemes. Its implementation is done in Python by representatives of Johns Hopkins University as part of their Advanced Research in Cryptography laboratory [16].

From the ABENC [17] schemes package of the library, several schemes were successfully adapted by having the 3 algorithms added to them. The transitioning from ABE to ASE with the 3 steps has not impacted any of the functionality available for the implemented ABE schemes, including user attribute revocation, user access policy adjustments, attribute accountability hiding, policy accountability hiding and other bonus functionalities of the given ABE schemes.

For demonstrative implementations, *SHA256* was used was used as the one-way hash function, due to its efficiency and the fact that it is collision-free, but *SHA256* hashes do not include the salting element. That makes the hashes more susceptible to dictionary-based cyber attacks [18]. Thus, while *SHA256* is more suitable for applications that require frequent interaction, *bcrypt* [18] is a better solution for safely storing the hashed policies at the expense of efficiency. The proposed extensions are not limiting to a certain hash function, policy structure or index secret document storage method and are meant to be a comprehensible base for the further development of ASE schemes, based on either CP-ABE, or KP-ABE schemes.

Taking that into consideration, the following schemes from the library were successfully adapted and tested:

- DAC-MACS [9], a multi-authority CP-ABE scheme with efficient decryption and authority revocation implementations, proposed by Kang Yang, Xiaohua Jia, K. Ren and B. Zhang;
- The next scheme proposed by Kang Yang, Xiaohua Jia, the Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage CP-ABE scheme [10];
- The KP-ABE lightweight attribute-based encryption scheme for the Internet of things [11], proposed by Xuanxia Yao, Zhi Chen and Ye Tian;
- From Attribute Based Encryption with Privacy Protection and Accountability for CloudIoT [12], proposed by Jiguo Li, Yichen Zhang, Jianting Ning, Xinyi Huang, Geong Sen Poh and Debang Wang, the first proposed pairing-based scheme for policy-hiding CP-ABE;
- One of the first proposed CP-ABE schemes, in Ciphertext-Policy Attribute-Based Encryption [13], by John Bethencourt, Brent Waters; this scheme is a basic pairing-based CP-ABE scheme;
- The Rouselakis - Waters Efficient Statically-Secure Large-Universe Multi-Authority Attribute-Based Encryption scheme [14], based on a bilinear pairing group of prime order;

The added algorithm which needed the most scheme specific implementation was the $GenerateQuery$ one, given the different formats of the private keys employed by each adapted scheme.

The development environment utilized for this project was PyCharm Community 2024 [19] with Python 3.12, running within a Parallels-managed virtual machine [20] on Ubuntu 22.04. The implementations were developed as prototypes, leveraging the open-source nature of the base ABE implementations designed for educational purposes in Python, which inherently did not prioritize maximum time efficiency.

Benchmark tests conducted post-adaptation indicated inconsistencies in execution times across multiple trials of the same code. Notably, there were instances where the more complex ASE implementations executed faster than expected, suggesting the influence of external factors related to the development environment. Consequently, these benchmark results should be interpreted with caution and are not wholly reliable.

V. Acknowledgments

I would like to express my gratitude towards my MSc dissertation supervisor, Prof. Dr. Ferucio Laurenţiu Ţiplea, for the guidance and support throughout the course of this research. Prof. Dr. Ferucio Laurenţiu Ţiplea's advice and insightful feedback were instrumental in shaping the direction and outcome of this study.

VI. Conclusion

The formalization of transforming Attribute-Based Encryption (ABE) schemes into Asymmetric Searchable Encryption (ASE) schemes is an important step forward in the field of cryptography. By treating attributes in ABE as keywords in ASE, efficient search capabilities are introduced, all while preserving the robust security properties of ABE. This theoretical framework for Ciphertext-Policy Attribute-Based Searchable Encryption (CP-ABSE) and Key-Policy Attribute-Based Searchable Encryption (KP-ABSE) demonstrates the feasibility and practicality of this approach, as evidenced by our successful adaptation of several existing ABE schemes.

The importance of this formalization lies in its ability to enable more secure and flexible search functionalities in encrypted databases, which is more and more necessary for multi-user environments with diverse access control requirements. Our analysis shows that the introduction of the index, query generation, and search algorithms does not compromise the confidentiality of encrypted documents or the privacy of access policies and does not introduce high performance overhead.

Moving forward, we aim to provide the formal security proof for both CP-ABSE and KP-ABSE schemes and further refine and optimize the three added steps: $CreateIndex$, $GenerateQuery$, and $Search$. Our goal is to enhance their efficiency, making them even more suitable for practical applications. Additionally, we plan to develop two specific algorithms based on this formalization: one for Key-Policy Attribute-Based Searchable Encryption (KP-ABSE) and one for Ciphertext-Policy Attribute-Based Searchable Encryption (CP-ABSE); and test performance benchmarks on relevant data. These algorithms will leverage the strengths of their respective ABE schemes, providing tailored solutions for different use cases and further advancing the capabilities of searchable encryption.

References

[1] Dawn Xiaoding Song, D. Wagner and A. Perrig, *Practical techniques for searches on encrypted data*, in Proceeding 2000 IEEE Symposium on Security and Privacy. S&P 2000, Berkeley, CA, USA, pp. 44-55, 2000, https://dx.doi.org/10.1109/SECPRI.2000.848445

[2] D. Boneh, G. Di Crescenzo, R. Ostrovsky and G. Persiano, *Public Key Encryption with Keyword Search*, in: Cachin, C., Camenisch, J.L. (eds) Advances in Cryptology - EUROCRYPT 2004. EUROCRYPT (Lecture Notes in Computer Science), vol 3027. Springer, Berlin, Heidelberg, 2004, https://doi.org/10.1007/978-3-540-24676-3_30

[3] B. G. Pillai and N. Dayanand Lal, *Blockchain-Based Searchable Asymmetric Encryption Scheme in Cloud Environment*, in 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, pp. 1-6, 2023 https://dx.doi.org/10.1109/ICAISC58445.2023.10201090

[4] M. Wang, L. Rui, S. Xu, Z. Gao, H. Liu and S. Guo, Shaoyong, *A multi-keyword searchable encryption sensitive data trusted sharing scheme in multi-user scenario*, in Elsevier North-Holland, Inc., USA, vol. 237, no. C, 2023 https://doi.org/10.1016/j.comnet.2023.110045

[5] M. Wang, L. Rui, S. Xu, Z. Gao, H. Liu and S. Guo, Shaoyong, *A multi-keyword searchable encryption sensitive data trusted sharing scheme in multi-user scenario*, in Elsevier North-Holland, Inc., USA, vol. 237, no. C, 2023 https://doi.org/10.1016/j.comnet.2023.110045

[6] R. Zhang, R. Xue, T. Yu and L. Liu, *PVSAE: A Public Verifiable Searchable Encryption Service Framework for Outsourced Encrypted Data*, in 2016 IEEE International Conference on Web Services (ICWS), San Francisco, CA, USA, pp. 428-435, 2016, https://dx.doi.org/10.1109/ICWS.2016.62

[7] L. Meng, L. Chen, Y. Tian, M. Manulis and S. Liu, *FEASE: Fast and Expressive Asymmetric Searchable Encryption*, in Cryptology ePrint Archive, Paper 2024/054, 2024, https://eprint.iacr.org/2024/054, in press

[8] T. Yarl, B.M. Goi, R. Komiya and S.Y. Tan, *A Study of Attribute-Based Encryption for Body Sensor Networks*, in Informatics Engineering and Information Science, vol. 251, pp. 238-247, 2011, https://doi.org/10.1007/978-3-642-25327-0_21

[9] K. Yang, X. Jia, K. Ren and B. Zhang, *DAC-MACS: Effective data access control for multi-authority cloud storage systems*, in 2013 Proceedings - IEEE INFOCOM, Turin, Italy, pp. 2895-2903, 2013, https://dx.doi.org/10.1109/INFCOM.2013.6567100

[10] K. Yang, X. Jia, *Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage*, in 2014 IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 7, pp. 1735-1744, 2014, https://dx.doi.org/10.1109/TPDS.2013.253

[11] X. Yao, Z. Chen, Y. Tian, *A Lightweight Attribute-Based Encryption Scheme for the Internet of Things*, Future Generation Computer Systems, vol. 49, pp. 104-112, 2015, https://dx.doi.org/10.1016/j.future.2014.10.010

[12] J. Li, Y. Zhang, J. Ning, X. Huang, G. S. Poh, D. Wang, *Attribute Based Encryption with Privacy Protection and Accountability for CloudIoT*, in 2022 IEEE Transactions on Cloud Computing, vol. 10, no. 2, pp. 762-773, 2022, https://dx.doi.org/10.1109/TCC.2020.2975184

[13] J. Bethencourt, A. Sahai, B. Waters, *Ciphertext-Policy Attribute-Based Encryption*, in 2007 IEEE Symposium on Security and Privacy (SP), pp. 321-334, 2007, https://dx.doi.org/10.1109/SP.2007.11

[14] Y. Rouselakis and B. Waters, *Efficient Statically-Secure Large-Universe Multi-Authority Attribute-Based Encryption*, in Financial Cryptography and Data Security (Lecture Notes in Computer Science), vol. 8975, Springer, Berlin, Heidelberg, Germany, 2015, https://dx.doi.org/10.1007/978-3-662-47854-7_19

[15] Charm: A Framework for Rapidly Prototyping Cryptosystems, https://github.com/JHUISI/charm/blob/dev/LICENSE.txt

[16] John Hopkins University: Advanced Research in Cryptography laboratory, https://arc.isi.jhu.edu/

[17] Charm ABENC schemes package, https://github.com/JHUISI/charm/tree/dev/charm/schemes/abenc

[18] M. Grigutytė, *What is bcrypt and how does it work?* , in NordVPN Blog, 2023, https://nordvpn.com/blog/what-is-bcrypt/

[19] PyCharm, https://www.jetbrains.com/pycharm/

[20] Parallels, https://www.parallels.com/

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 25–32 DOI: 10.15439/2024F8304 ISSN 2300-5963 ACSIS, Vol. 40

Relative Performance of Neural Networks and Binary Logistic Regression in a Variable Selection Framework

Castro Gbêmêmali Hounmenou 0000-0002-2306-6083*, Émile C. Agbangba 0000-0001-5280-6397†,
Génevieve Amagbégnon 0009-0001-9809-7191‡ and Reine Marie Ndéla Marone 0000-0003-0787-1936 §
*Laboratoire de Biomathématiques et d'Estimations Forestières, University of Abomey-Calavi (LABEF/UAC), Benin
Centre de Recherche et de Formation en Infectiologie de Guinée, Université Gamal Abdel Nasser de Conakry, Guinea
University of Labé
Email: castrohounmenou@gmail.com
† Department of Environmental Engineering, Polytechnic School of Abomey-Calavi, Benin
Laboratoire de Biomathématiques et d'Estimations Forestières, University of Abomey-Calavi, Benin
Email: agbangbacodjoemile@gmail.com
‡Département de la Statistique et d'Economie Sectorielle, Ecole Nationale d'Economie Appliquée et de Management,
University of Abomey-Calavi (ENEAM/UAC), Benin
Email: genevieveamagbegnon@gmail.com
§ Ecole des Bibliothécaires, Archivistes et Documentalistes, Cheikh Anta Diop University (UCAD), Sénégal
Email:reinemarie.marone@ucad.edu.sn

***Abstract*—This study evaluates the predictive capabilities of a binary response variable using Multilayer Perceptron neural networks (BLMLP) and binary logistic regression (BLR) in a variable selection context. The data used was related to the identification of prenatal factors linked to premature birth in women already in labor. The stepwise selection method on BLR and the Olden selection method based on the neural network approach were used to select the most relevant variables to predict the probability of premature birth by women. Then, the two selection methods were combined with BLR and BLMLP models. Using performance criteria such as sensitivity, precision, classification accuracy, F-score, and Area Under the Curve, the selection methods were compared to identify the best model. It appears from the analysis that the best procedure for selecting variables in a binary variable prediction is the use of the Stepwise procedure followed by multilayer perceptron neural networks.**

***Index Terms*—Binary logistic regression, neural network, multilayer perceptron, selection of variables, prediction**

I. Introduction

THE DURATION of a full-term pregnancy is 41 weeks of amenorrhoea. However, premature birth is defined as a baby born alive before 37 weeks of amenorrhea. [1] There are three levels of prematurity: (i) extreme prematurity (less than 28 weeks); (ii) great prematurity (between 28 and 32 weeks) and (iii) average or late prematurity (between 32 and 37 weeks). The World Health Organisation estimates that in 2018 there are 15 million premature babies each year, which represents more than one in 10 babies. Nearly one million children die each year from complications related to prematurity [2]. Many survivors suffer lifelong disabilities, including learning, visual and hearing impairments. Apart from the health problems and the number of lives lost as a result of premature birth , the consequences of premature birth for women in labour present enormous health, psychic and psychological risks [3], [4] that need to be mastered in order to develop better prevention solutions. Therefore, it is important not only to know the most significant factors responsible for preterm birth in women, especially in labour, but also to predict from a number of the most relevant prenatal factors whether women already in labour will conceive prematurely or not. For this purpose, binary logistic regression models are most commonly used.

Binary Logistic Regression (BLR) is one of the most widely used statistical modeling techniques in practice to predict or to explain a binary response variable [5], [6]. BLR models are more flexible than other techniques like parametric discriminant analysis, multi-channel frequency analysis, among other techniques [7]. The optimal conditions for good performance of BLR are: absence or very weak presence of multicollinearity between the explanatory variables, linearity of the independent variables and logarithm of odds ratios, a sufficient number of events per independent variable and absence of outliers having a strong influence [8], [9].

In real world situations, these conditions may not always met. Current models are more complex and often non-linear [10]–[13]. Among new tools to handle the complexity of the relationship between variables and possible noises in data are Multilayer Perceptron Neural Networks (MLP). MLP methods do not require verification of the assumptions and do not impose any restrictions on input variables. MLPs belong to a very rich family of continuous functions, the main characteristic of which is to allow great modeling flexibility. In addition, they have demonstrated their effectiveness in predicting empirical data compared to traditional methods and are applied in various fields [14]–[17]. Moreover, another important point for the establishment of any model is the selection of the variables to be included in the model in order to improve its explanatory and/or predictive power [18]. The selection of variables offers several advantages, such as: (i) facilitates the understanding or visualization of data, (ii) facilitate deployment, (iii) reduces physical storage and sizing requirements, (iv) improves the ratio of number of observations and dimension of representation , (v) reduces running time, (vi) improves knowledge of the phenomenon of causality between descriptors and the variable to be predicted and (vii) improves prediction performance [18].

There are several selection approaches (Manual, Backward, Forward, Stepwise, Olden, Garson, etc.) classified in two main categories: methods dependent on a model (wrapper methods), which allow the selection of a subset of variables resulting into construction of a good prediction model and the filter methods which ensure the search for relevant variables and then possibly their ordering [5], [6], [18]. The latter the user and who has the possibility of eliminating one of two variables which are significantly linked. However, the knowledge of the user is not sufficient to fully understand the underlying causalities, to discern the true links of simple artefacts, highlight the interactions, among others. Likewise, when the number of candidate variables is high, this knowledge-based approach or manual selection is not easy in practice. In this case, it is necessary to turn to automatic approaches (wrapper methods). However, there is a panoply of selection approaches and given the characteristics presented by the available data, it is up to the user to sort the method most suited to the available data and which leads to the lowest possible error rate. A method frequently used in classical logistic regression is the stepwise technique, which is more efficient compared to Backward and Forward selection methods since it is a combination of these two methods [19].

MLP approach assesses the importance of a variable as the product of the raw input-hidden and output-hidden connections between each input and output neuron and adds the product across all neurons [20]. The variable selection approaches do not necessarily lead to the same types of explanatory variables selected or to the same number. Under these conditions, what are the best subset of explanatory variables to consider? And in which model to include them for prediction purposes? This paper aims to answer these questions by comparing the prediction of binary variables by multilayer perceptron neural networks and binary logistic regression in a variable selection framework for binary response prediction.

The rest of the document is structured as follows. Section 2 briefly describes the data source, provides the specifications of the models considered, offers a brief synthesis of variable selection approaches, outlines the statistical performance criteria used, and details the data analysis methodology. The results are presented in Section 3 and discussed in Section 4. Finally, Section 5 concludes the paper.

II. Methodology

A. Data source

The data used in this study focuses on prenatal factors (medical and personal) associated with preterm delivery in women already in preterm labor. They are recorded in an array of dimension 390×14 and can be accessed at [1]. They aim to get a better understanding and prediction of this threat to boost medical analysis. The summary of these data was generated by means of the calculation of some descriptive statistics parameters such as mean (the standard error) for the quantitative variables and the absolute frequency (the relative frequency) for the qualitative variables (Table I).

TABLE I
Description of variables, $n = 390$

Variable: Description	**Nature**	**Statistics**
Predictive variables		
GEST: Gestational age in weeks at the start of the study	Quantitative	30.30 (2.50)
DILATE: Cervical dilation in cm	Quantitative	1.24 (1.31)
EFFACE: Erasure of the collar in %	Quantitative	43.98 (34.86)
CONSIS: Consistency of the neck (1 = soft, 2 = medium, 3 = firm)	Ordinal qualitative	1: 55(14.10) 2: 127 (32.56) 3: 208 (53.33)
CONTR: Presence (= 1) or not (= 0) of contraction	Binary qualitative	1: 355 (91.03) 0: 35 (8.97)
MEMBRAN: Ruptured membranes(= 1) or not (= 0)	Binary qualitative	1: 91 (23.33) 0: 299 (76.67)
AGE: Patient's age	Quantitative	26.34 (5.31)
STRAT: Period of pregnancy	Quantitative	3.23 (0.83)
GRAVID: Gravidity (number of previous pregnancies including the current one)	Quantitative	2.30 (1.45)
PARIT: Parity (number of previous term pregnancies)	Quantitative	0.78 (1.01)
DIAB: Presence (=1) or non (=0) of a diabetes problem	Binary qualitative	1: 11 (2.56) 0: 380 (97.44)
TRANSF: Transfer (= 1) or not (= 0) transfer to a specialized care hospital	Binary qualitative	1: 188 (48.21) 0: 202 (51.79)
GEMEL: Simple pregnancy (= 1) or multiple (= 0)	Binary qualitative	1: 351 (90.00) 0: 39 (10.00)
Variable to predict		
PREMATURE: Premature delivery (= 1) or not (= 0)	Binary qualitative	1: 266 (68.21) 0: 124 (31.79)

B. Specification of models

1) Binary Logistic Regression (BLR): The relationship between the binary response variable, the premature by women in labor has two classes (premature delivery versus non-premature delivery) and various potential predictors (a collection of continuous, discrete and binary variables) is modeled by Binary logistic regression (BLR). If Y_i denotes the premature for the i^{th} woman in a sample of size $n = 390$ ($Y_i = 1$ if the woman in labor gives birth prematurely, and $Y_i = 0$ otherwise), and $\mathbf{X_i} = (X_{i1}, \cdots, X_{ia}) \in \mathbb{R}^a$ with $a \in \mathbb{N}^*$ denotes the corresponding predictors, the logistic regression model expresses the relationship between Y_i and

[1] http://eric.univ-lyon2.fr/~ricco/cours/slides/prematures.xls

$\mathbf{X_i}$ in term of the conditional probability $P(Y = 1|\mathbf{X_i} = \mathbf{x_i})$ of premature, as:

$$P(Y = 1|\mathbf{X_i} = \mathbf{x}) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x_i})}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{x_i})} \tag{1}$$

where $\boldsymbol{\beta}^\top \mathbf{x_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_a x_{ia}$ is a linear combination between the vector $\mathbf{x_i}$ of predictor variables: $\mathbf{x_i} = (x_{i0}, x_{i1}, \cdots, x_{ia})' \in \mathbb{R}^{a+1}$ and the vector of logistic regression model coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_a)^\top \in \boldsymbol{B} \subset \mathbb{R}^{a+1}$; x_0 is an additional component of unit vector and β_0 is the intercept in the model.

By applying the logistic transformation and using the equation (Eq.1), we get the linear relation between the logarithm of the odds ratio (odds = $\exp(\boldsymbol{\beta}^\top \mathbf{x_i})$) and the independent variables (Eq.2).

$$\begin{aligned}\text{logit}\Big(P(Y = 1|\mathbf{X_i} = \mathbf{x_i})\Big) &= \ln\Big(\frac{P(Y = 1|\mathbf{X_i} = \mathbf{x_i})}{1 - P(Y = 1|\mathbf{X_i} = \mathbf{x_i})}\Big) \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_a x_{ia}. \end{aligned} \tag{2}$$

Assuming that we have n independent observations: $y_1, \cdots, y_n$, and that the i^{th} observation is a realization of the random response variable Y, the probability density function of Y is given by [21]:

$$f(y_i|\boldsymbol{\beta}) = P(Y = 1|\mathbf{X_i} = \mathbf{x_i})^{y_i}(1 - P(Y = 1|\mathbf{X_i} = \mathbf{x_i}))^{1-y_i} \tag{3}$$

and the conditional likelihood function is written:

$$\mathsf{L}(\boldsymbol{\beta}|y_i) = \prod_{i=1}^{n} P(Y_i = 1|X_i = x_i)^{y_i}(1 - P(Y_i = 1|X_i = x_i))^{1-y_i} \tag{4}$$

To simplify the maximization of the equation (4), which allows to obtain the values of $\boldsymbol{\beta}$, its logarithm is used:

$$\begin{aligned}\ln \mathsf{L}(\boldsymbol{\beta}|y_i) &= \sum_{Y_i=1} \ln P(Y = 1|\mathbf{X_i} = \mathbf{x_i}) \\ &+ \sum_{Y_i=0} ln(1 - P(Y = 1|\mathbf{X_i} = \mathbf{x_i}))\end{aligned} \tag{5}$$

And replacing the expression $P(Y = 1|\mathbf{X_i} = \mathbf{x_i})$ (see equation (1)) in equation (5), we obtain:

$$\ln \mathsf{L}(\boldsymbol{\beta}|y_i) = \sum_{i=1}^{n} \Big(y_i(\mathbf{x_i}\boldsymbol{\beta}) - \ln\big(1 + \exp(\mathbf{x_i}\boldsymbol{\beta})\big)\Big) \tag{6}$$

The maximization of the relation ((6)) gives the estimation of $\boldsymbol{\beta}$ and this includes partial differentiation using iterative procedures as Newton-Raphson algorithm, Fisher scoring method, etc. [22], [23].

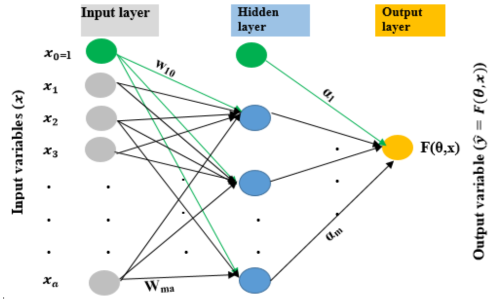


Fig. 1. An example of binary multilayer perceptron neural network model, BLMLP(a,m,1)

2) Formalism of Binary Logistic Multilayer Perceptron Neural Network (BLMLP): Binary Logistic Multilayer Perceptron Neural Networks are mathematical models inspired by human brain function and represented as directed graph (Fig.1). They are made up of neurons organized in successive layers. The first layer is called the input layer, the last output layer, and the middle layers are called the hidden layers. Neurons are interconnected with each other by synaptic weights (model parameters) and on the same layer, neurons cannot interconnect. Considering $n \in \mathbb{N}^*$, the number of women to give birth in the sample where $i\,(i = 1, \dots, n)$ represents any women in the sample, after through passing the examples $(x_i, y_i)_{1 \leq i \leq n}$ in the network, the output F (the likelihood of a woman delivering a premature baby or not) is calculated using the following equation [24]:

$$F(\theta, x) = f(\sum_{k=1}^{m} \alpha_k f(\sum_{l=1}^{a} w_{kl} x_l + w_{k0}) + \alpha_0) \tag{7}$$

where $F(.,.) : \Theta \times \mathbb{R}^{a+1} \longrightarrow [0, 1]$; $\theta = (w_{10}, \dots, w_{m0}; w_{11}, \dots, w_{1a}, \dots, w_{m1}, \dots w_{ma}; \alpha_0; \alpha_1, \dots, \alpha_m) \in \Theta \subset \mathbb{R}^{m(a+2)+1}$ et $f(.) : \mathbb{R} \longrightarrow [0, 1]$ (real value function) are respectively the output of the model, the vector of parameters of the model and the activation function of the output unit and each hidden unit $(f(z) = \dfrac{1}{1 + e^{-z}})$. $w_k = (w_{k0}, \dots, w_{ka})' \in \mathbb{R}^{a+1}$ is a vector of parameters of a hidden unit k with $k \in [\![1, m]\!]$; et $\alpha = (\alpha_0, \dots, \alpha_m)' \in \mathbb{R}^{m+1}$ a vector of parameters for the single output unit.

The parameter θ is estimated by minimizing the cross-entropy error function defined by :

$$E(\theta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \ \log(F(\theta, x_i) + (1 - y_i) \log(1 - F(\theta, x_i))] \tag{8}$$

For this purpose, different algorithms are used and based on the gradient descent procedure. The basic idea is to calculate the partial derivatives $\partial(\theta)/\partial w_k$ et $\partial E(\theta)/\partial \alpha_k$ using the chain rule. There are two steps: The first is propagation learning, which calculates the error and partial derivatives, and the second is reverse propagation learning, which calculates the update of the resulting weight. From one algorithm to another, only the second step changes. We briefly present the one used in this study which is the resilient backpropagation algorithm (Rprop) as well as a local adaptive learning program [25].

$$\theta(k + 1) = \theta(k) + \triangle\theta(k) \tag{9}$$

$$\begin{cases} \eta^+ \times \triangle(k-1) \ \ if \dfrac{\partial E(\theta)}{\partial \theta}(k-1) \times \dfrac{\partial E(\theta)}{\partial \theta} > 0 \\ \eta^- \times \triangle(k-1) \ \ if \dfrac{\partial E(\theta)}{\partial \theta}(k-1) \times \dfrac{\partial E(\theta)}{\partial \theta} < 0 \\ \triangle\theta(k-1) \ \ else \end{cases} \tag{10}$$

where $k =$ number of iterations; η^- et η^+ are reduction and increase factors, $0 < \eta^- < \eta^+$. These factors are fixed at $\eta^+ = 1,2$ et $\eta^- = 0,5$ based on theoretical considerations and empirical assessments. This reduces the number of free parameters to two, namely $\triangle_0$ and $\triangle_{max}$. The computation is slightly more expensive than the ordinary back-propagation but is an answer to the problems of convergence and over-adjustment.

C. Variable selection

Variable selection eliminates irrelevant variables from the model to improve its accuracy and also reduce the risk of overfitting [26]. For logistic regression models, it is possible to test the statistic of the significance of the coefficients associated with the variables in the model [27]. These tests can be used to build models step by step. The three most common approaches are to start with an empty model and successively add variables (forward selection), to start with the complete model and remove variables (backward selection) or by adding and removing covariates (stepwise selection). Due to the nonlinear nature of multilayer perceptron neural networks, the statistical tests for the significance of the coefficients that are used in classical logistic regression cannot be applied here. We can use the automatic relevance determination [28] or the sensitivity analysis [20], [29] to heuristically evaluate the importance of the input variables on the target variable. One method used for the selection of variables is the Olden method. This method is similar to Garson's [30] algorithm modified by [31] in that the connection weights between layers of a neural network form the basis for determining varying importance. This Olden method calculates the importance of a variable as the product of the raw input-cached and output hidden connections between each input and output neuron and adds the product across all the hidden neurons. An advantage of this approach is that the relative contributions of each connection weight are maintained in terms of amplitude and sign with respect to Garson algorithm which only takes into account the absolute amplitude. Moreover, the need to reduce the number of input variables was not linked only to the performance of neural network models. Indeed, before the work of [32], neural networks were treated as a " black box " because they provided little information to explain the influence of independent variables in prediction process. Thus, [32] have proposed and demonstrated a randomization approach to statistically assess the importance of axon connection weights and input variables contribution to the neural network. Researchers have the possibility of eliminating null connections between neurons whose weights do not significantly influence the output of the network thus facilitating the interpretation of the individual and interactive contributions of the input variables in the network. By using this randomization procedure, the mechanism of the " black box" is clarified and improves the predictive ability of neural networks.

Variable selection methods, particularly the Olden procedure and the stepwise procedure, are favored in this work due to their numerous advantages. The Olden method stands out for its interpretability, allowing for easy analysis of variable importance and their contributions to predictions, its ability to account for correlations between variables, and its flexibility in application to various types of complex and even nonlinear models. On the other hand, the stepwise procedure offers an automatic selection process that simplifies modeling, strikes a balance between complexity and performance to avoid overfitting, while producing simpler and more generalizable models, as well as a solid statistical foundation to justify variable choices. It is particularly more utilized in the health field, where understanding the impact of each variable is crucial for clinical decision-making. In comparison to other methods, filtering methods evaluate variables independently of the model, risking the neglect of interactions and potentially leading to less relevant selection, whereas Olden and stepwise analyze the effect of variables within the model, promoting better selection. Clustering selection methods, while effective in reducing the number of variables, may omit crucial information by grouping features without considering their individual importance; in contrast, Olden and stepwise assign a distinct value to each variable. Finally, wrapper methods can produce excellent results but are often computationally expensive, while Olden and stepwise prove to be more efficient and suitable for large datasets while maintaining good performance.

D. Statistical performance criterion

To evaluate and select the best performing model, goodness of fit statistics such as sensitivity, precision, F-score, classification accuracy (Accuracy) and the area under the curve (AUC) are used. The closer the values of these criteria are to 1, the better the model. They are calculated from a confusion matrix (Table II). The notations in this table are as follows: all true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) [33]. In the table above, True Positives are observations that have been rated positive and actually are. False Positives are individuals classified as positive and who are in fact negative. Likewise, False negatives are individuals classified as negative but who are actually positives and True negatives are observations that have been classified as negative and are actually negative.

TABLE II
CONFUSION MATRIX

	Predict: No (0)	Predict: Yes(1)
Actual: No (0)	True negatives (TN)	False positives (FP)
Actual: Yes (1)	False positives (FN)	True positives (TP)

Sensitivity: It measures the proportion of current positives that are correctly identified. The formula is as follows :

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

Accuracy: It is the proportion of the total number of predictions that are correct.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

Precision: This is the proportion of positives that are correctly identified.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

F-score: It is the combination of sensitivity and positive predictive value, which can be further called precision.

$$F - score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{14}$$

The AUC criterion of ROC: It expresses the probability of placing a positive individual in front of a negative individual

$$AUC = \frac{W_1 - \frac{n_1.(n_1+1)}{2}}{n_1.n_0} \tag{15}$$

where :W_1: the sum of the ranks of mis-classified individuals; n_1: the number of misclassified individuals ; n_0: the number of well-ranked individuals

Akaike Information Criterion (AIC): is a measure of the quality of a statistical model. It applies to models estimated by the maximum likelihood approach such as logistic regressions. It is defined by :

$$AIC = -2log(L) + 2a \tag{16}$$

where L: the likelihood of the model and a the number of parameters in the model. It is a criterion for penalizing the log likelihood taking into account the number of explanatory variables. The best model is the one with the lowest AIC.

E. Data analysis methods

The data analysis was done in 5 steps :

1^{st} *Step* : Data processing

Initial data (x_{ij}, y_i) (1≤i≤390 and 1≤j≤13) are normalized using the formula (Eq.17). Therefore, they are partitioned into training data (70%) and test data (30%). The training data is used to establish models and test data is used to assess the model generalization abilities.

$$new_v = \frac{v - min_z}{max_z - min_z} \tag{17}$$

2^{nd} *Step* : Establishment of models

Two different models were considered for the prediction of preterm delivery. First, the binary logistic regression (BLR) model using the regression (Eq.2) with the function “glm” from the default package “stat” of R software [34] and based on binomial distribution. Second, multilayer perceptron neural networks, MLP (see Eq. 7) were used by varying the number of hidden neurons (2, 5, 8, 11, 15 and 20). The Rprop algorithm is applied. The function ”neuralnet” from the package ”neuralnet” of R software [34] is used [35]. The best MLP architecture is obtained based on the performance criteria value close to 1.

3^{rd} *Step* : Variables selection (identification of the determinants of preterm birth)

The variable selection methods used for an effective prediction of preterm delivery in women are : the Stepwise procedure applied on the BLR model with the ”stepAIC” function from ”MASS” package of R software [34] and the AIC fit statistic is used to measure the fit of the model during the variable selection process. The best model is the one with the lowest value of AIC.

The Olden procedure applied on the MLP identified in step 2 as best. The ”olden” function of the ”NeuralNetTools” package [36] is used and the higher the value of importance of an explanatory variable, the more this variable affects the response variable and the better the results . Since the number of input variables has decreased, a new MLP architecture has been chosen again taking into account the variables selected by the Olden procedure.

4^{th} *Step* : Effective prediction of premature baby delivery with selected variables and identification of the best model.

Four types of models have been developed but with regard to the use of MLPs, the number of hidden neurons has always been varied. These models are: (i) MLP on the variables selected from the Olden procedure, (ii) MLP on the variables selected from the Stepwise procedure, (iii) BLR on the variables selected from the Olden procedure and (iv) BLR on the variables selected from the Stepwise procedure. Based on the performance criteria value near 1, the best model is identified.

5^{th} *Step* : Analysis of the variables of preterm delivery according to the best approach.

III. 3. Results

A. Determination of the best architecture of multilayer perceptron neural networks and establishment of classical binary logistic regression

The performance of BLMLPs models varies depending on the number of neurons in the hidden layer (Table III). The BLMLP model $(13, 20, 1)$ provided the best architecture with 13 input variables, 20 hidden neurons and an output variable (value closed to 1 for all performance criteria: $TBC = 0.99$, $Sensitivity = 0.99$, $Precision = 1$, $F - score = 0.99$ and $AUC = 0.99$).

Regarding the binary logistic regression model, the residual deviance (246.11) is deviated from the degrees of freedom (274) and their ratio is equal to 0.90, $AIC = 274.11$, $TBC = 0.76$, $Sensitivity = 0.88$, $Precision = 0.80$, $F - score = 0.84$, $AUC = 0.84$.

B. Identification of selected variables according to Olden and Stepwise procedures

Fig. 2 provides information on the importance of the explanatory variables compared to the variable explained by the

TABLE III
IDENTIFYING THE BEST NEURAL NETWORK

Architecture	**Sensitivity**	**Precision**	**F-score**	**Accuracy**	**AUC**
BLMLP(13,2,1)	0.75	0.85	0.80	0.72	0.81
BLMLP(13,5,1)	0.91	0.90	0.90	0.87	0.88
BLMLP(13,8,1)	0.95	1.00	0.97	0.96	0.95
BLMLP(13,11,1)	0.96	0.99	0.97	0.96	0.95
BLMLP(13,15,1)	0.96	0.99	0.97	0.96	0.95
BLMLP(13,20,1)	0.99	1.00	0.99	0.99	0.99

Olden procedure. It reveals that a subset of 5 explanatory variables are retained among the initial 13 . These are: GEMEL, TRANSF, GRAVID, PARIT and DILATE (importance value greater than 0). With the Stepwise procedure, 8 explanatory variables are selected (AIC = 266.91, lower than that of the full model, AIC = 274.11): CLEAR, MEMBRAN, STRAT, DIAB, in addition to the 4 variables obtained by the Olden procedure except by GRAVID.

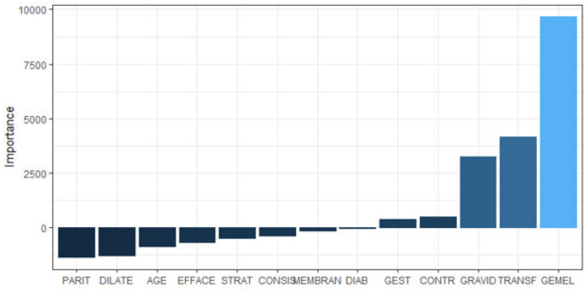


Fig. 2. "Importance diagram of explanatory variables derived from the Olden procedure

C. Comparative analysis of modeling approaches for an efficient prediction of premature

The five variables selected with the Olden procedure and the 8 resulting from the Stepwise procedure were used as input for the BLMLPs with variation in the number of hidden neurons (Tables IV and V). The analysis of the performance criteria reveals that the best architectures of the binary logistic multilayer perceptron neural network are respectively $BLMLP(5,2,1)$ and $BLMLP(8,20,1)$ for a good prediction of the PREMATURE.

So for comparisons, the pure neural network approach ($BLMLP_{\mathrm{Olden}}$, $BLMLP(5,2,1)$) and the approach BLMLP and Stepwise ($BLMLP_{\mathrm{Olden}}$, $BLMLP(8,20,1)$) are retained. Added to this are the binary logistic regression models with the 8 variables retained by stepwise procedure ($BLR_{stepwise}$) and the one with the 5 variables retained by Olden procedure (BLR_{Olden}).

The comparison of predictive performances for these four models (Table VI): Sensitivity, precision, F-score, rate of good classification and AUC showed that the model $BLMLP_{stepwise}(8,20,1)$ is the best model (Table VI). Therefore, stepwise selection gives the neural network better performance in terms of prediction. Fig. 3 presents this network. We can therefore retain that stepwise procedure is better compared to Olden procedure. Thus, the relevant variables to better predict the premature delivery of a baby are :

- DIAB: presence or absence of a diabetes problem
- GEMEL: single or multiple pregnancy
- STRAT: period of pregnancy
- TRANSF: transfer or not to a hospital for specialized care
- DILATE: cervical dilation
- PARIT: parity (number of previous term pregnancies)
- EFFACE: the erasure of the collar
- MEMBRAN: rupture of membranes

TABLE IV
IDENTIFICATION OF THE BEST NETWORK WITH THE SELECTED INPUT VARIABLES WITH OLDEN PROCEDURE

Architecture	**Sensitivity**	**Precision**	**F-score**	**Accuracy**	**AUC**
BLMLP(5,2,1)	1	0.82	0.90	0.71	0.73
BLMLP(5,5,1)	1	0.81	0.90	0.68	0.72
BLMLP(5,8,1)	1	0.79	0.88	0.70	0.72
BLMLP(5,11,1)	1	0.80	0.89	0.64	0.68
BLMLP(5,15,1)	1	0.80	0.89	0.70	0.69
BLMLP(5,20,1)	0	0.78	0.00	0.54	0.58

TABLE V
IDENTIFICATION OF THE BEST NETWORK WITH THE VARIABLES SELECTED WITH THE STEPWISE PROCEDURE

Architecture	**Sensitivity**	**Precision**	**F-score**	**Accuracy**	**AUC**
BLMLP(8,2,1)	1	0.86	0.92	0.75	0.82
BLMLP(8,5,1)	1	0.85	0.92	0.80	0.88
BLMLP(8,8,1)	1	0.85	0.93	0.82	0.88
BLMLP(8,11,1)	1	0.88	0.88	0.83	0.89
BLMLP(8,15,1)	1	0.90	0.95	0.86	0.94
BLMLP(8,20,1)	1	0.91	0.95	0.88	0.97

TABLE VI
COMPARISON OF ANALYTICAL APPROACHES IN THE CONTEXT OF VARIABLE SELECTION

Models	**Sensitivity**	**Precision**	**F-score**	**Accuracy**	**AUC**
BLR$_{stepwise}$	1	0.81	0.90	0.77	0.82
BLR$_{Olden}$	1	0.77	0.87	0.73	0.73
BLMLP$_{stepwise}$	1	0.91	0.95	0.88	0.97
BLMLP$_{Olden}$	1	0.82	0.90	0.71	0.73

IV. DISCUSSION

The predictive performance of empirical data based on binary logistic multilayer perceptron neural network (BLMLP) model is better than that of classical logistic regression (BLR), taking into account all the starting independent variables (full model). Likewise, for the same subset of variables resulting from the same variable selection procedure and serving as input for the BLMLP and BLR models, BLMLPs give the best prediction performance. These results could be justified by the fact that BLMLPs are semi-parametric classifiers and are more flexible than parametric models. They

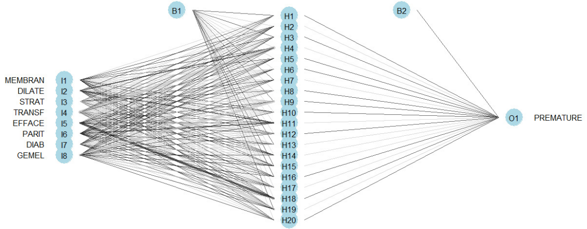


Fig. 3. Best model for PREMATURE prediction

use learning by example which makes them more powerful in pattern recognition and have more ability to mimic complicated patterns than classical logistic regression [37], [38]. In addition, they do not require a hypothesis [39] and are able to find models despite the presence of noisy data or missing data and even in the presence of multi-collinearities between the descriptors [8], [17], [40]. Furthermore, the use of BLR models requires satisfaction of many assumptions which may not be true in some real cases. This is probably the case with the PREMATURE data on which the study is focused. Failure to respect these assumptions can affect the predictive performance of BLR models and consequently lead to errors in predictions [8], [9], [18] Likewise, several studies have shown that multilayer perceptron neural networks have better prediction skills compared to classical binary logistic regression. [41]–[45]. But since the sample size of our data is not large, this result is contrary to those obtained by [44], [46] who worked on a large sample size where BLMLPs and logistic regression classic have almost similar performance although PCMs are powerful in concept. However, logistic regression requires large sample sizes to make maximum likelihood estimates powerful. [9]. The independent variables selected vary according to the selection procedure and as well as their numbers. This observation is certainly due to the approaches used which are related to the estimation criterion (AIC for the Stepwise procedure and Importance for Olden procedure). With the Olden procedure, we can know the order of importance and the direction of influence of each identified descriptor, which is not the case with Stepwise where we can only know the group of significant descriptors. Considering the selected variables by the Stepwise procedure as input variables for the BLMLP model ($BLMLP_{Stepwise}$) has good predictive power than the models $BLMLP_{Olden}$, $BLR_{Stepwise}$ and BLR_{Olden}. This could be justified by the fact that Stepwise procedure got rid of all the explanatory variables not relevant than Olden procedure. These irrelevant variables could make the estimates numerically unstable and negatively affect the predictive capacity of the BLMLP and BLR models [47]. This approach seems to give a result contrary to the principle of Occam's Razor, which in favor of selecting, for the same number of observations, a model with few variables with a better chance of being more robust in generalization. However, the number of descriptors identified with the Stepwise procedure is higher than that obtained with Olden by considering the same number of observations. This contraction could be explained by the complicity of the data or of the possible interactions existing between them, that the MLPs models have the capacity to manage [44], [45]. Although the selected variable prediction approach $BLMLP_{Stepwise}$ gives better predictive performance, it would be advantageous for a study to compare Olden procedure to stepwise one depending on the complexity of the relationship between variables. Another advantage may be to vary the sample size and the dimension of the variables to see how the four models will behave as the sample size increases. Another important aspect of networks is the choice of hyper-parameters (activation functions in hidden layers, number of layers and hidden neurons, learning rate, learning algorithm, etc.). The latter influence the performance of neural networks and would be useful to explore them for the selection of variables with Olden procedure. Moreover, the comparisons were based on empirical data and it would be important to repeat them on several databases through a simulation in order to generalize the conclusions.

V. Conclusion

In this study, two models of prediction of a binary variable (binary logistic regression and multilayer perceptron neural networks) were combined with two variable selection procedures (stepwise and Olden) in order to propose a new prediction approach.Starting from the example of predicting the premature or non-premature delivery of a baby, binary logistic multilayer perceptron neural network (BLMLP) models best predict these data compared to classical logistic regression (BLR) models with all the starting independent variables (full model). Also, for the same group of variables resulting from the same variable selection procedure and serving as input for the PCM and BLR models, the BLMLPs give the best prediction performance. Moreover, the use of the variables selected by the stepwise selection procedure as input variables to neural networks has a good predictive power that the models $BLRMLP_{Olden}$, $BLR_{Stepwise}$ and BLR_{Olden}.

References

[1] D. Wolke, "Preterm birth: high vulnerability and no resiliency?" *Journal of Child Psychology and Psychiatry*, vol. 59, no. 11, pp. 1201–1204, 2018.

[2] L. Liu, S. Oza, D. Hogan, Y. Chu, J. Perin, J. Zhu, J. E. Lawn, S. Cousens, C. Mathers, and R. E. Black, "Global, regional, and national causes of under-5 mortality in 2000–15: An updated systematic analysis with implications for the sustainable development goals," *The Lancet*, vol. 388, no. 10063, pp. 3027–3035, 2016.

[3] R. Alijahan, S. Hazrati, M. Mirzarahimi, F. Pourfarzi, and P. A. Hadi, "Prevalence and risk factors associated with preterm birth in ardabil, iran," *Iranian journal of reproductive medicine*, vol. 12, no. 1, p. 47, 2014.

[4] A. T. Deressa, A. Cherie, T. M. Belihu, and G. G. Tasisa, "Factors associated with spontaneous preterm birth in addis ababa public hospitals, ethiopia: cross sectional study," *BMC pregnancy and childbirth*, vol. 18, no. 1, pp. 1–5, 2018.

[5] C. J. Peng and T. H. So, *Logistic regression analysis and reporting: A Primer*, 2002.

[6] P. K. Josepha and A. Ame, "Effect of testing logistic regression assumptions on the improvement of the propensity scores," *International Journal of Statistics and Applications*, vol. 8, no. 1, pp. 9–17, 2018. [Online]. Available: http://doi:10.5923/j.statistics.20180801.02

[7] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*, 2007.

[8] J. C. Stoltzfus, "Logistic regression: A brief primer," *ACADEMIC EMERGENCY MEDICINE*, vol. 18, pp. 1099–1104, 2011. [Online]. Available: http://doi:10.1111/j.1553-2712.2011.01185.x

[9] H. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J Korean Acad Nurs*, vol. 43, no. 2, pp. 154–164, 2013. [Online]. Available: https://doi.org/10.4040/jkan.2013.43.2.154

[10] D. M. Bates and D. G. Watts, *Nonlinear regression analysis and its applications*. Wiley New York, 1988, vol. 2.

[11] J. K. Lindsey, *Nonlinear models in medical statistics*. Oxford University Press on Demand, 2001.

[12] S. Wang, L. Zheng, J. Dai *et al.*, "Empirical likelihood diagnosis of modal linear regression models," *Journal of Applied Mathematics and Physics*, vol. 2, no. 10, p. 948, 2014.

[13] J. Hagenauer, H. Omrani, and M. Helbich, "Assessing the performance of 38 machine learning models: The case of land consumption rates in bavaria, germany," *International Journal of Geographical Information Science*, vol. 33, no. 7, pp. 1399–1419, 2019. [Online]. Available: https://doi.org/10.1080/13658816.2019.1579333

[14] M. Cottrell, M. Olteanu, F. Rossi, J. Rynkiewicz, and N. Villa-Vialaneix, "Neural networks for complex data," *KI-Künstliche Intelligenz*, vol. 26, no. 4, pp. 373–380, 2012.

[15] G. Daniel, *Principles of artificial neural networks*. World Scientific, 2013, vol. 7.

[16] O. Asogwa and A. Oladugba, "On the comparison of artificial neural network (ann) and multinomial logistic regression (mlr)," *West African Journal of Industrial and Academic Research*, vol. 13, no. 1, pp. 3–9, 2015.

[17] A. Mollalol, M. K. Rivera, and B. Vahedi, "Artificial network modeling of novel coronavirus (covid 19) incidence rates across the continental united states," *Int. J. Environ. Res. Public Health*, vol. 17, no. 4204, pp. 1–13, 2020. [Online]. Available: http://doi:10.3390/ijerph17124204

[18] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014. [Online]. Available: http://dx.doi.org/10.11613/BM.2014.003

[19] P. Du Jardin, "Prevision de la defaillance et reseaux de neurones : L'apport des methodes numeriques de selection de variables," in *These de Doctorat, Universite Nice Sophia Antipolis, France*. HAL, 2007.

[20] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological modelling*, vol. 178, no. 3-4, pp. 389–397, 2004.

[21] J. D. Jobson, *Applied multivariate data analysis: volume II: Categorical and Multivariate Methods*. Springer Science & Business Media, 2012.

[22] S. A. Czepiel, "Maximum likelihood estimation of logistic regression models: theory and implementation," *Available at czep. net/stat/mlelr. pdf*, pp. 1 825 252 548–1 564 645 290, 2002.

[23] A. Diop, A. Diop, and J.-F. Dupuy, "Maximum likelihood estimation in the logistic regression model with a cure fraction," *Electronic journal of statistics*, vol. 5, pp. 460–483, 2011.

[24] C. G. Hounmenou, R. Tohoun, K. E. Gneyou, and R. GLèLè Kakaï, "Empirical determination of optimal configuration for characteristics of a multilayer perceptron neural network in nonlinear regression," *Afrika Statistika*, vol. 15, no. 3, pp. 2413–2429, 2020.

[25] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *IEEE international conference on neural networks*. IEEE, 1993, pp. 586–591.

[26] C.-K. Chen and H. John Jr, "Using ordinal regression model to analyze student satisfaction questionnaires. ir applications." *Association for Institutional Research (NJ1)*, vol. 1, 2004.

[27] D. Hosmer and S. Lemeshow, "Applied logistic regression 2nd edition wiley," *New York*, 2000.

[28] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[29] J. M. Zurada, A. Malinowski, and I. Cloete, "Sensitivity analysis for minimization of input data dimension for feedforward neural network," in *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94*, vol. 6. IEEE, 1994, pp. 447–450.

[30] D. G. Garson, "Interpreting neural network connection weights," *Artificial Intelligence Expert*, vol. 9, no. 3, pp. 46–51, 1991.

[31] A. T. Goh, "Back-propagation neural networks for modeling complex systems," *Artificial Intelligence in Engineering*, vol. 9, no. 3, pp. 143–151, 1995.

[32] J. Olden and D. Jackson, "Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks," *Ecological Modelling*, vol. 154, pp. 135–150, 2002. [Online]. Available: https://doi.org/10.1016/S0304-3800(02)00064-9

[33] J. Zarifis, V. Grammatikou, M. Kallistratos, A. Katsivas, and I. of the Prospective Noninterventional Observational Study of the Antianginal Efficacy of Ivabradine During a 4-Month Treatment of a Greek Population With Coronary Artery Disease, "Treatment of stable angina pectoris with ivabradine in everyday practice: a pan-hellenic, prospective, non-interventional study," *Clinical cardiology*, vol. 38, no. 12, pp. 725–732, 2015.

[34] R. R Core Team, "A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. version 3.3.6," *https://www.R-project.org/*, 2019.

[35] S. Fritsch, F. Guenther, and M. F. Guenther, "Package ?neuralnet?" *Training of Neural Networks. Recuperado de https://cran. r-project. org/web/packages/neuralnet/neuralnet. pdf*, 2019.

[36] M. W. Beck, "Neuralnettools: Visualization and analysis tools for neural networks," *Journal of statistical software*, vol. 85, no. 11, p. 1, 2018.

[37] J. Song, S. Venkatesh, E. Conant, P. Arger, and C. Sehgal, "Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masse," *Acad Radiol*, vol. 12, pp. 487–495, 2005. [Online]. Available: https://doi.org/10.1016/j.acra.2004.12.016

[38] S. R. D. N. C. Barwar, "A comparative study of multilayer perceptron, radial basis function networks and logistic regression for healthcare data classification," 2016.

[39] P. M. West, P. L. Brockett, and L. L. Golden, "A comparative analysis of neural networks and statistical methods for predicting consumer choice," *Marketing Science*, vol. 16, no. 4, pp. 370–391, 1997.

[40] B. Eftekhar, K. Mohammad, H. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Med Inform Decis Mak*, vol. 5, no. 3, p. 20, 2005. [Online]. Available: https://doi.org/10.1186/1472-6947-5-3

[41] P. J. Adeodato, G. C. Vasconcelos, A. L. Arnaud, R. A. Santos, R. C. Cunha, and D. S. Monteiro, "Neural networks versus logistic regression: A comparative study on a large data set." in *ICPR (3)*, 2004, pp. 355–358.

[42] V. Bourdès, S. Bonnevay, P. Lisboa, R. Defrance, D. Pérol, S. Chabaud, T. Bachelot, T. Gargi, and S. Négrier, "Comparison of artificial neural network with logistic regression as classification models for variable selection for prediction of breast cancer patient outcomes," *Advances in Artificial Neural Systems*, vol. 2010, 2010.

[43] C.-p. LI, X.-y. Zhi, M. Jun, C. Zhuang, Z.-l. Zhu, C. Zhang, and L.-p. HU, "Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus," *Chinese medical journal*, vol. 125, no. 5, pp. 851–857, 2012.

[44] M. Parsaeian, K. Mohammad, M. Mahmoudi, and H. Zeraati, "Comparison of logistic regression and artificial neural network in low back pain prediction: second national health survey," *Iranian journal of public health*, vol. 41, no. 6, p. 86, 2012.

[45] M. García, C. Valverde, M. I. López, J. Poza, and R. Hornero, "Comparison of logistic regression and neural network classifiers in the detection of hard exudates in retinal images," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 5891–5894.

[46] A. Kazemnejad, Z. Batvandi, and J. Faradmal, "Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes," *EMHJ-Eastern Mediterranean Health Journal*, vol. 16, no. 6, pp. 615–620, 2010.

[47] S. Menard, *Applied logistic regression analysis (Second Edition)*. Sage Publications, 2002, vol. 106.

DOI: 10.15439/2024F2991
ISSN 2300-5963 ACSIS, Vol. 40

Towards the actual deployment of robust, adaptable, and maintainable AI models for sustainable agriculture

Giacomo Ignesti
0000-0003-2389-3086
Institute of Information Science
and Technologies (ISTI)
National Research Council of Italy (CNR)
Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy
University of Pisa
Second Floor, Largo Bruno Pontecorvo, 3, 56127 Pisa, Italy
Email: giacomo.ignesti@isti.cnr.it

Davide Moroni, Massimo Martinelli
0000-0002-5175-5126
0000-0001-7419-5099
Institute of Information Science
and Technologies (ISTI)
National Research Council of Italy (CNR)
Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy
Email: {davide.moroni, massimo.martinelli}@isti.cnr.it

***Abstract*—In the past two decades, computer vision and artificial intelligence (AI) have made significant strides in delivering practical solutions to aid farmers directly in the fields, thereby contributing to the integration of advanced technology in precision agriculture. However, extending these methods to diverse crops and broader applications, including low-resource situations, raises several concerns. Indeed, the adaptability of AI methods to new cases and domains is not always straightforward. Moreover, the dynamic global panorama requires a continuous adaptation and refinement of artificial intelligence models. In this position paper, we examine the current opportunities and challenges, and propose a new approach to address these issues, currently in the implementation phase at CNR-ISTI.**

***Index Terms*—Sustainable Agriculture; Artificial Intelligence; Deep Learning; Crowd-sensing; Citizen science**

I. Introduction

IN RECENT years, the emergence of deep learning, combined with the increasingly widespread use of visual monitoring technologies for crops, has significantly contributedto the advancement of precision agriculture [1]. Uncrewed Aerial Vehicles (UAVs) equipped with colour or multispectral/hyperspectral cameras, as well as other robotic platforms designed for close-range operations with crops, have paved the way for the introduction of AI-assisted, data-driven approaches in agriculture [2]. This has permitted the implementation of precise monitoring, treatment, and harvesting techniques. However, these advancements have primarily impacted a narrow range of cultivated crops, particularly specialized ones yielding high revenues, such as high-end wine production [3].

It is clear that Artificial Intelligence (AI) and Machine Learning (ML), including Deep Learning (DL) methods, are versatile methodologies capable of being applied to disparate fields, including agriculture, where their potential impact has yet to be fully realized. However, the transfer from specific domains to new ones is not always feasible or cost-effective due to the associated efforts required for designing and developing new models.

Numerous research and academic initiatives focus on a wide range of crops, encompassing intensive cultivation practices that have a significant impact on the global food supply [4]. In these contexts, DL models have demonstrated unparalleled performance on standardized datasets [5].

Concerning such topics, the state of research on AI applications in agriculture is wide. There is still no standardized approach, but the literature encompasses a lot of strategies that are all focused on improving crop quality and production. While modern DL models excel in image analysis for product quality enhancement, other critical agricultural domains, like water control [6], soil management, and production chain optimization, primarily rely on tabular data or emerging multimodal approaches. Real-time object detection is a prominent AI application in agriculture, though classification algorithms often demonstrate superior performance [7] in specific contexts.

Recent works, as [8], try to employ knowledge-distillation techniques to improve weed mapping, adapting complex transformer architecture to the agricultural domain. At the same time, other studies [9] analyse various detection algorithms and design possible edge computing solutions for their real-time applications in precision agriculture. Image acquisition modality plays a pivotal role in plant analysis studies [10], as shown by the advancements in multi-modal imaging techniques that enhance the accuracy of trait estimation and facilitate the analysis of plant morphology and development. For instance, integrating visible light, fluorescence, and near-infrared imaging allows for a comprehensive assessment of plant structures, improving the segmentation and quantification of traits critical for phenotyping. These diverse imaging modalities not only provide complementary information. But also address challenges related to variable illumination and plant colouration, ultimately leading to more robust phenotypic data extraction and analysis. Object detection and segmentation algorithms are usually more complex than their classification counterpart; therefore, translating these models and approaches into practical use for corporate farmers of all scales presents challenges, as real-world variability differs from the conditions in static benchmark datasets. To date, while there is a right to benchmarking agricultural datasets, no foundational models have been trained in this domain, making only possible transfer learning strategies and training from scratch solutions. Non-technological factors, including user acceptability, also hinder the widespread adoption of the latest research findings [7].

In this context, there is a growing need for developing new methodologies to overcome the current limitations of AI-assisted technologies. Specifically, these necessitate broadening their application to new crops and different scales of cultivation to support niche, small-scale, local, and organic productions while preserving biodiversity and environments through sustainable resource management. These demands come from various stakeholders, including farmers and policymakers (such as the European Community [11]). At the same time, they also originate from the Sustainable Development Goals set by the United Nations [12], particularly Goal 2 "Zero Hunger". This goal includes targets such as doubling agricultural productivity (Target 2.3), ensuring sustainable food production systems, implementing resilient farming practices, and improving land and soil quality (Target 2.4), as well as maintaining genetic diversity through well-managed seed and plant banks (Target 2.5).

This *position paper* intends to present prospective ideas that might contribute to achieving the Sustainable Development Goals and fulfilling the requirements for the widespread adoption and implementation of practical artificial intelligence. While AI has potential applications across various domains, we focus specifically on using image-based intelligent systems to support farmers in their day-to-day operations. These systems can act as effective assistants, enabling informed decision-making and promoting the best practices for increased yet sustainable production.

The paper is organized as follows. In Section II, we critically review previous experiences, including ours, and highlight their limitations. In Section III, we enumerate a set of challenges and research questions that should be addressed to reach the scope described in this introduction. In Section IV, we analyze the current opportunities provided by technological advances and then explain the proposed approach rationale. Section V concludes the paper with remarks for further analysis and prospective implementation.

II. Previous Experiences

In light of advancements in image processing, computer vision, and machine learning, considerable research efforts have been directed toward developing intelligent systems to support agriculture. These efforts include the creation of algorithms for detecting, classifying, and quantifying crops and various potential threats such as weeds, diseases, insects, and other stressors that could impact successful harvesting. The focus has been on analyzing remote sensing images captured by UAVs and close-range photography obtained through handheld devices or robot platforms.

The curation of benchmark datasets, particularly those released as open data, has played a pivotal role in enhancing the reproducibility and extensibility of research across different domains. Surveys on existing datasets, as documented in the work by Lu et al. [13], have become readily available. For instance, the PlantVillage dataset [14] has emerged as a de facto benchmark for leaf disease classification even though images, while numerous, may not fully represent the entirety of natural variability. Consequently, the performance of deep learning models on such datasets has been exceptional, with approaches achieving maximal accuracy levels [15].

Significant endeavours have been put forth within the AGROSAT+ project, sponsored by Barilla, to address detecting and classifying weeds. Under this initiative, collaborative efforts between CNR-ISTI and CNR-IBE have led to curating a dataset specific to cereal crop weeds [16]. This dataset might be valuable for weed detection and classification problems through close-range imaging or high-resolution UAV surveys. Additionally, its suitability for machine learning methods has been demonstrated in [17], where again the top performance was obtained. While intriguing and of great importance for advancing research, the current approaches have limitations regarding practical applications. The models' ability to generalize when processing uncontrolled, real-world images is unsatisfactory, with a significant performance degradation of over 20%. This lack of reliability and inconsistent performance may be unacceptable to users in real-world deployments, leading to distrust in artificial intelligence and overall dissatisfaction, ultimately resulting in the technology's failure to be adopted.

In the context of the AGROSAT+ project, an additional initiative was undertaken to address these challenges, leading to the development of an app called "GranoScan". This app is designed to serve as an expert system that can be used directly in the field to identify plant diseases and stress, as well as detect weeds, insects, and other potential threats simply by using pictures captured through the smartphone camera. The app's backend is driven by deep learning models that handle various visual recognition tasks [18]. One notable aspect of the app is its approach, which is somewhat independent of the specific computational models employed. In more detail, following an intensive period of initial data collection to train the machine learning models, GranoScan has now entered the production stage. Since then, a continuous stream of images from diverse users has been processed, with user consent, and stored to augment the dataset. This data has provided a wealth of information that can be leveraged to enhance and refine the models developed over the years using semi-assisted and semi-supervised methods. The experience is still ongoing.

III. Challenges

Based on the previous experiences reported in Section II, a critical gap in the current AI technology for sustainable agriculture is the absence of a well-established methodology for the rapid deployment of models, namely of trained deep learning architecture for solving visual tasks related to agronomical problems. These AI models must satisfy various requirements, including robustness, adaptability, and maintainability, while being versatile enough to address various crops. Notably, the methodology should also ensure that the models can be easily transferred across different domains while maintaining their effectiveness and accuracy. For example, the models should be capable of adapting from one crop variety to another, regardless of similarities or differences in cultivation practices based on geographical location, climate, and other environmental conditions such as soil quality, water availability, and farming methods (e.g. organic, with biological or natural pest control, traditional). Developing such a methodology involves confronting several key challenges outlined below.

One of the primary challenges in deploying AI models in agriculture is the *limited availability of comprehensive and high-quality datasets*. Indeed, as shown in the survey [13], agricultural datasets, particularly those related to specific crops or regions, are often sparse, fragmented, or inconsistent (see, for instance, the dataset proposed for the challenge [19]). As we have seen, thanks to data augmentation strategies and the definition of ad hoc architectures, such a scarcity has not prevented the realization of performant AI models on static benchmark datasets. However, the generalization capabilities observed in practice have been, in our experience, somewhat disappointing.

Additionally, the agricultural environment is highly dynamic and is influenced by seasonal variations, pest outbreaks, and other temporal factors. To ensure that AI models can effectively generalize, it is crucial to train them using data collected over multiple growing seasons in order to capture these variations accurately. *Longitudinal studies* that span several agricultural cycles can provide valuable insights into long-term trends and enhance the model's ability to generalize across different conditions and time periods. Such longitudinal assessment is feasible when analyzing routine remote sensing images captured by satellite-borne sensors. However, when considering the smaller scale of details (e.g., airborne sensors and close-range images), there are currently no relevant and accessible datasets that span multiple harvest seasons.

Climate change introduces significant unpredictability into agricultural systems, affecting crop yields, pest prevalence, and overall farm productivity. For this reason, AI models need to be capable of not only interpolating within the known data but also *extrapolating* to predict the impact of unprecedented climate scenarios. This is a feature that should be taken into account when selecting the deep learning architecture or other machine learning paradigm to be used in a classification or regression task. Indeed, some methods are only suited to analyze data within the convex hull of the training set, producing in output something within the convex hull of the labels in the training data. Although most of the classification and object detection tasks are not apparently conditioned by these issues, in general, reasoning about crop status, these issues should be taken into account. In particular, this might require integrating climate models with agricultural data to create AI systems that can adapt to changing climatic conditions and provide reliable recommendations for farmers.

The ultimate objective of utilizing AI-based systems in agriculture is to convert predictive insights into *actionable knowledge* that farmers can easily put into practice. This involves not only creating user-friendly interfaces and providing effective training for farmers but also ensuring that the AI recommendations are reliable, practical, and economically feasible. In addition, there is a need for processes that facilitate continuous feedback from the field to refine and update the models, ensuring that their relevance and accuracy in real-world applications remain stable without being affected by potential non-stationary conditions.

IV. Proposed Approach

Having discussed the challenges towards the implementation and actual deployment of robust, adaptable, and manageable AI models for tackling agronomic tasks, it is important to note that several opportunities are linked to technological advances that can ease the identification of possible solutions.

From one side, indeed, there has been a flourishing of research towards identifying highly efficient and robust AI models with improved insensitivity to data variability [20].

Secondly, methods have also been analyzed from the point of view of carbon footprint, [21] taking into account not only the training and inference costs but also the overhead linked, for instance, to data transfer. This is an aspect in deciding where to collocate computationally intensive tasks over the computational continuum, determining whether to process directly near the node where the data has been captured (i.e. directly on the smartphone capturing the image or on a robotic platform) with no transfer overhead or, conversely, on the cloud (with variable transfer costs). In such a context, progress in hardware also allows for more freedom in such design choices, given the general availability of computational resources, including GPU resources, along the computational continuum.

Finally, a third opportunity arises from the successful implementation of crowd-sensing that can be attributed to two key aspects: - the first aspect is technical, in which modern accessible devices, such as smartphones, now offer enhanced sensing capabilities, including LiDAR technology, multiple camera lenses, and advanced geolocation features; - the second aspect relates to the growing awareness and willingness of individuals to participate in citizen science initiatives.

In this section, we propose the envisaged rationale and then discuss in detail the three main points it leverages.

A. Rationale

The rationale of the approaches is based on the use of three main levers that are considered to be able to effectively contribute to fast and efficient deployments, respecting the requirements discussed in the previous section. The first aspect is based on the provision, not only of statistic classification or number produced by ML/DL models, but also on integrating these methods with Decision Support Services (Section IV-B). This is envisaged to respond to the need to translate insights into actionable knowledge. Indeed, not only the output of the image processing will be produced, but it is necessary to accompany this output with an explanation (in an explainability effort) and suggestions on how this output may be used in practice to optimize treatment, for instance, by devising an adaptive treatment plan. Secondly, a better tradeoff between performance and generalization capabilities should be sought (Section IV-C). This attains research efforts in ML/DL where new methods that have already proved promising, based on ensembling, can achieve improved generalization capabilities and allow for a faster domain transfer. A third ingredient is represented by a more strategic approach to filtering crowd-sensed information, considering uncertainty in their evaluation since they originate from non-authoritative sources (Section IV-D). In this case, new methodologies can be enlisted to determine data quality and define the confidence level the new data has to enter into the decisional processes.

In the current envisaged activities such rationale is going to be validated (see Figure 1) in a variety of cases addressing i) plant position detection, ii) plant count, iii) control of the growing phase (e.g. pre/post-germination, developed, budding, pre-flowering, pre-fruiting, ripening depending on the cultivation) and iv) anomaly detection (abnormal growth compared to market standards, sufficient/insufficient gems,. . .) and v) plant threats (weeds, pests, and diseases). In addition, vi) time (of budding, flowering, fruiting, ripening,. . .) and vii) and volume predictions (number of plants/flowers/fruits/biomass) as well as viii) quality of the final product will be considered.

B. Integration with DSS

A DSS must consider several factors depending on the plant species, including sprout number, flowering time, loss of first flowering, and other variables.

The proposed model envisages the DSS's intervention point as at least twofold. Indeed, the DSS intervenes before and after the AI models, (a) first to decide which ones to run based on historical data, context conditions, seasons, situations, and others, and (b) then to provide suggestions based on the results.

A hybrid DSS integrates different technologies and information types in order to provide greater flexibility, scalability and efficiency in helping make the right decision, the correct application, and the proper treatment in the right place and at the right time: knowledge-based modules allow a semantic representation of data to extract and infer helpful information and can include Data Mining and predictive analytics to identify hidden patterns and relationships between data, providing high quality and a clear explanation of decisions; model-based modules allow optimizing the internal decision processes by analyzing specific issues, such as the irrigation scheduling or the crop prediction processing data, when the target audience/stakeholder is not interested in understanding the decision-making process but only in the results produced.

DSS can also be utilized to communicate and present information as needed. For example, AI tools that predict future outcomes based on historical data and trends (e.g. forecasting flowering or fruiting times or spreading a disease) can be activated proactively in response to specific events. The resulting output can then be promptly presented to the user, allowing for optimal treatment and harvesting planning.

C. Model adaptation, generalization capabilities and continuous learning

A key factor is the ability of AI models to adapt to new data, to generalize their knowledge, to apply them to new contexts, to ensure that models are able to function properly in different situations and to improve their accuracy over time. At the same time, we need a model capable of learning fast without relying on an extensive corpus of knowledge, represented in this specific case by a dataset annotated with ground truth. In DL, the capabilities of transfer learning are well known: deep models trained on a dataset belonging to a certain domain, often general purpose such as in the case of ImageNet, are then capable of adapting more quickly and with better performance to new domains with respect to the same architectures initialized in a random way. In addition, zero-shot and few-shot learning have been considered in several contexts, achieving classification with minimal training data [22].

In our view, we aim to address these elements by exploiting *adaptive ensembling* and *continuous learning*.

More specifically, in adaptive ensembling [5], a few weak models are trained in parallel, resulting in a set of specialized modules. Such weak models are based on DL models and, specifically, in architectures belonging to the EfficientNet family [23]. As such, they comprise a first set of layers, performing feature extraction and a final layer-producing classification. In our approach, such weak models are combined together to produce a strong classifier at the deep feature level. Namely, the original classification layer of each weak model is neglected, and a new global classification layer, taking into input the concatenation of the feature vectors provided by all the weak models, is introduced and trained to obtain the desired *ensemble*. Such an approach has been proven to give promising results in domain adaptation, as in the case of olive diseases [7], but extended analysis and diverse dataset partition methodology should be studied to assess the added value in robustness.

Ensembling is also suitable to support continuous learning. As already introduced based on the yearly campaign or on a steady stream of data coming from the field, the concept of a static dataset has to be surpassed. The data flow indeed offers the opportunity to update models based on deep learning to provide increasingly accurate answers by taking advantage of the expansion of the available case studies. To this end, it is neither practical nor convenient to retrain the models from

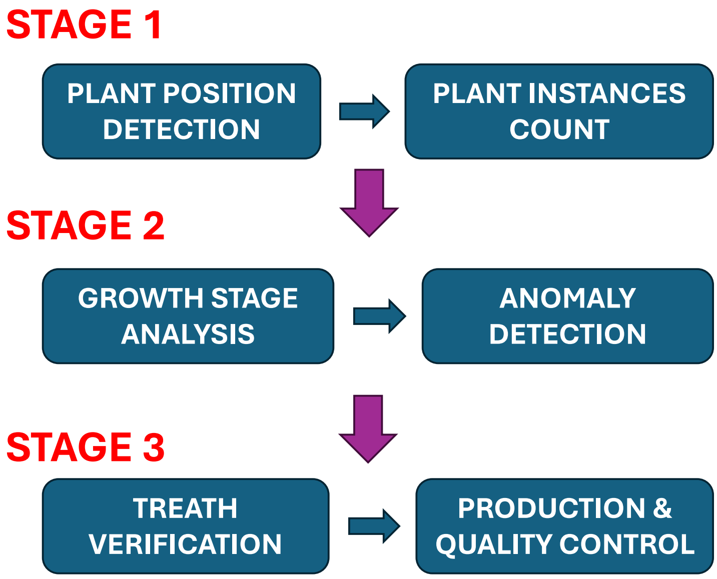


Fig. 1. Key steps diagram of a possible chain of activities as a rationale for plant monitoring and analysis

scratch at each update, but it is advisable to use a continuous learning approach.

The possibility of shifting toward a continual learning paradigm has significant potential: beyond providing constant retraining, it also enables enhanced models through continuous updates, making the system more resilient to unseen threats. This approach is more accurate and trustworthy than considering all boundary conditions simultaneously. While classical supervised deep learning algorithms can detect seasonal patterns, they often fail to accurately predict anomalous conditions. Moreover, they often fail to detect points of instability, which can adversely impact the evolution of the studied environment and potentially lead to catastrophic consequences.

From a technical perspective outside of research contexts, the use of ensemble methods is often not aligned with company objectives and means because it requires continuous resources. Other strategies, such as using state-of-the-art machine learning models with a priori studies of data distribution, can effectively produce one-shot models with an initial better performance. Ultimately, the support from advanced techniques demonstrates that moving beyond conventional methods can lead to developing more effective models, such as those achieved through ensemble approaches.

In the main studies of the AGROSAT+ project, it has become clear that transferring technology and know-how from the public to the private sector plays an important role. Even though large companies have the possibility and the means to sustain the production of high levels, in AGROSAT+, the resources employed in the developed DL model are far lower than the computational necessity of Large Language Models (LLMs). Indeed, the training of a state-of-the-art model [15] required only a mid-range workstation (equipped with two RTX QUADRO 5000 GPUs, which have now become an example of affordable accelerators), and the inference of the trained model worked on the CPU of this machine. This API solution lets users control their production directly with their phone (basic technologies approach). The proposed ensemble model was also used successfully in other scientific fields [24], showing the potential of open-access research.

Accuracy, Precision, Recall, F1 & R1 score and any other method largely treated in the statistical literature are the main methods to evaluate the goodness of a DL model. Still, the black-box nature of these algorithms hinders trust in their performance. The public is sceptical of their benefits since it is impossible to fully understand their inner working. For the same reason, the scientific community, with their government counterparts, is questioning the danger, limits, and rightfulness of the DL models. Good practices, such as strict control of no train-test data contamination, augmentation strategies, and eXplainable Artificial Intelligence (XAI), are common methods to ensure that the systems are accurate but also trustworthy and plausible. Knowledge-based DL algorithms are other possible solutions; in genomic and molecular biology, AlphaFold [25] is a good example of how to evaluate the quality of a model. AlphaFold architecture combines the transformer attention mechanism in pairs with the Evoformer module; this processes correctly evaluate the data of the biological sequence and the pair representation to output a new possible structure. Another possible solution is Physic-Informed Neural Networks (PiNNs) [26] that guide the systems' output towards valid output thanks to the incorporation of the boundary conditions of the described problem. The listed procedures suggest that leveraging information from crop traits could provide an intrinsic validation method for the model, as the proposed approach aligns with natural observations. Last, it is worth mentioning the possible benefits of incorporating continual learning strategies to validate the model over time. Continual learning enhances the adaptability of DL algorithms by enabling them to incrementally acquire information from new data while retaining the old ones used for the previous state. This approach not only mitigates the risk of catastrophic forgetting but also allows for dynamic updates, thereby outputting an unbiased overall accuracy of real-world phenomena. Consequently, the ability to control and fine-tune the model's performance across diverse tasks and datasets is significantly enhanced, ensuring that the model remains robust and effective over time.

D. Filtering and analysis of crowd-sensed data

In our previous experience with the AGROSAT+ project, researchers dealt with the quality of data collected from voluntary users. While the information provided, including new images to enhance the datasets, was effective in meeting the need for more varied spatial and temporal data, it is essential to implement suitable filtering to avoid errors or biases due to the non-authoritative nature of the information. To this end, one of the first elements integrated into the GranoScan app is a deep learning method, achieved through supervised learning, to differentiate relevant images from those that may not be suitable for a specific computer vision task. However, this approach can be improved and expanded by: a) incorporating blind general-purpose image quality assessment methods, such as those based on deep learning (e.g., [27], [28]), and b) developing appropriate object detectors to verify that the image is relevant to the computer vision task (for example, if the visual task involves identifying leaf diseases, there should be at least one leaf in the picture, and it should occupy a significant area). After passing through the specified filters and if the user provides feedback, the processed image can be stored in an expanded version of the datasets suitable for potential model updates and fine-tuning, also according to online procedures and to the continuous learning approach described in Section IV-C. Furthermore, additional filtering should be conducted to analyze the cross-correlation between contextual and image data. This is primarily focused on identifying potential anomalies within the data, such as a disease reported in a region of the world or during a time of year when the disease is not expected. While such anomalies may indicate the nonstationarity of the observed global situation (also as an outcome of climate change), they should be carefully reviewed by additional AI agents and, ultimately, by human observers. This is somewhat related to the continuous monitoring of the expert system in the operational phase to prevent biases and drifts and contribute to the overall maintainability of the system.

V. Conclusions

In this position paper, we have revised and enumerated challenges and opportunities for developing AI models that can tackle visual tasks relevant to agronomy. These models must exhibit high levels of robustness, adaptability, and maintainability to be considered trustworthy for deployment across various scenarios. Our proposed approach focuses on three key elements: developing technologies for model domain adaptation, utilizing crowd-sensing with awareness of uncertainty, and integrating with reasoning and recommendation systems to transform computational intelligence outputs into actionable knowledge. The synergy among these three points is also inspired by the general principles of responsibility, accountability, explainability, and trustworthiness, which collectively enhance the acceptability of our proposed solutions by addressing both technical and non-technical requirements.

Work is currently underway as part of the STRIVE project, and it will continue over the next two years. During this time, experiments will be conducted to test the proposed approach, evaluate its effectiveness, and understand its limitations. Additional measures will involve working with the community of farmers to raise awareness and encourage engagement.

The additional benefits of engaging the farming community in this precision approach to agricultural practices include building trust and improving perceptions of this tool.

An active community can guarantee a steady flow of data, allowing the continual learning implementation part of our solution, and communicate additional information, enabling real-time adjustments to the predictive component of the employed algorithm, which would otherwise not be possible.

In the future, we may consider utilizing Generative AI and LLMs to enhance communication and interaction with end users. However, we will proceed cautiously, as these technologies are not yet fully mature, language support is not consistent, and the portability and sustainability of the technology still need to be assessed. Therefore, we may need to postpone their application in scenarios where actual deployment is being pursued.

Acknowledgments

This work has been partially supported by the CNR FOE Project DIT.AD022.207 "La Scienza per le TRansizioni Industriale, Verde, Energetica - STRIVE" (Science for Industrial, Green and Energy Research).
This work is part of a project that has received funding from the European Union's Horizon Europe - Topic: DIGITAL-2022-SKILLS-03-SPECIALISED-EDU – Project ID: 101123258 - Project name: Digital agriculture for sustainable development(AGRITECH EU).

References

[1] T. Saranya, C. Deisy, S. Sridevi, and K. S. M. Anbananthen, "A comparative study of deep learning and internet of things for precision agriculture," *Engineering Applications of Artificial Intelligence*, 2023. doi: 10.1016/j.engappai.2023.106034

[2] I. Zualkernan, D. A. Abuhani, M. H. Hussain, J. Khan, and M. El-Mohandes, "Machine learning for precision agriculture using imagery from unmanned aerial vehicles (uavs): A survey," *Drones*, 2023. doi: 10.1016/j.compag.2020.105760

[3] S. Koul, "Machine learning and deep learning in agriculture," *Smart Agriculture: Emerging Pedagogies of Deep Learning, Machine Learning and Internet of Things*, 2021. doi: 10.1201/b22627-1

[4] R. Priya and D. Ramesh, "Ml based sustainable precision agriculture: A future generation perspective," *Sustainable Computing: Informatics and Systems*, 2020. doi: 10.1016/j.suscom.2020.100439

[5] B. Antonio, D. Moroni, and M. Martinelli, "Efficient adaptive ensembling for image classification," *Expert Systems*, 2022. doi: 10.1111/exsy.13424

[6] Ł. Błaszczyk, M. Mizura, A. Płocharski, and J. Porter-Sobieraj, "Simulating large-scale topographic terrain features with reservoirs and flowing water," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F2137

[7] A. Bruno, D. Moroni, and M. Martinelli, "Efficient deep learning approach for olive disease classification," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F4794

[8] G. Castellano, P. De Marinis, and G. Vessio, "Applying knowledge distillation to improve weed mapping with drones," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F960

[9] N. Iqbal, C. Manss, C. Scholz, D. König, M. Igelbrink, and A. Ruckelshausen, "Ai-based maize and weeds detection on the edge with cornweed dataset," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F2125

[10] S. Kolhar and J. Jagtap, "Plant trait estimation and classification studies in plant phenotyping using machine vision–a review," *Information Processing in Agriculture*, vol. 10, no. 1, pp. 114–135, 2023.

[11] European Commission, "Work programme 2023-2025 – 9. food, bioeconomy, natural resources, agriculture and environment," 2024. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2023-2024/wp-9-food-bioeconomy-natural-resources-agriculture-and-environment_horizon-2023-2024_en.pdf

[12] "THE 17 GOALS," https://sdgs.un.org/goals, 2015, [Online; accessed 11-June-2024].

[13] Y. Lu and S. Young, "A survey of public datasets for computer vision tasks in precision agriculture," *Computers and Electronics in Agriculture*, 2020. doi: 10.1016/j.compag.2020.105760

[14] D. Hughes, M. Salathé *et al.*, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060*, 2015. doi: 10.48550/arXiv.1511.08060

[15] A. Bruno, D. Moroni, R. Dainelli, L. Rocchi, S. Morelli, E. Ferrari, P. Toscano, and M. Martinelli, "Improving plant disease classification by adaptive minimal ensembling," *Frontiers in Artificial Intelligence*, 2022. doi: 10.3389/frai.2022.868926

[16] R. Dainelli, M. Martinelli, A. Bruno, D. Moroni, S. Morelli, M. Silvestri, E. Ferrari, L. Rocchi, and P. Toscano, "A phenotyping weeds image dataset for open scientific research," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7598372

[17] ——, *49. Recognition of weeds in cereals using AI architecture*. Wageningen Academic, 2023.

[18] R. Dainelli, A. Bruno, M. Martinelli, D. Moroni, L. Rocchi, S. Morelli, E. Ferrari, M. Silvestri, S. Agostinelli, P. La Cava *et al.*, "Granoscan: an ai-powered mobile app for in-field identification of biotic threats of wheat," *Frontiers in Plant Science*, 2024. doi: 10.3389/fpls.2024.1298791

[19] T. Kattenborn, "Planttraits2023," 2023. [Online]. Available: https://kaggle.com/competitions/planttraits2023

[20] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?" *arXiv preprint arXiv:2112.00639*, 2021. doi: 10.48550/arXiv.2112.00639

[21] K. Kirkpatrick, "The carbon footprint of artificial intelligence," *Communications of the ACM*, 2023. doi: 10.1145/3603746

[22] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, 2020. doi: 10.48550/arXiv.1904.05046

[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019. doi: 10.48550/arXiv.1905.11946

[24] A. Bruno, C. Caudai, G. R. Leone, M. Martinelli, D. Moroni, and F. Crotti, "Medical waste sorting: a computer vision approach for assisted primary sorting," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023. doi: 10.48550/arXiv.2303.04720

[25] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, 2021. doi: 10.1038/s41586-021-03819-2

[26] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, 2021. doi: 10.1038/s42254-021-00314-5

[27] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE transactions on neural networks and learning systems*, 2014. doi: 10.1109/TNNLS.2014.2336852

[28] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, 2018. doi: 10.1007/s11760-017-1166-8

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 41–48
DOI: 10.15439/2024F7887
ISSN 2300-5963 ACSIS, Vol. 40

The Synergy of Interpolative Boolean Algebra and Ordinal Sums of Conjunctive and Disjunctive Functions in Stock Price Trend Prediction

Pavle Milošević, Aleksandar Rakićević, Milica Zukanović
0000-0002-5943-6023
0000-0002-8917-7229
0000-0003-3650-8327
University of Belgrade– Faculty of Organizational Sciences, Jove Ilića 154, 11000 Belgrade, Serbia
Email: {pavle.milosevic, aleksandar.rakicevic, milica.zukanovic,@fon.bg.ac.rs}

Miroslav Hudec
0000-0002-2868-0322
Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1. 852 35 Bratislava, Slovakia
Faculty of Economics, VSB - Technical University of Ostrava, 17. listopadu 15, 708 00 Ostrava-Poruba, Czech Republic
Email: miroslav.hudec@euba.sk

Nina Barčáková, Eva Rakovská
0000-0002-7382-9701
0000-0003-6191-184X
Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1. 852 35 Bratislava, Slovakia
Email: {nina.barcakova, eva.rakovska,@euba.sk}

***Abstract*—Stock price prediction is crucial for accurate investment decision-making and widely regarded as one of the most important tasks in finance. Investors and financial professionals rely on a wide range of input data, such as market information, technical analysis, and fundamental analysis, to make informed decisions. When it comes to financial data, it is important to incorporate the logical dependencies of inputs into the modeling and prediction process. Therefore, logic-based approaches are considered adequate for solving such problems. This paper proposes a novel logic-based approach to stock price trend prediction based on Interpolative Boolean algebra (IBA) and ordinal sums of conjunctive and disjunctive (OSCD) functions. This is the very first paper that aims to explore the synergy of these two approaches in a real-world setting, utilizing their comparative advantages in different phases of modeling. The proposed approach is tested on a sample of 23 companies from the S&P500 over the past three years. The paper also presents the results of the application of the proposed model for the analyzed companies.**

***Index Terms*—Interpolative Boolean Algebra, Ordinal Sums of Conjunctive and Disjunctive Functions, Price Trend Forecasting, S&P 500.**

I. Introduction

FINANCIAL markets have been an attractive research field for application of artificial intelligence (AI) techniques. The most challenging task in financial markets is to forecast prices because of their dynamic, complex, evolutionary, nonlinear, nonparametric, and chaotic nature [1].

To predict stock market movements, researchers use different types of structured and unstructured inputs. Structured inputs are based on market information (such as stock prices, volumes, spread), technical analysis (including technical indicators and chart patterns), and fundamental analysis (macroeconomic indicators, financial statement ratios). The unstructured inputs include news (general financial news, company news) and posts from social networks (such as Twitter, Reddit, Facebook). Still, technical indicators and financial statement ratios are the most commonly used inputs for predicting market movements [2] – [4].

Due to its ability to recognize patterns in data, AI and machine learning (ML) techniques are extensively used for this purpose [2], [5] – [7] with deep learning models strongly emerging as the most promising predictors [8] – [11]. Being flexible and able to fit complex data, ML models don't require a theoretical understanding of the problem, nor strong assumptions about the data. However, they lack interpretability, which complicates the extraction of knowledge from the data. In the real-world setting, the predictions given by the algorithm have to be fully understandable and interpretable to financial decision makers. To overcome this issue, one could use logic for modeling to imitate human reasoning. Keeping in mind the weak theoretical background of stock price movements, researchers usually combine (fuzzy) logic with some learning/optimization algorithms to create hybrid prediction models like neuro-fuzzy systems [12], [13], deep convolutional-fuzzy systems [14], evolutionary-fuzzy systems [15], etc. Still, there are very few authors who have tried to use an expert-based (fuzzy) logic models solely for financial forecast. One of the examples is the application of Interpolative Boolean algebra (IBA) for portfolio selection [16].

On the other hand, data aggregation underpins almost every stock price trend prediction system. The careful selection of the function for aggregating financial indicators into a single indicator is of paramount importance. In other words, it is necessary to choose a function that has the desired mathematical properties, and at the same time is comprehensible even to decision makers with a not so strong mathematical background. One possible direction is to employ simple yet effective aggregation functions, such as weighted sum [17] or order weighted sum operators [18]. The other prominent approach implies using logic-based aggregation methods in a

This study was supported by the Ministry of Science, Technological Development and Innovation of Serbia, and the Slovak Research and Development Agency under Grant No. 337-00-3/2024-05/18 on the Serbian side and APVV No. SK-SRB-23-0007 on the Slovak side. The partial support of the SGS project No. SP2024/083 of the Ministry of Education, Youth and Sports of the Czech Republic is kindly appreciated.

Topical area: Advanced Artificial Intelligence in Applications

broader sense, e.g. Choquet integral [19]. Finally, pure logic-based approaches based on fuzzy or multi-valued are frequently used [20]. However, mixed aggregation functions adapted to data to cover conjunctive, disjunctive and averaging behavior might be beneficial.

In this study, we utilize logic-based aggregation to model future stock price movements based on a selected set of financial ratios, primarily employing fundamental analysis. We engage two logic-based aggregation methods: Interpolative Boolean algebra and Ordinal Sums of Conjunctive and Disjunctive Functions (OSCD), to construct models for price trend prediction. OSCD serves for the aggregation of financial ratios within each group. Subsequently, IBA-based aggregation is employed to amalgamate the ordinal sums of the groups. To evaluate the proposed meta-model, we utilize a dataset comprising market information and financial ratio data of the S&P 500 companies over a three-year period on a quarterly basis. The price trend prediction task is formulated as a binary classification problem.

It's important to recognize that there are logical connections or dependencies among financial indicators within individual groups. For example, when the values of two indicators are high, they reinforce each other, leading to an upward trend (or increased satisfaction). Conversely, when values are low, it results in downward reinforcement. When some indicators are high while others are low, the satisfaction level falls somewhere in between. Therefore, for such cases, we require aggregation functions with mixed behavior [21]. One option is uninorms [21], but due to the non-continuity of representative uninorms [22], an alternative could be ordinal sums of conjunctive and disjunctive functions [23].

The next question is how to aggregate the higher-level results using ordinal sums of the indicator groups. We have chosen logical aggregation (LA) based on IBA for several reasons [24]. First, LA is a sophisticated multi-valued aggregation approach within the Boolean framework that provides clear guidelines for data aggregation, from verbal descriptions to the final mathematical aggregation model. Finally, aggregation models based on LA are fully transparent and interpretable for decision-makers.

The structure of this paper is given as follows. In the next two sections, a short overview of a theoretical background for the two aggregation methods, OSCD and logical aggregation based on IBA, is given. In Section 4 we explain the problem setup, describe the dataset, and propose the model. Finally, in the last two sections we present and discuss the experimental results and conclude the paper.

II. Interpolative Boolean Algebra

Interpolative Boolean Algebra is a consistent multi-valued realization of Boolean algebra, preserving all the laws on which Boolean algebra rests [25]. Namely, IBA is proposed as an answer for the disregard of the law of exclusion of the third and contradiction in the classical phase of logic and it serves as a fundamental component for various multi-valued techniques and methods [26]-[28].

IBA consists of two levels: the symbolic and the value levels. At the symbolic level of IBA, the structure of attributes is taken into account, while at the value level, values are assigned and the final resulting value of the expression in the Boolean framework is calculated [24]. The principle of structural functionality dictates that IBA transformations are performed at the symbolic level before introducing values. This ensures that negation is treated differently compared to traditional fuzzy approaches. Focusing on the structure preserves all Boolean laws in the multivalued case, including the laws of excluded middle and contradiction, which is the main contribution of IBA.

In the IBA framework, attributes/inputs are called primary attributes. These attributes are elements of logical functions within IBA. The value realization of primary attributes within the classical Boolean algebra implies the use of two values 0 and 1, while within the IBA primary attributes have [0,1]-valued realization. All logical functions over primary attributes represent elements of IBA. On the other hand, IBA is based on atomic elements of Boolean algebra [25]. Atomic elements are the simplest elements of Boolean algebra. They are logical functions that do not contain any other Boolean element except itself and 0 constant, e.g. Boolean algebra over two attributes has four atoms and they are $p_1 \wedge p_2, \neg p_1 \wedge p_2, p_1 \wedge \neg p_2$ and $\neg p_1 \wedge \neg p_2$.

The inclusion of atoms in a logical expression is the basis for the introduction of the structure of a logical expression and the structural level of IBA. Any logical expression is interpreted as a scalar product [16]:

$$\varphi(p_1, \dots, p_n) = P \cdot S(p_1, \dots, p_n) \tag{1}$$

where P is vector of atomic elements and $S(p_1, \dots, p_n)$ is the structural vector.

At the symbolic level, the given logical expression is transformed into a generalized Boolean polynomial (GBP) based on the transformation rules [24]:

$$(\alpha(p_1, \dots, p_n) \wedge \beta(p_1, \dots, p_n))^{\otimes} = \alpha^{\otimes}(p_1, \dots, p_n) \otimes \beta^{\otimes}(p_1, \dots, p_n) \tag{2}$$

$$(\alpha(p_1, \dots, p_n) \vee \beta(p_1, \dots, p_n))^{\otimes} = \alpha^{\otimes}(p_1, \dots, p_n) + \beta^{\otimes}(p_1, \dots, p_n) - \alpha^{\otimes}(p_1, \dots, p_n) \otimes \beta^{\otimes}(p_1, \dots, p_n) \tag{3}$$

$$(\neg\alpha(p_1, \dots, p_n))^{\otimes} = 1 - \alpha^{\otimes}(p_1, \dots, p_n) \tag{4}$$

where $\alpha(p_1, \dots, p_n)$ and $\beta(p_1, \dots, p_n)$ are complex elements of Boolean algebra.

When it comes to the primary attributes $p_1, \dots, p_n$, the following transformation rules applies [24]:

$$(p_i \wedge p_j)^{\otimes} = \begin{cases} p_i \otimes p_j, & i \neq j \\ p_i, & i = j \end{cases} \tag{5}$$

$$(p_i \vee p_j)^{\otimes} = p_i + p_j - p_i \otimes p_j \tag{6}$$

$$(\neg p_i)^{\otimes} = 1 - p_i \tag{7}$$

IBA transformations on the symbolic level enable the preservation of Boolean laws in general case. Furthermore, the first GBP transformation rule for primary attributes (5) ensures idempotency within the IBA framework.

At the value level, each element of Boolean algebra is realized by GBP. In GBP, alongside standard arithmetic addition and subtraction operations, and the generalized product (GP). GP, a binary operator on the unit interval, belongs to a subclass of t-norms that satisfies the non-negativity condition. It takes precedence as the highest priority operation within the expression. GP can be any function greater than the Lukasiewicz operator and less than the minimum [29]:

$$\max(p_1 + p_2 - 1{,}0) \leq p_1 \otimes p_2 \leq \min(p_1, p_2) \qquad (8)$$

In the IBA framework, the choice of operators for generalized products depends on the nature of the attributes and their correlation. Specifically, the Lukasiewicz operator is utilized for aggregating attributes of opposite nature, i.e., negatively correlated variables. The standard product operator is employed for uncorrelated variables. Finally, the minimum operator is applied for the aggregation of attributes of the same or similar nature, i.e., highly correlated variables.

From the practical standpoint, the two most successful application areas of IBA are logical aggregation and the IBA similarity measure, which have been used in various fields [27], [28], [30] – [33].

Logical aggregation (LA) is a Boolean consistent and fully transparent aggregation technique based on IBA [24]. It implies that the normalized values of the input variables are aggregated using GBP or weighted sum of GBPs into the resulting, globally representative value. The main advantage of LA compared to traditional aggregation methods is that it enables modelling of complex logical connections that may exist in problems where the human factor is particularly important [34].

Hence, understanding and analysing LA functions is simple, and it has found application in many areas of finance. For instance, IBA-based methods for portfolio selection showed promising results and their average monthly returns were aligned with the performance of the S&P500 index [16]. Also, a system for automatic trading on the stock market based on IBA is proposed [35]. Moreover, authors in [30], [36] used logic-based approach for financial ratio analysis of a company's performance.

III. Ordinal Sums

Mixed aggregation functions are able to adjust to data in the sense of reinforcing or averaging. One category are ordinal sums of conjunctive and disjunctive functions depicted in Fig 1.

The ordinal sum considering two attributes is an aggregation function on [0,1] [23]:

$$A(\mathrm{x},\mathrm{y}) = A_1(\mathrm{a} \wedge \mathrm{x}, \mathrm{a} \wedge \mathrm{y}) + A_2(\mathrm{a} \vee \mathrm{x}, \mathrm{a} \vee \mathrm{y}) + \mathrm{a} \qquad (9)$$

where

- for $x, y \in [0, a]^2$ we get $\mathrm{A}(\mathrm{x},\mathrm{y}) = A_1(\mathrm{x},\mathrm{y})$
- for $x, y \in [a, 1]^2$ we get $\mathrm{A}(\mathrm{x},\mathrm{y}) = A_2(\mathrm{x},\mathrm{y})$
- for $x, y \in [0, a] \times [a, 1]$ we get $A(\mathrm{x},\mathrm{y}) = x + y - a$
- for $x, y \in [a, 1] \times [0, a]$ we get $\mathrm{A}(\mathrm{x},\mathrm{y}) = x + y - a$

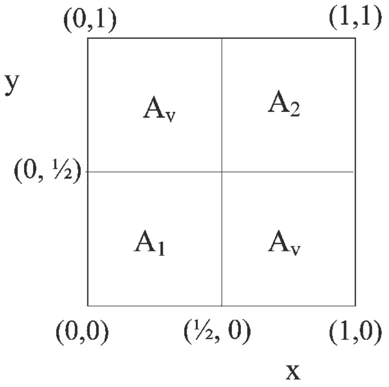


Fig. 1. Graphical illustration of ordinal sums

When A_1 is conjunctive and A_2 is disjunctive function, then in the remaining two sub squares are described with averaging function:

$$\mathrm{A}(\mathrm{x},\mathrm{y}) = A_v(x, y) = x + y - a \qquad (10)$$

Observe that, for conjunctive function we should keep neutral element $A_1(\mathrm{a},\mathrm{a}) = \mathrm{a}$ [23]. For a=0.5 and product t-norm we get:

$$A_1(\mathrm{x},\mathrm{y}) = 2 \cdot \mathrm{x} \cdot \mathrm{y} \qquad (11)$$

Analogously holds for disjunctive function. Hence, for probabilistic sum t-conorm we get:

$$A_2(\mathrm{x},\mathrm{y}) = -1 + 2 \cdot \mathrm{x} + 2 \cdot \mathrm{y} - 2 \cdot \mathrm{x} \cdot \mathrm{y} \qquad (12)$$

The same observation holds for other t-norms and t-conorms. We introduce here only functions used in this work. Lukasiewicz t-norm (as a representative of nilpotent functions) is:

$$A_1(\mathrm{x},\mathrm{y}) = \max(0, \mathrm{x} + \mathrm{y} - 0.5) \qquad (13)$$

while its dual t-conorm is:

$$A_2(\mathrm{x},\mathrm{y}) = \min(0, \mathrm{x} + \mathrm{y} - 0.5) \qquad (14)$$

In this way, we are able to upwardly reinforce high values (or assign them value 1), downwardly reinforce low values (or even assign them value 0) and assign average value for the other cases. In the case of averaging behavior, we can manage inclination towards conjunctive, or disjunctive behavior. In the case of careful or pessimistic evaluation (conjunctive inclination), we adopt, e.g., geometric mean for averaging part as:

$$A_v(\mathrm{x},\mathrm{y}) = 2 \cdot \mathrm{x} \cdot \mathrm{y} \qquad (15)$$

Observe that, even though (11) and (15) have the same structure, they are applied on different sub squares and therefore behave differently.

IV. Materials and Methods

A. The problem setup

Financial prediction poses a complex and challenging problem, inherently fraught with uncertainty and risk. Consequently, it may not always accurately foresee future outcomes due to unforeseen events, market volatility, or changes in underlying assumptions. Achieving accurate predictions is a primary challenge, alongside the imperative of rendering them comprehensible to decision-makers and fully interpretable.

Various analytical techniques are employed in financial prediction, encompassing fundamental, technical, and sentiment analysis, alongside machine learning or statistical tools. However, despite the breadth of methodologies, financial predictions remain vulnerable to uncertainty and risk. Thus, ensuring their accuracy, comprehensibility, and interpretability becomes paramount.

In this paper, we aim to address a stock trend prediction problem, i.e., forecasting the future direction of stock prices by identifying patterns in historical stock price data. For the purpose of this paper, we will consider this problem as a binary classification. Namely, the main goal is to determine which companies' stock should be bought and which stock should be sold. Stocks that should be bought are the ones which will increase in price in the future, while the stocks with decreasing value are considered to be sold.

The success of the proposed models will be measured using standard metrics for binary classification: receiver operating characteristic (ROC) curve / area under the curve (AUC), together with precision, recall and F1 metric.

B. *Dataset*

This study employs a dataset consisting of financial ratios for 23 companies that constitute the S&P500 index. Companies are selected to cover different industries, such as the Energy and Materials sectors, to Media & Entertainment companies. Furthermore, companies are chosen as a balanced sample in terms of upward and downward price trends during the observed period of time. Finally, each company is described with fundamental financial indicators collected on a quarterly basis for the period of three years (from December 2021 to December 2023). The final dataset consists of 201 instances, reflecting the availability of data and the fact that some of the chosen companies were not in the S&P 500 index throughout the entire observed period. A detailed overview of the companies and the industries they belong to is provided in Table I.

Based on the literature review and recommendations of experts in the field, eight financial indicators were chosen as inputs for the experiment. The selected attributes differ in nature and cover four aspects of a company's performance and financial health: activity, liquidity, cash flow, and investment ratios. From each group, the two most significant indicators were chosen. Net operating assets to total assets (a_1) and asset turnover (a_2) are identified as representatives of activity ratios, cash and cash equivalents to total assets (l_1) and working capital to total assets (l_2) from liquidity indicators, ratio of operating cash flow to total assets (c_1) and free cash flow yield (c_2) from cash flow indicators and earnings per share to price (i_1) and enterprise value growth rate (i_2) from investment ratios.

TABLE I.
OVERVIEW OF COMPANIES AND INDUSTRIES

Company	**Industry**
Accenture (ACN)	Software & Services
Align Technology (ALGN)	Health Care Equipment & Services
Amcor (AMCR)	Materials – Containers & Packaging
Broadcom Inc. (AVGO)	Semiconductors & Semiconductor Equipment
Bristol Myers Squibb (BMY)	Pharmaceuticals, Biotechnology & Life Sciences
Corteva (CTVA)	Materials - Chemicals
Dow Inc. (DOW)	Materials – Chemicals
Fox Corp. Class B (FOX)	Media & Entertainment
Fox Corp. Class A (FOXA)	Media & Entertainment
Fortinet (FTNT)	Software & Services
Hasbro (HAS)	Consumer Durables & Apparel
Gartner (IT)	Materials – Containers & Packaging
Mastercard (MA)	Financial Services
3M (MMM)	Capital Goods
Paramount Global (PARA)	Media & Entertainment
Paychex (PAYX)	Commercial & Professional Services
Pool Corp. (POOL)	Consumer Discretionary Distribution & Retail
Phillips 66 (PSX)	Energy – Oil, Gas & Consumable Fuels
Qorvo (QRVO)	Semiconductors & Semiconductor Equipment
Uber (UBER)	Transportation
Vici Properties (VICI)	Equity Real Estate Investment Trusts (REITs)
Warner Bros. Discovery (WBD)	Media & Entertainment
Welltower Inc (WELL)	Equity Real Estate Investment Trusts (REITs)

As data are on different scales, the usual step for any data mining task is normalization. It is also relevant in mining patterns from data by logic aggregation usually working on the unit interval. A simple normalization might cause outliers skew the normalization [37]. Hence, we applied normalization based on the inter quartile distribution in the following way. All values lower than $L = \mathrm{L} = \mathrm{Q1} - 1.5 \cdot (\mathrm{Q3} - \mathrm{Q1})$ are transformed to 0, while all values greater than $\mathrm{H} = \mathrm{Q3} + 1.5 \cdot (\mathrm{Q3} - \mathrm{Q1})$ are transformed into 1, where *Q1* and *Q3* are the first and third quartile, respectively.

Inner values are normalized as (see Fig. 2).

$$x^* = \frac{x-L}{H-L} \tag{16}$$

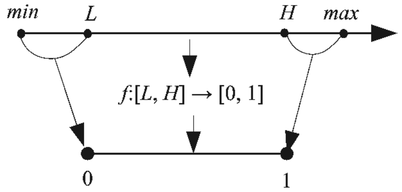


Fig. 2. Normalization into the unit interval adopted from **[35]**

The output variable in our case are returns of a stock price, i.e. the change in the price of a stock over a certain period of time. According to the returns, instances are divided into two classes depending on whether they are positive or negative. Accordingly, 1 is assigned to positive and 0 for negative returns.

C. Meta-model

The majority of models for financial prediction are based on a black-box approach, lacking interpretability, which is often crucial for decision-makers. Additionally, data on financial markets often incorporate a certain extent of uncertainty and vagueness that may hinder decision-making. Therefore, logic-based approaches seem to be a natural direction for the development of our model.

The proposed model respects the hierarchy of selected inputs; that is, first, we will obtain aggregated indicators of activity (A), liquidity (L), cash flow (C) and investment ratios (I), and then the final assessment for a company ($score$). The model is entirely based on logic-based techniques.

On one hand, we have chosen OC as an aggregation operation at the group level, owing to its capability to generate a minimized aggregation score for small attribute values and maximize aggregation scores for attributes with large values. This would result in a clear separation of preferred values from those representing potentially risky situations for further aggregation at the group level. Therefore, scores at the level of the groups are calculated as follows:

$$A = OSDC(a_1, a_2) \tag{17}$$

$$L = OSDC(l_1, l_2) \tag{18}$$

$$C = OSDC(c_1, c_2) \tag{19}$$

$$I = OSDC(i_1, i_2) \tag{20}$$

On the other hand, we have implemented IBA-based logical aggregation as a final aggregation function due to the mathematical characteristics of IBA, along with its explainability. The final score is obtained as an LA of the scores at the level of the group:

$$score = LA(A, L, C, I) \tag{21}$$

From a mathematical standpoint, the main benefit of using IBA-based aggregation/decision-making models lies in their ability to perform a fine gradation of instances using the [0,1] approach, which remains consistent with the Boolean frame. From a practical standpoint, IBA-based models are fully interpretable and transparent. The process of IBA expert-based modeling starts with formulating a clear verbal model that comprehensively addresses the needs and preferences of decision-makers. This model can be easily articulated and explained to executives lacking significant mathematical knowledge or familiarity with IBA. Furthermore, the verbal model can be easily interpreted as a logical/mathematical model, i.e., a single logical aggregation. Subsequently, the logical model is transformed into a suitable GPB, either manually or utilizing existing software solutions [26].

$$score = GBP(A, L, C, I) \tag{22}$$

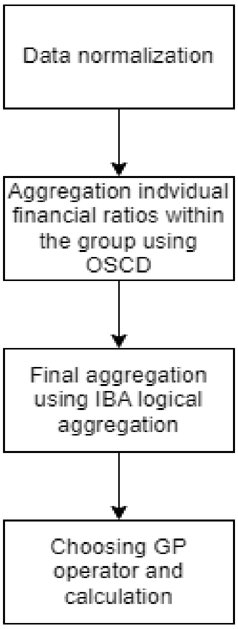

Data normalization
Aggregation indvidual financial ratios within the group using OSCD
Final aggregation using IBA logical aggregation
Choosing GP operator and calculation

Fig 3. The proposed meta-model

As a final step, the GP operator is chosen based on data correlation, and the resulting value for each instance is calculated. Finally, the proposed metamodel is illustrated in Fig. 1.

D. Model realizations

In order to calculate scores at the group level, three different OSCD operators are used: 1) OSCD based on product t-norm, probabilistic sum, and arithmetic mean; 2) OSCD based on product t-norm, probabilistic sum, and geometric mean; 3) Lukasiewicz t-norm and t-conorm. These OSCD operators are presented in Figures 4-6.

The first two OSCD operators differ only in the mean operator, while the third one is based on a different t-norm. All three operators are thoroughly discussed and elaborated from a theoretical point of view in the literature [23]. Still, this may be seen as an attempt to investigate their practical value from the perspective of the presented problem. In other words, the experiment will show which of these operators is the best choice for score calculation at the group level. The value of the parameter a was chosen according to the literature [23].

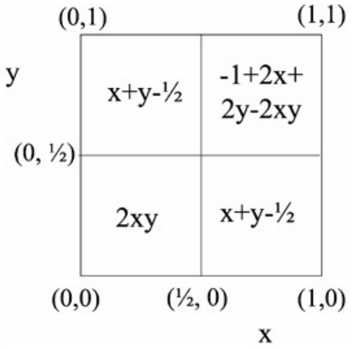


Fig 4. The graphical interpretation of OSCD with standard product t-norm, probabilistic sum and arithmetic mean **[23]**

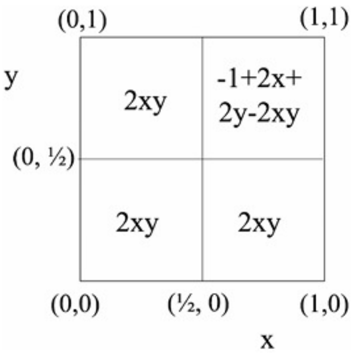


Fig 5. The graphical interpretation of OSCD with standard product t-norm, probabilistic sum and geometric mean **[23]**

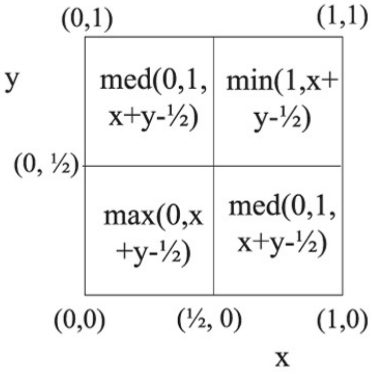


Fig 6. The graphical interpretation of OSCD with Lukasiewicz t-norm and t-conorm **[23]**

The verbal model may be easily translated into the following logical aggregation model:

$$LA(A,L,C,I) = \big(\neg I \wedge C \wedge (A \vee L)\big) \vee (I \wedge \neg C \wedge L) \quad (23)$$

Afterwards, the logical aggregation model is treated within the IBA framework and transformed into the corresponding GBP.

$$score = GBP(A,L,C,I) = A \otimes C + L \otimes C + C \otimes I - A \otimes L \otimes C - A \otimes C \otimes I - 2L \otimes C \otimes I + A \otimes L \otimes C \otimes I \quad (24)$$

The final step in the proposed approach involves choosing the GP operator. Considering that various groups of financial ratios are to be aggregated, the standard product appears to be a natural choice. This choice is supported by correlation analysis. Indeed, correlations between groups of aggregated indicators are illustrated in Figures 7-9. Since correlation coefficients are not high, the product operator is selected for the GP [39]. Therefore, the final aggregation score is calculated using the following expression:

$$score = A \cdot C + L \cdot C + C \cdot I - A \cdot L \cdot C - A \cdot C \cdot I - 2 \cdot L \cdot C \cdot I + A \cdot L \cdot C \cdot I \quad (25)$$

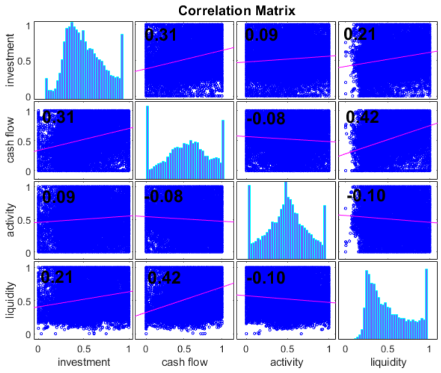


Fig 7. Correlation Matrix of scores at the level of the group calculated using OSCD with standard product t-norm, probabilistic sum and arithmetic mean

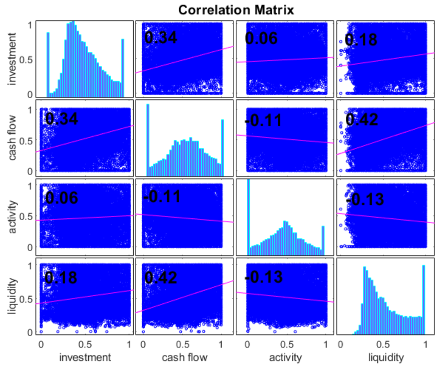


Fig 8. Correlation Matrix of scores at the level of the group calculated using OSCD with standard product t-norm, probabilistic sum and geometric mean

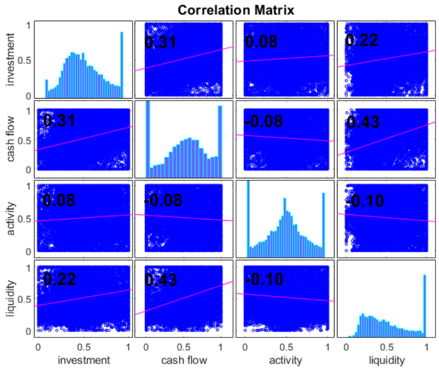


Fig 9. Correlation Matrix of scores at the level of the group calculated using OSCD with Lukasiewicz t-norm and t-conorm

V. Experimental Results

The proposed approach was validated by comparing the predicted values with the actual class of stock trend. The acquired dataset is balanced, consisting of 101 instances (50.25%) with a negative trend, assigned class 0, while 100 instances (49.75%) exhibit a positive trend, assigned class 1.

The performance of classification models was evaluated using AUC/ROC. These metrics provide a comprehensive measure of model performance in terms of binary classification. The model with the group calculated using OSCD with Lukasiewicz t-norm and t-conorm (M3) achieved an AUC of 0.7223, indicating a good ability to distinguish between positive and negative stock trends. The model with the group calculated using OSCD with standard product t-norm, probabilistic sum, and arithmetic mean (M1) followed with 0.7206, and the model with the group calculated using OSCD with standard product t-norm, probabilistic sum, and geometric mean (M2) with 0.7036. The ROC curves for each model were plotted to visualize their performance. In Fig. 10, it can be seen that the curves for M1 and M3 overlap, while the ROC for M2 is clearly worse.

Therefore, we may conclude that the choice of OSCD functions has a significant influence on the results of classification. It seems that the geometric mean operator has a negative influence on the results. On the other hand, the t-norm/t-conorm operator did not have a strong influence, since M1 and M2 utilize different operators.

In order to perform more detailed analysis, confusion matrices for all three models are calculated for a chosen threshold, while classification performance metrics are presented in Table II. The best values are marked in bold font.

For the chosen threshold, the model with the group calculated using OSCD with Lukasiewicz t-norm and t-conorm outperformed the remaining two models. Bearing in mind that we do not rely on ML approaches and the difficulty of the problem, AUC of 0.7223 is satisfactory. High precision in this domain is crucial because false positives can lead to significant costs. Together with recall and F1 score, the models suggest their reliability and effectiveness in real-world applications.

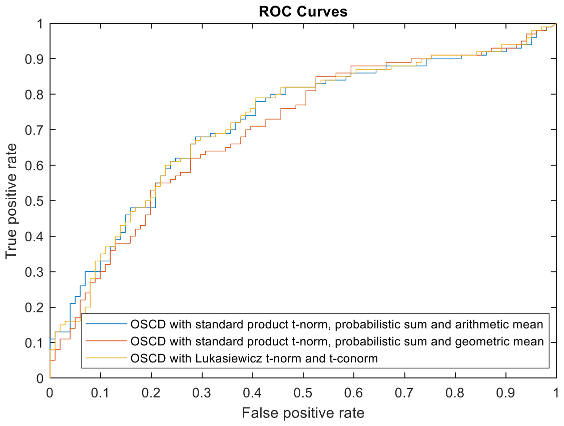


Fig 10. ROC curves for proposed models

TABLE III.
PERFORMANCE OF CLASSIFICATION

	M1	M2	M3
AUC	0.7206	0.7036	**0.7223**
Precision	0.6923	0.6737	**0.7021**
Recall	0.6300	0.6400	**0.6600**
F1 score	0.6597	0.6564	**0.6804**

VI. CONCLUSION

In this research, a combination of two logic-based aggregation methods is proposed for stock trend prediction. To the best of our knowledge, this is the first time such an approach with IBA and OSCD has been proposed. Although these approaches share a similar background, they have different mathematical properties and the potential to model different real-world situations. In this paper, on one hand, the financial indicators within the groups are aggregated by OSCD to perform the maximization/minimization according to the input. On the other hand, IBA, as a consistent real-valued generalization of classical Boolean algebra, is used as a framework for modeling and the logical aggregation of groups of financial indicators.

In this study, we have collected and utilized financial data from 23 companies across various industries to test the proposed approach. The overall results provide evidence that the proposed approach can be used as a tool for investment decision-making in any industry sector. Moreover, based on the obtained results, companies can be analyzed and ranked.

The synergy of interpolative Boolean algebra and ordinal sums of conjunctive and disjunctive functions is explored for the first time in this research. The initial results are optimistic, so in future work, we can consider parametric classes of these functions as suggested in [40] to fine-tune data evaluation. For this task, we need a domain expert to express the desired inclination and a higher amount of data to learn parameters, for example, using genetic algorithms. Furthermore, future work will be oriented towards the development of a more complex system that includes a broader range of financial indicators covering different aspects of a company's financial performance. Additionally, in this paper, two different aggregation methods are used together.

REFERENCES

[1] 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 Systems with Applications*, vol. 115, 2019, pp. 635-655, https://doi.org/10.1016/j.eswa.2018.08.003

[2] O. Bustos, and A. Pomares-Quimbaya, "Stock market movement forecast: A systematic review," *Expert Systems with Applications*, vol. 156, 2020, pp. 113464, https://doi.org/10.1016/j.eswa.2020.113464

[3] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no.4, 2020, pp. 3007-3057, https://doi.org/10.1007/s10462-019-09754-z

[4] T. W. Lee, P. Teisseyre, and J. Lee, "Effective exploitation of macroeconomic indicators for stock direction classification using the multimodal fusion transformer," *IEEE Access*, *11*, 2023, 10275-10287, https://doi.org/10.1109/ACCESS.2023.3240422

[5] B.M. Henrique, V.A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, 2019, pp.226-251, https://doi.org/10.1016/j.eswa.2019.01.012

[6] D.P. Gandhmal, and K. Kumar, "Systematic analysis and review of stock market prediction techniques," *Computer Science Review*, vol. 34, 2019, pp.100190, https://doi.org/10.1016/j.cosrev.2019.08.001

[7] M.M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review", *Expert Systems with Applications*, vol. 197, 2022, pp.116659, https://doi.org/10.1016/j.eswa.2022.116659

[8] O. B. Sezer, M. U. Gudelek, and A.M. Ozbayoglu, “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019,” *Applied soft computing*, vol. 90, 2020, pp. 106181, https://doi.org/10.1016/j.asoc.2020.106181

[9] W. Jiang, “Applications of deep learning in stock market prediction: recent progress,” *Expert Systems with Applications*, vol. 184, 2021, pp.115537, https://doi.org/10.1016/j.eswa.2021.115537

[10] R. Corizzo, and J. Rosen, “Stock market prediction with time series data and news headlines: a stacking ensemble approach*,” J Intell Inf Syst, 62* , 2024, pp. 27–56, https://doi.org/10.1007/s10844-023-00804-1

[11] J. Borst, L. Wehrheim, A. Niekler, and M. Burghardt, “An Evaluation of a Zero-Shot Approach to Aspect-Based Sentiment Classification in Historic German Stock Market Reports,” In *FedCSIS (Communication Papers)*, 2023, pp. 51-60, http://dx.doi.org/10.15439/2023F3725

[12] G.S. Atsalakis, I.G. Atsalaki, F. Pasiouras, and C. Zopounidis, “Bitcoin price forecasting with neuro-fuzzy techniques*,” European Journal of Operational Research*, vol. 276, no. 2, 2019, pp.770-780, https://doi.org/10.1016/j.ejor.2019.01.040

[13] S. Rajab, and V. Sharma, “An interpretable neuro-fuzzy approach to stock price forecasting,” *Soft Computing*, vol. 23, 2019, pp.921-936, https://doi.org/10.1007/s00500-017-2800-7

[14] L.X. Wang, ”Fast training algorithms for deep convolutional fuzzy systems with application to stock index prediction,” *IEEE Transactions on fuzzy systems*, vol. 28, no. 7, 2019, pp.1301-1314, https://doi.org/10.1109/TFUZZ.2019.2930488

[15] P. Hajek, and J. Novotny, “Fuzzy rule-based prediction of gold prices using news affect,” *Expert Systems with Applications*, vol. 193, 2022, pp.116487, https://doi.org/10.1016/j.eswa.2021.116487

[16] A. Rakićević, P. Milošević, A. Poledica, I. Dragović, and B. Petrović, “Interpolative Boolean approach for fuzzy portfolio selection,” *Applying fuzzy logic for the digital economy and society*, 2019, pp. 23-46, https://doi.org/10.1007/978-3-030-03368-2_2

[17] K. Park, and H. Shin, "Stock price prediction based on a complex interrelation network of economic factors*," Engineering Applications of Artificial Intelligence*, vol. 26, no. 5-6, 2013, pp. 1550-1561, https://doi.org/10.1016/j.engappai.2013.01.009

[18] C. H. Cheng, L. Y. Wei, J. W. Liu, and T. L. Chen, "OWA-based ANFIS model for TAIEX forecasting," *Economic Modelling*, vol. 30, 2013, pp. 442-448, https://doi.org/10.1016/j.econmod.2012.09.047

[19] Y. Cao, "Aggregating multiple classification results using Choquet integral for financial distress early warning," *Expert Systems with Applications*, vol. 39, no. 2, 2012, pp. 1830-1836, https://doi.org/10.1016/j.eswa.2011.08.067

[20] F. Zhang, and Z. Liao, "Stock Price Forecasting Based on Multi-Input Hamacher T-Norm and ANFIS," in *Proceedings of the Ninth International Conference on Management Science and Engineering Management,* Berlin: Springer Heidelberg, 2015, pp. 55-66, https://doi.org/10.1007/978-3-662-47241-5_3

[21] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions. Encyclopedia of Mathematics and its Applications*. Cambridge: Cambridge University Press, 2009.

[22] M. Munar, M. Hudec, S. Massanet, E. Mináriková, and D. Ruiz-Aguilera, “On an Edge Detector Based on Ordinal Sums of Conjunctive and Disjunctive Aggregation Functions,” *In Proc. Conference of the European Society for Fuzzy Logic and Technology (Eusflat 2023)*, Palma, 2023, pp. 271-282, https://doi.org/10.1007/978-3-031-39965-7_23

[23] M. Hudec, E. Mináriková, R. Mesiar, A. Saranti, and A. Holzinger, “Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions”, *Knowledge-Based Systems*, vol. 220, id. 106916, 2021, https://doi.org/10.1016/j.knosys.2021.106916

[24] D. Radojevic, “Logical Aggregation Based on Interpolative Boolean Algebra,” *Mathware & Soft Computing*, vol. 15, 2008, pp. 125-141.

[25] D. Radojević, “[0,1]-valued logic: A natural generalization of Boolean logic,” *Yugoslav Journal of Operations Research*, vol. 10, no. 2, 2000, pp. 185-216.

[26] P. Milošević, B. Petrović, D. Radojević, and D. Kovačević, “A software tool for uncertainty modeling using Interpolative Boolean algebra,” *Knowledge-Based Systems*, vol. 62, 2014, pp. 1-10, https://doi.org/10.1016/j.knosys.2014.01.019

[27] I. Dragović, N. Turajlić, D. Pilčević, B. Petrović, and D. Radojević, “A Boolean consistent fuzzy inference system for diagnosing diseases and its application for determining peritonitis likelihood,” *Computational and Mathematical Methods in Medicine*, 2015.

[28] I. Dragović, N. Turajlić, D. Radojević, and B. Petrović, “Combining Boolean consistent fuzzy logic and AHP illustrated on the web service selection problem,” *International Journal of Computational Intelligence Systems*, vol. 7, Suppl 1, 2014, pp. 84-93, https://doi.org/10.1155/2015/147947

[29] P. Milošević, A. Poledica, A. Rakićević, V. Dobrić, B. Petrović, and D. Radojević, “IBA-based framework for modeling similarity,” *International Journal of Computational Intelligence Systems*, vol. 11, 2018, pp. 206-218, https://doi.org/10.2991/ijcis.11.1.16

[30] A. Rakićević, P. Milošević, B. Petrović, and D. Radojević, “DuPont Financial Ratio Analysis Using Logical Aggregation,” in *V. E. Balas, L. C. Jain & B. Kovačević (Eds.), Soft Computing Applications. Advances in Intelligent Systems and Computing,* vol. 357, Berlin: Springer, 2016, pp. 727-739, https://doi.org/10.1007/978-3-319-18416-6_57

[31] J. Kostić, M. Bakajac, P. Milošević, and A. Poledica, “Ranking of Banks Using Logical Aggregation,” in *N. Mladenović, G. Savić, M. Kuzmanović, D. Makajić-Nikolić & M. Stanojević (Eds.), Proceedings of the 11th Balkan Conference on Operational Research*, Belgrade: Faculty of Organizational Sciences, 2013, pp. 3-11.

[32] P. Milošević, I. Nešić, A. Poledica, D. Radojević, and B. Petrović, “Logic-based aggregation methods for ranking student applicants,” *Yugoslav Journal of Operational Research*, vol. 27, no. 4, 2017, pp. 461-477, https://doi.org/10.2298/YJOR161110007M

[33] M. Jeremić, A. Rakićević, and I. Dragović, “Interpolative Boolean algebra based multicriteria routing algorithm,” *Yugoslav Journal of Operations Research*, vol. 25, no. 3, 2015, pp. 397-412, https://doi.org/10.2298/YJOR140430029J

[34] A. Poledica, P. Milošević, I. Dragović, B. Petrović, and D. Radojević, “Modeling consensus using logic-based similarity measures*,” Soft Computing*, vol. 19, no. 11, 2015, pp. 3209-3219, https://doi.org/10.1007/s00500-014-1476-5

[35] A. Rakićević, V. Simeunović, B. Petrović, and S. Milić, "An automated system for stock market trading based on logical clustering," *Tehnički vjesnik*, vol. 25, no. 4, 2018, pp. 970–978, https://doi.org/10.17559/TV-20160318145514

[36] A. Rakićević, P. Milošević, and A. Poledica, "Logic-based system for evaluation of corporate financial performance." *InfoM Časopis za informacione tehnologije i multimedijalne sisteme,* vol. 51, 2014.

[37] M. Bramer, *Principles of data mining*. London: Springer – Verlag, 2020.

[38] M. Hudec, R. Mesiar, and E. Mináriková, “Applicability of Ordinal Sums of Conjunctive and Disjunctive Functions in Classification,”*In Proc. Conference of the European Society for Fuzzy Logic and Technology (Eusflat 2021)*, Bratislava, 2021, https://doi.org/10.2991/asum.k.210827.081

[39] O. Anđelić, P. Milošević, I. Dragović, & Z. Rakićević, “Logic-based Evaluation of production scheduling rules Using Interpolative Boolean Algebra”, *In Proc. 32nd International Conference on Information Systems Development (ISD 2024)*, Gdansk, 2024, accepted for publication.

[40] M. Hudec, E. Mináriková, and R. Mesiar, “Aggregation Functions in Flexible Classification by Ordinal Sums,” In *Proc. Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2022)*, Milan, 2023, Part I, pp. 372-383, https://doi.org/10.1007/978-3-031-08971-8_31

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 49–54
DOI: 10.15439/2024F3026
ISSN 2300-5963 ACSIS, Vol. 40

Applications of new q-rung orthopair fuzzy rough distance measures in pattern recognition and disease dignosis problems

Dragan Pamucar
Department of Operations Research and Statistics
Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia,
Department of Mechanics and Mathematics, Western Caspian University, Baku, Azerbaijan
Email: dpamucar@gmail.com

Arunodaya Raj Mishra
Department of Mathematics, Government College Raigaon, Satna, Madhya Pradesh, India,
Email: arunodaya87@outlook.com

Pratibha Rani
Department of Engineering Mathematics,
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Andhra Pradesh-522302, India,
Email: pratibha138@gmail.com

***Abstract*—As the combined version of rough sets (RSs) and q-rung orthopair fuzzy sets (q-ROFSs), the idea of q-rung orthopair fuzzy rough sets (q-ROFRSs) is more flexible to deal with inaccurate, uncertain and incomplete data. In this manuscript, we propose various q-rung orthopair fuzzy rough distance measures for computing the distance between q-ROFRSs. Some examples are discussed to exemplify the efficacy of developed q-ROFR-distance measures over existing ones. We further demonstrate its utility in pattern recognition and crop disease diagnosis problems. We also establish the superiority of developed distance measures over existing distance measures on q-ROFRSs in view of the structured linguistic variables.**

***Index Terms*—q-rung orthopair fuzzy rough set; distance measure; pattern recognition; medical diagnosis.**

I. Introduction

TO HANDLE the uncertain knowledge, Pawlak (1982) introduced a mathematical approach, named as rough set theory (RST), which has been widely implemented for various purposes (Sayed et al., 2024; Hosny et al., 2024). Dubois & Prade (1990) invented an idea of fuzzy rough set (FRS) to deal with granuality, incompleteness and uncertainty of knowledge in information measures. As an extended version, Zhang et al. (2012) pioneered the intuitionistic FRSs and implemented to the decision-making area. Further, Sun & Ma (2014) combined soft set and FRSs, and developed the notion of soft fuzzy rough sets (SFRSs). A q-rung orthopair fuzzy set (q-ROFS) (Yager, 2017) is an extended version of fuzzy set (FS) in which q^{th} powers sum of membership grade (MG) and nonmembership grade (NG) is ≤ 1, where $q \geq 1$. Yager (2017) pointed out that the space of acceptable orthopairs increases as value of q increases, therefore, the q-ROFS offers more choice to scholars in stating their confidence. The doctrine of q-ROFSs is more authoritative than the FS (Zadeh, 1965), intuitionistic fuzzy set (IFS) (Atanassov, 1986), Pythagorean fuzzy set (PFS) (Yager, 2014) and Fermatean fuzzy set (FFS) (Senapati & Yager, 2020) since all types of sets are contained in the space of q-ROFSs (Yager, 2017).

Khoshaim et al. (2021) integrated the notions of rough set and q-ROFS and gave a new idea namely q-rung orthopiar fuzzy rough set (q-ROFRS). A q-ROFRS offers benefits of q-ROFS as well as rough set. For the first time, Khoshaim et al. (2021) presented the basic aggregation operators (AOs) to unite the q-ROFRNs into a single q-ROFRN. Ashraf et al. (2021) proposed some AOs based on the combination of Einstein norms and q-ROFRNs. Further, a q-ROFR Einstein AOs-based EDAS approach has been presented for robotic agrifarming assessment problem. In a study, Liu et al. (2021) gave an axiomatic definition of distance measure for q-ROFRSs. Based on the distance measure, score function and AOs, they introduced a hybrid decision support system and its application in major infrastructure projects assessment. To assess the ship energy alternatives, Qahtan et al. (2023) presented a fuzzy decision with opinion score model under q-ROFRS environment. Moreover, the weights of evaluation criteria have been determined through fuzzy-weighted zero-inconsistency model. With the use of q-ROFRSs, Mishra et al. (2024) studied a combined multiple-criteria group decision-making (MCGDM) model consisting of symmetry point of criterion (SPC) tool for objective weight of indicators, ranking comparison (RANCOM) tool for subjective weight of indicators and multi-attribute multi-objective optimization based on ratio assessment (MULTIMOORA) approach to evaluate and rank the sustainable enterprise resource planning systems.

Distance measure is a vital mathematical way to compute degree of discrimination between two objects. This concept has widely been utilized to the medical dignosis, MCGDM and pattern recognition problems (Alrasheedi et al., 2023; Gogoi et al., 2023; Rani et al., 2024). Using distance measure, Wang et al. (2019) planned distance measure and FRS-based approach for reducing the number of attributes. They developed some iterative forms to determine fuzzy rough dependency and improtance degree of attributes and introdcued iterative assessemnt framework using variable distance parameter. Based on granular distance, An et al. (2021) studied a robust FRS approach and applied it in feature selection problem. Sahu et al. (2021) studied distance measure on picture fuzzy rough sets (PFRSs) and applied for career selection of students. Tiwari & Lohani (2023) studied a conflict distance measure between interval-valued IFSs and its application in MCGDM problem. Using weighted FRSs, Wang et al. [24] presented distance measure between the sample and other samples in a feature selection problem.

In the context of q-ROFRSs, Khoshaim et al. [12] gave the distance measure for computing the dissimilarity between considered criteria during the assessment of emergency MCGDM problem. Liu et al. [14] gave an idea of Euclidean q-ROFR-distance measure and discussed its application. Khan et al. [25] presented the hamming q-ROFR-distance measure and its utility in the evaluation of positive and negative ideal solutions. Some of these measures are unable to make the difference between q-ROFRSs. To overcome drawbacks of extant distances (Liu et al. [14]; Khoshaim et al. [12], Khan et al. [25]), this work introduces some distance measures for q-ROFRSs, which take into account the lower approximation and upper approximation MG and NG functions. Further, a utility of introduced distance measures are discussed on pattern recognition, crop disease diagnosis and medical diagnosis problems.

Other sections are presented in the following way. Section 2 presents the fundamental definitions related to q-ROFRSs. Section 3 introduces three distance measures for computing the degree of distance between q-ROFRSs. Section 4 applies the developed q-ROFR-distance measures to pattern recognition and crop disease diagnosis problem. Section 5 accomplishes the whole work.

II. Preliminaries

In the section, we first present basic notions related to q-ROFRSs.

Definition 2.1 [7]. Let $R=\{r_1,r_2,...,r_n\}$ be a fixed discourse set. A q-ROFS *G* on *R* is mathematically defined as

$$G=\left\{\left(r_i,\mu_G\left(r_i\right),\nu_G\left(r_i\right)\right)\middle| r_i\in R\right\},\tag{1}$$

wherein $\mu_G:R\to[0,1]$ and $\nu_G:R\to[0,1]$ denote MG and NG of an object $r_i\in R,$ respectively, with constraints $0\le\mu_G\left(r_i\right)\le 1,\quad 0\le\nu_G\left(r_i\right)\le 1,\quad 0\le\left(\mu_G\left(r_i\right)\right)^q+\left(\nu_G\left(r_i\right)\right)^q\le 1,$ $q\ge 1,\forall r_i\in R.$ For $r_i\in R,$ a hesitancy grade is defined as $\pi_G\left(r_i\right)=\sqrt[q]{1-\left(\mu_G\left(r_i\right)\right)^q-\left(\nu_G\left(r_i\right)\right)^q}.$

Definition 2.2 [13]. Consider *R* be a fixed discourse set and $\zeta\in R\times R$ be a crisp relation. Then

(i) ζ is reflexive if $(\wp,\wp)\in\zeta,\ \forall\wp\in R,$

(ii) ζ is symmetric if $\wp,\partial\in R$ and $(\wp,\partial)\in\zeta,$ then $(\partial,\wp)\in\zeta,$

(iii) ζ is transitive if $\wp,\partial,\ell\in R,$ $(\wp,\partial)\in\zeta$ and $(\partial,\ell)\in\zeta,$ then $(\wp,\ell)\in\zeta.$

Definition 2.3 [12]. Let $\zeta\in R\times R$ be defined as any arbitrary relation over *R*. Now, define a mapping $\zeta^*:R\to P(R)$ as

$$\zeta^*\left(\wp\right)=\left\{\partial\in R:\left(\wp,\partial\right)\in\zeta\right\},\ \text{for }\wp\in \boldsymbol{R},\tag{2}$$

where $\zeta^*\left(\wp\right)$ is an object's successor neighborhood $\wp$ with respect to $\zeta.$ Crisp approximation space (AS) is described as a pair $(R,\zeta).$ The lower and upper approximation of $\Im$ over $(R,\zeta),$ for each $\Im\in \boldsymbol{R}$ are given by

$$\underline{\zeta}\left(\Im\right)=\left\{\wp\in R:\zeta^*\left(\wp\right)\subseteq\Im\right\},\tag{3}$$

$$\bar{\zeta}\left(\Im\right)=\left\{\wp\in R:\zeta^*\left(\wp\right)\cap\Im\ne\phi\right\}.\tag{4}$$

The pair $\left(\underline{\zeta}\left(\Im\right),\bar{\zeta}\left(\Im\right)\right)$ is stated as a rough set (RS) and $\underline{\zeta}\left(\Im\right),\bar{\zeta}\left(\Im\right):P\left(R\right)\to P\left(R\right)$ are lower and upper approximation operators, respectively.

Definition 2.4 [12]. Let *R* be a fixed discourse set and $\zeta\in q-ROFS\left(R\times R\right)$ be any q-ROF-relation on *R*. Then

(i) ζ is reflexive if $\mu_\zeta\left(\wp,\wp\right)=1$ and $\nu_\zeta\left(\wp,\wp\right)=0,\forall\wp\in R,$

(ii) ζ is symmetric if $\left(\wp,\partial\right)\in R\times R,$ $\mu_\zeta\left(\wp,\partial\right)=\mu_\zeta\left(\partial,\wp\right)$ and $\nu_\zeta\left(\wp,\partial\right)=\nu_\zeta\left(\partial,\wp\right),$

(iii) ζ is transitive if $\left(\wp,\ell\right)\in R\times R,$ $\mu_\zeta\left(\wp,\ell\right)\ge\vee_{\partial\in R}\left[\mu_\zeta\left(\wp,\partial\right)\vee\mu_\zeta\left(\partial,\ell\right)\right]$ and $\nu_\zeta\left(\wp,\ell\right)=\wedge_{\partial\in R}\left[\nu_\zeta\left(\wp,\partial\right)\wedge\nu_\zeta\left(\partial,\ell\right)\right].$

Definition 2.5 [12]. Let *R* be a fixed discourse set and $\zeta\in q-ROFS\left(R\times R\right)$ be any non-empty q-rung orthopair fuzzy relation on *R*. The pair $\left(R,\zeta\right)$ is therefore stated as a q-rung orthopair fuzzy approximation space (q-ROFAS). The lower and upper approximation of $\Im$ over AS $\left(R,\zeta\right)$ are two q-ROFSs for any $\Im\subseteq q-ROFS\left(R\right),$ given by

$$\underline{\zeta}\left(\Im\right)=\left\{\left(\wp,\mu_{\underline{\zeta}}\left(\wp\right),\nu_{\underline{\zeta}}\left(\wp\right)\right):\wp\in \boldsymbol{R}\right\},\tag{5}$$

$$\bar{\zeta}\left(\Im\right)=\left\{\left(\wp,\mu_{\bar{\zeta}}\left(\wp\right),\nu_{\bar{\zeta}}\left(\wp\right)\right):\wp\in \boldsymbol{R}\right\},\tag{6}$$

where $\mu_{\underline{\zeta}(\Im)}\left(\wp\right)=\underset{\partial\in R}{\wedge}\left[\mu_\zeta\left(\wp,\partial\right)\wedge\mu_\Im\left(\partial\right)\right],$

$\nu_{\underline{\zeta}(\Im)}\left(\wp\right)=\underset{\partial\in R}{\vee}\left[\nu_\zeta\left(\wp,\partial\right)\vee\nu_\Im\left(\partial\right)\right],$

$\mu_{\bar{\zeta}(\Im)}\left(\wp\right)=\underset{\partial\in R}{\vee}\left[\mu_\zeta\left(\wp,\partial\right)\vee\mu_\Im\left(\partial\right)\right],$

$\nu_{\bar{\zeta}(\Im)}\left(\wp\right)=\underset{\partial\in R}{\wedge}\left[\nu_\zeta\left(\wp,\partial\right)\wedge\nu_\Im\left(\partial\right)\right],$

satisfying $0\le\left(\mu_{\underline{\zeta}(\Im)}\left(\wp\right)\right)^q+\left(\nu_{\underline{\zeta}(\Im)}\left(\wp\right)\right)^q\le 1$ and $0\le\left(\mu_{\bar{\zeta}(\Im)}\left(\wp\right)\right)^q+\left(\nu_{\bar{\zeta}(\Im)}\left(\wp\right)\right)^q\le 1,\ q\ge 1.$ As $\bar{\zeta}\left(\Im\right)$ and $\bar{\zeta}\left(\Im\right)$ are q-ROFSs, $\underline{\zeta}\left(\Im\right),\bar{\zeta}\left(\Im\right):q-ROFS\left(R\right)\to q-ROFS\left(R\right)$ are lower and upper approximation operators. Thus, a pair $\zeta\left(\Im\right)=\left(\underline{\zeta}\left(\Im\right),\bar{\zeta}\left(\Im\right)\right)$ $=\left\{\left(\wp,\left(\mu_{\underline{\zeta}(\Im)}\left(\wp\right),\nu_{\underline{\zeta}(\Im)}\left(\wp\right)\right),\left(\mu_{\bar{\zeta}(\Im)}\left(\wp\right),\nu_{\bar{\zeta}(\Im)}\left(\wp\right)\right)\right):\wp\in R\right\}$ is referred as q-ROFRS. For ease,

$$\zeta(\Im)=\left\{\left\langle \wp,\left(\mu_{\underline{\zeta}(\Im)}(\wp),\nu_{\underline{\zeta}(\Im)}(\wp)\right),\left(\mu_{\bar{\zeta}(\Im)}(\wp),\nu_{\bar{\zeta}(\Im)}(\wp)\right)\right\rangle:\wp\in R\right\}$$

is defined as $\zeta(\Im)=\left(\left(\underline{\mu},\underline{\nu}\right),\left(\bar{\mu},\bar{\nu}\right)\right)$ and named as q-rung orthopair fuzzy rough number (q-ROFRN) and its collection is acknowledged as q-ROFRS(R).

Definition 2.6 [14]. Let $\zeta(\Im_1)=\left(\underline{\zeta}(\Im_1),\bar{\zeta}(\Im_1)\right)=\left(\left(\underline{\mu}_1,\underline{\nu}_1\right),\left(\bar{\mu}_1,\bar{\nu}_1\right)\right)$ and $\zeta(\Im_2)=\left(\underline{\zeta}(\Im_2),\bar{\zeta}(\Im_2)\right)=\left(\left(\underline{\mu}_2,\underline{\nu}_2\right),\left(\bar{\mu}_2,\bar{\nu}_2\right)\right)$ be two q-ROFRNs and $\alpha>0$ be a real number. Then, Liu et al. [14] defined some operations on q-ROFRNs, given as

(i) $\left(\zeta(\Im_j)\right)^c=\left(\left(\bar{\zeta}(\Im_j)\right)^c\times\left(\underline{\zeta}(\Im_j)\right)^c\right)=\left(\left(\bar{\nu}_j,\bar{\mu}_j\right),\left(\underline{\nu}_j,\underline{\mu}_j\right)\right)$,

(ii) $\zeta(\Im_1)+\zeta(\Im_2)=\left(\underline{\zeta}(\Im_1)\oplus\underline{\zeta}(\Im_2),\bar{\zeta}(\Im_1)\oplus\bar{\zeta}(\Im_2)\right)$,

(iii) $\zeta(\Im_1)\times\zeta(\Im_2)=\left(\underline{\zeta}(\Im_1)\otimes\underline{\zeta}(\Im_2),\bar{\zeta}(\Im_1)\otimes\bar{\zeta}(\Im_2)\right)$,

(iv) $\alpha\zeta(\Im_j)=\left(\alpha\underline{\zeta}(\Im_j),\alpha\bar{\zeta}(\Im_j)\right),\ j=1,2,$

(v) $\left(\zeta(\Im_j)\right)^\alpha=\left(\left(\underline{\zeta}(\Im_j)\right)^\alpha,\left(\bar{\zeta}(\Im_j)\right)^\alpha\right),\ j=1,2,$

(vi) $\dfrac{\zeta(\Im_1)}{\zeta(\Im_2)}=\zeta(\Im_1)\times\left(\zeta(\Im_2)\right)^c$

$$=\left(\underline{\zeta}(\Im_1)\otimes\left(\bar{\zeta}(\Im_2)\right)^c,\bar{\zeta}(\Im_1)\otimes\left(\underline{\zeta}(\Im_2)\right)^c\right)$$

Definition 2.7 [14]. Let G, H and T be three q-ROFRSs. A q-ROFR distance measure $d:q-ROFRSs(R)\times q-ROFRSs(R)\to[0,1]$ is a real-valued mapping which satisfies the given axioms:

(i) $d(G,H)\geq 0,$

(ii) $d(G,H)=0$ iff $G=H,$

(iii) $d(G,H)=d(H,G),$

(iv) If $G\subseteq H\subseteq T,$ then $d(G,H)\leq d(G,T)$ and $d(H,T)\leq d(G,T).$

III. Proposed q-ROFR-Distance Measures

This section develops some distance measures to calculate the degree of dissimilarity between q-ROFRSs. Moreover, some examples are discussed to illustrate the usefulness of developed distance measures over extant distance measures (Khoshaim et al. [12], Liu et al. [14], Khan et al. [25]).

Let G and H be the q-ROFRSs. Then three q-ROFR-distance measures are given as

$$d_1(G,H)=\sqrt{\frac{1}{4n}\sum_{i=1}^{n}\left(\begin{array}{l}\left|\sqrt{\underline{\mu}_G^q(r_i)}-\sqrt{\underline{\mu}_H^q(r_i)}\right|+\left|\sqrt{\underline{\nu}_G^q(r_i)}-\sqrt{\underline{\nu}_H^q(r_i)}\right|\\+\left|\sqrt{\bar{\mu}_G^q(r_i)}-\sqrt{\bar{\mu}_H^q(r_i)}\right|+\left|\sqrt{\bar{\nu}_G^q(r_i)}-\sqrt{\bar{\nu}_H^q(r_i)}\right|\end{array}\right)}. \quad (7)$$

$$d_2(G,H)=\sqrt{\frac{1}{4n}\sum_{i=1}^{n}\left(\begin{array}{l}\left|\sqrt{\underline{\mu}_G^q(r_i)}-\sqrt{\underline{\mu}_H^q(r_i)}\right|+\left|\sqrt{\underline{\nu}_G^q(r_i)}-\sqrt{\underline{\nu}_H^q(r_i)}\right|\\+\left|\sqrt{\underline{\pi}_G^q(r_i)}-\sqrt{\underline{\pi}_H^q(r_i)}\right|+\left|\sqrt{\bar{\mu}_G^q(r_i)}-\sqrt{\bar{\mu}_H^q(r_i)}\right|\\+\left|\sqrt{\bar{\nu}_G^q(r_i)}-\sqrt{\bar{\nu}_H^q(r_i)}\right|+\left|\sqrt{\bar{\pi}_G^q(r_i)}-\sqrt{\bar{\pi}_H^q(r_i)}\right|\end{array}\right)}. \quad (8)$$

$$d_3(G,H)=\sqrt{\frac{3}{4n}\sum_{i=1}^{n}\left(\begin{array}{l}\dfrac{\left(\underline{\mu}_G^q(r_i)-\underline{\mu}_H^q(r_i)\right)^2}{\underline{\mu}_G^q(r_i)+\underline{\mu}_H^q(r_i)+2}+\dfrac{\left(\underline{\nu}_G^q(r_i)-\underline{\nu}_H^q(r_i)\right)^2}{\underline{\nu}_G^q(r_i)+\underline{\nu}_H^q(r_i)+2}\\+\dfrac{\left(\bar{\mu}_G^q(r_i)-\bar{\mu}_H^q(r_i)\right)^2}{\bar{\mu}_G^q(r_i)+\bar{\mu}_H^q(r_i)+2}+\dfrac{\left(\bar{\nu}_G^q(r_i)-\bar{\nu}_H^q(r_i)\right)^2}{\bar{\nu}_G^q(r_i)+\bar{\nu}_H^q(r_i)+2}\end{array}\right)}. \quad (9)$$

Property 3.1. For two q-ROFRSs G and H, $0\leq d_j(G,H)\leq 1,$ where j = 1, 2, 3.

Property 3.2. For two q-ROFRSs G and H, $d_j(G,H)=0$ iff $G=H,$ where j = 1, 2, 3.

Property 3.2. $d_j(G,H)=d_j(H,G),$ where G and H are two q-ROFRSs and j = 1, 2, 3.

Property 3.4. Let G, H and T be three q-ROFRSs. If $G\subseteq H\subseteq T,$ then $d_j(G,H)\leq d_j(G,T)$ and $d_j(H,T)\leq d_j(G,T),$ where j = 1, 2, 3.

Next, we present an example as Example 3.1 consisting of six different pairs of q-ROFRSs. Through this example, we highlight the drawbacks of extant q-ROFR-distance measures (Khoshaim et al. [12], Liu et al. [14], Khan et al. [25]). To this aim, we firstly recall the existing measures by Khoshaim et al. [12], Liu et al. [14], Khan et al. [25], given as follows:

Khoshaim et al.'s q-ROFR-DM [12]:

$$d_4(G,H)=\frac{1}{2}\left(\begin{array}{l}\left|\left(\underline{\mu}_G\right)^2-\left(\underline{\mu}_H\right)^2\right|^p+\left|\left(\underline{\nu}_G\right)^2-\left(\underline{\nu}_H\right)^2\right|^p\\+\left|\left(\bar{\mu}_G\right)^2-\left(\bar{\mu}_H\right)^2\right|^p+\left|\left(\bar{\nu}_G\right)^2-\left(\bar{\nu}_H\right)^2\right|^p\end{array}\right)^{1/p}. \quad (10)$$

Liu et al.'s q-ROFR-DM [14]:

$$d_5(G,H)=\frac{1}{4}\left(\begin{array}{l}\left(\left(\underline{\mu}_G^q\right)-\left(\underline{\mu}_H^q\right)\right)^2+\left(\left(\underline{\nu}_G^q\right)^2-\left(\underline{\nu}_H^q\right)\right)^2\\+\left(\left(\bar{\mu}_G^q\right)^2-\left(\bar{\mu}_H^q\right)\right)^2+\left(\left(\bar{\nu}_G^q\right)-\left(\bar{\nu}_H^q\right)\right)^2\\+\left(\left(\underline{\pi}_G^q\right)-\left(\underline{\pi}_H^q\right)\right)^2+\left(\left(\bar{\pi}_G^q\right)-\left(\bar{\pi}_H^q\right)\right)^2\end{array}\right)^{1/2}. \quad (11)$$

Khan et al.'s q-ROFR-DM [25]:

$$d_6(G,H)=\frac{1}{4}\left(\begin{array}{l}\left|\left(\underline{\mu}_G\right)-\left(\underline{\mu}_H\right)\right|+\left|\left(\underline{\nu}_G\right)-\left(\underline{\nu}_H\right)\right|\\+\left|\left(\bar{\mu}_G\right)-\left(\bar{\mu}_H\right)\right|+\left|\left(\bar{\nu}_G\right)-\left(\bar{\nu}_H\right)\right|\\+\left|\left(\underline{\pi}_G\right)-\left(\underline{\pi}_H\right)\right|+\left|\left(\bar{\pi}_G\right)-\left(\bar{\pi}_H\right)\right|)\end{array}\right). \quad (12)$$

Example 3.1. Consider the six different pairs of q-ROFRSs, which are given as *Set-1: {G =(0.26,0.36), (0.36,0.46), H =(0.36,0.26), (0.46,0.36)}, Set-2: {G = ((1,0), (1,0)), H = (0,1), (0,1)}, Set-3: {G = ((1,0), (1,0)), H =*

((0,0), (0,0))}, Set-4: {G = ((0.5,0.5), (0.5,0.5)), H = ((0,0), (0,0))}, Set-5: {G = (0.36,0.16), (0.46,0.26), H = (0.46,0.26), (0.56,0.36)} and *Set-6: {G = ((0.36,0.16), (0.46,0.26)), H = ((0.46,0.16), (0.56,0.26))}*. Next, we compute the degree of distance between these pairs of sets through the proposed and existing q-ROFR-distance measures (Khoshaim et al. [12], Liu et al. [14], Khan et al. [25]).

Table I presents required computational results of the q-ROFR-distance measures. On account of the obtained results, we draw the following conclusions:

- For two sets (Set-2 and Set-3), it can be observed that the q-ROFR-distance measure by Liu et al. [14] obtains the same value "1.803". It means that Liu et al.'s distance measure does not fulfil the postulate (*i*) of Definition 2.7.
- The distance measure by Khan et al. [25] is unable to differentiate two different pairs of sets (Set-1 and Set-6) as it obtains the same value "0.1". For three sets (Set-2, Set-3 and Set-4), Khan et al.'s [25] distance measure is unable to describe the difference between two different q-ROFRSs.
- The proposed q-ROFR-distance measure satisfies the axiomatic requirements of distance measure, given in Definition 2.7. For very similar but different q-ROFRSs, the proposed distance measure provides clear and rational data, which shows its effectiveness and rationality over extant measures (Khoshaim et al. [12], Liu et al. [14], Khan et al. [25]).

IV. Various Applications on q-ROFR Environment

To verify the rationality of introduced q-ROFR-distance measure given in Eq. (7)-Eq. (9), we present their utility in the field of pattern recognition and crop disease diagnosis.

A. Application to Pattern Recognition

Let us assume four known patterns R_1, R_2, R_3 and R_4, which have classifications S_1, S_2, S_3 and S_4, respectively. The known patterns are characterized by given q-ROFRSs in $\boldsymbol{R} = \{r_1, r_2\}$:

$$R_1 = \{(r_1,(0.9,0.5),(0.6,0.7)),(r_2,(0.8,0.6),(0.5,0.7))\}, \quad (13)$$

$$R_2 = \{(r_1,(0.4,0.7),(0.5,0.6)),(r_2,(0.7,0.5),(0.4,0.6))\}, \quad (14)$$

$$R_3 = \{(r_1,(0.3,0.6),(0.6,0.5)),(r_2,(0.5,0.7),(0.7,0.3))\}, \quad (15)$$

$$R_4 = \{(r_1,(0.5,0.8),(0.2,0.6)),(r_2,(0.7,0.5),(0.6,0.6))\} \quad (16)$$

Given an unknown pattern is defined as

$$T = \{(r_1,(0.6,0.5),(0.5,0.5)),(r_2,(0.3,0.7),(0.5,0.6))\}. \quad (17)$$

The objective is to identify that which class does the unknown pattern's T belong to. In accordance with the doctrine of minimum distance measure between q-ROFRSs, the procedure of assigning T to S_{k^*} is defined as

$$k^* = \arg\min_k \{d_\alpha(R_k, T)\}, \alpha = 1,2,3. \quad (18)$$

Table II shows computational outcomes of q-ROFR-distance measures. Based on the obtained results, it has been observed that the pattern T is being classified to S_3 as it has least degree of distance on known pattern R_k and unknown pattern T.

Table III.

Degree of distance measure $d_\alpha(R_k,T), k \in \{1,2,3,4\}$

Pattern	R_1	R_2	R_3	R_4
T	0.437	0.699	**0.393**	0.449

B. Application to Crop Disease Diagnosis

Here, we apply the proposed q-ROFR-distance measures for diagnosing the crop disease in an Indian region. This study consists of sets of crops, diseases and factors, which are represented by $\boldsymbol{P}$ = {Wheat, Rice, Carrot, Onion red}, $\boldsymbol{H}$ = {Viroid, Fungal, Nematodes, Bacterial, Phytoplasmal} and $\boldsymbol{V}$ = {Temperature, Soil moisture, Insect, pH value, Humidity}, respectively. Table III displays related features of considered diseases and Table IV presents the symptoms features of given crops in terms of q-ROFRNs.

In order to do a proper diagnosis, we compute for each crop $p_i \in P$, where $i \in \{1,2,3,4\}$, the degree of q-ROFR-distance measure $d_\alpha(f(p_i), h_k)$ on crop symptoms and set of symptoms that are feature for each diagnosis $h_k \in H$ with $k \in \{1,2,3,4,5\}$. Similar to Eq. (18), the proper diagnosis h_{k^*} for ith crop is determined as follows:

Table I.

Comparative results by different q-ROFR-distance measures

Sets	Distance measures	d_1	d_2	d_3	d_4	d_5	d_6
Set-1	G = ((0.26,0.36), (0.36,0.46)) H = ((0.36,0.26), (0.46,0.36))	0.3	0.333	0.1	0.092	0.079	0.1
Set-2	G = ((1,0), (1,0)) H = ((0,1), (0,1))	1.0	1.307	1.0	1.260	**1.803**	**1.0**
Set-3	G = ((1,0), (1,0)) H = ((0,0), (0,0))	0.707	0.924	0.707	1.0	**1.803**	**1.0**
Set-4	G = ((0.5,0.5), (0.5,0.5)) H = ((0,0), (0,0))	0.595	0.648	0.399	0.315	0.419	**1.0**
Set-5	G = ((0.36,0.16), (0.46,0.26)) H = ((0.46,0.26), (0.56,0.36))	0.298	0.335	0.099	0.099	0.153	0.2
Set-6	G = ((0.36,0.16), (0.46,0.26)) H = ((0.46,0.16), (0.56,0.26))	0.225	0.262	0.077	0.093	0.124	0.1

$$k^{*} = \arg\min_{k}\left\{d_{\alpha}\left(f(p_i), h_k\right)\right\}, \alpha = 1,2,3. \quad (19)$$

We allocate to the i^{th} crop the diagnosis whose symptoms have lowest degree of distance measure from crop symptoms. Table V shows the required computational results of crop disease diagnosis.

It can be observed from Table V that "Wheat" is most affected by Bacterial, "Rice" is most affected by Fungal disease, "Carrot" is affected by Nematodes and "Onion red" is most affected by Fungal disease.

C. Applications, gaps and future directions of q-ROFRSs

In the thematic assessment, numerous emerging ideas have been surfaced, contributed to the developing landscape of q-ROFRSs literature. These concepts incorporate various disciplines namely correlation coefficient, similarity measure assessment on q-ROFRSs, calculations relating q-ROFRNs, the generalization of q-ROFRS with 2-Tuple linguistic approach and the application of q-ROFRSs in healthcare, digital technology mainly in the evaluation of challenges and barriers. Briefing understudied regions in q-ROFRSs literature, Table VI is presented by the authors, helps as a concise reference for future direction.

V. Conclusion

In the paper, we have introduced three new distance measures for q-ROFRSs with their enviable properties in the context of q-ROFRSs. We have discussed the consistency and efficacy of the developed distance measures through a comparative example consisting of six different pairs of q-ROFRSs. In addition, we have highlighted the counter-intuitive cases of Khoshaim et al. [12], Liu et al. [14] and Khan et al. [25] q-ROFR-distance measures. It has been obtained that in some circumstances, developed q-ROFR-distance measures perform better than some of the existent distance measures for some sets of q-ROFRSs. Further, the developed q-ROFR-distance measures has been implemented to the pattern recognition and crop disease diagnosis problems. In future, the developed q-ROFR-distance measures can be used to solve texture extraction and medical diagnosis problems. In addition, the proposed measures can be extended under different fuzzy environments such as interval-valued q-ROFRSs, linear Diophantine fuzzy rough sets, hypersoft rough sets and others.

Table IVII.
Symptoms-diseases q-rung orthopair fuzzy rough relation

Symptoms	Viroid	Fungal	Nematodes	Bactarial	Phytoplasmal
Temperature	((0.6, 0.7), (0.2, 0.5))	((0.8, 0.4), (0.4, 0.6))	((0.9, 0.2), (0.4, 0.3))	((0.6, 0.5), (0.5, 0.2))	((0.8, 0.5), (0.3, 0.6))
Soil Moisture	((0.9, 0.4), (0.6, 0.3))	((0.7, 0.6), (0.4, 0.2))	((0.5, 0.7), (0.5, 0.1))	((0.8, 0.4), (0.5, 0.2))	((0.7, 0.6), (0.5, 0.1))
Insect	((0.7, 0.5), (0.4, 0.2))	((0.8, 0.3), (0.5, 0.2))	((0.6, 0.5), (0.5, 0.3))	((0.4, 0.9), (0.4, 0.3))	((0.9, 0.4), (0.5, 0.3))
pH value	((0.9, 0.3), (0.5, 0.3))	((0.6, 0.5), (0.3, 0.5))	((0.8, 0.4), (0.5, 0.2))	((0.7, 0.6), (0.5, 0.2))	((0.6, 0.8), (0.5, 0.4))
Humidity	((0.3, 0.9), (0.2, 0.7))	((0.8, 0.4), (0.5, 0.3))	((0.7, 0.6), (0.6, 0.2))	((0.6, 0.5), (0.2, 0.6))	((0.4, 0.7), (0.6, 0.2))

Table IIIV.
Crops-Symptoms q-rung orthopair fuzzy rough relation

Crops	Temperature	Soil moisture	Insect	pH value	Humidity
Wheat	((0.3, 0.6), (0.2, 0.8))	((0.4, 0.8), (0.5, 0.7))	((0.5, 0.9), (0.3, 0.4))	((0.4, 0.9), (0.4, 0.7))	((0.6, 0.5), (0.5, 0.4)
Rice	((0.4, 0.5), (0.5, 0.9))	((0.2, 0.6), (0.5, 0.4))	((0.7, 0.4), (0.2, 0.8))	((0.5, 0.7), (0.3, 0.5))	((0.4, 0.4), (0.5, 0.6))
Carrot	((0.4, 0.6), (0.3, 0.5))	((0.3, 0.6), (0.2, 0.5))	((0.5, 0.6), (0.7, 0.4))	((0.5, 0.4), (0.6, 0.3))	((0.4, 0.6), (0.3, 0.4))
Onion red	((0.5, 0.7), (0.5, 0.4))	((0.6, 0.4), (0.5, 0.5))	((0.7, 0.2), (0.6, 0.5))	((0.4, 0.6), (0.5, 0.3))	((0.8, 0.4), (0.5, 0.2))

Table V.
Degree of distance measure on each crop symptoms and considered set of possible diagnoses

Crops	Viroid	Fungal	Nematodes	Bacterial	Phytoplasmal
Wheat	0.567	0.504	0.513	**0.47**	0.497
Rice	0.53	**0.441**	0.514	0.493	0.454
Carrot	0.454	0.47	**0.44**	0.455	0.449
Onion red	0.471	**0.371**	0.429	0.429	0.436

TABLE VI.
OUTLINE OF UNDERSTUDIED REGIONS IN Q-ROFRSS LITERATURE OFFERING FOUNDATION FOR FUTURE RESEARCH DIRECTIONS

q-ROFRSs dimensions	Understudied Regions
Aggregation operators (AOs)	a) New AOs defined on q-ROFRS or its generalizations, b) Integration of some extant AOs for finding more powerful and flexible AOs.
q-ROFRSs linguistic rating	Linguistic rating mapping to deal linguistic information of q-ROFRSs generalizations
MCDM methods	a) Proposing hybrid MCDM approaches to evade the drawbacks of single MCDM models, b) Integrating mathematical model or optimization with q-ROFRS MCDM models.
Application regions	a) Healthcare, b) Agriculture/agro-farming, c) Finance/Economy, d) Manufacturing, e) Technology innovation, f) Emergency decision-making, and others.
Application objectives	a) Advancing and designing urban areas, b) Manufacturing robot design and evaluation, c) Digital technology evaluation, and others.

REFERENCES

[1] Z. A. Pawlak, “Rough sets”, *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[2] G. I. Sayed, E. I. A. El-Latif, A. E. Hassanien, V. Snasel, “Optimized long short-term memory with rough set for sustainable forecasting renewable energy generation”, *Energy Reports*, vol. 11, pp. 6208-6222, 2024.

[3] R. A. Hosny, R. Abu-Gdairi, M. K. El-Bably, “Enhancing Dengue fever diagnosis with generalized rough sets: Utilizing initial-neighbourhoods and ideals”, *Alexandria Engineering Journal*, vol. 94, pp. 68-79, 2024.

[4] D. Dubois, H. Prade, “Rough fuzzy sets and fuzzy rough sets”, *International Journal of General Systems*, vol. 17, no. 2–3, pp. 191–209, 1990.

[5] X. Zhang, B. Zhou, P. Li, “A general frame for intuitionistic fuzzy rough sets”, *Information Sciences*, vol. 216, pp. 34–49, 2012.

[6] B. Sun, W. Ma, “Soft fuzzy rough sets and its application in decision making”, *Artificial Intelligence Review*, vol. 41, no. 1, pp. 67–80, 2014.

[7] R. R. Yager, “Generalized orthopair fuzzy sets”, *IEEE Transactions on Fuzzy Systems*, vol. 25, pp. 1222-1230, 2017.

[8] L. A. Zadeh, “Fuzzy sets”, *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[9] K. T. Atanassov, “Intuitionistic fuzzy sets”, *Fuzzy Sets and Systems*, vol. 20, no. 1, pp. 87–96, 1986.

[10] R. R. Yager, “Pythagorean membership grades in multicriteria decision making”, *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 4, pp. 958–965, 2014.

[11] T. Senapati, R. R. Yager, “Fermatean fuzzy sets”, *J Ambient Intell Human Comput.*, vol. 11, pp. 663–674, 2020. https://doi.org/10.1007/s12652-019-01377-0.

[12] A. B. Khoshaim, S. Abdullah, S. Ashraf, M. Naeem, “Emergency decision-making based on q-rung orthopair fuzzy rough aggregation information. Computers”, *Materials & Continua*, vol. 69, no. 3, pp. 4077-4094, 2021.

[13] S. Ashraf, N. Rehman, A. Hussain, A. Al-Salman, A. H. Gumaei, “q-Rung orthopair fuzzy rough einstein aggregation information-based EDAS method: applications in robotic agrifarming”, *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 5520264), pp. 01-27, 2021.

[14] F. Liu, T. Li, J. Wu, T. Liu, „Modification of the BWM and MABAC method for MAGDM based on q-rung orthopair fuzzy rough numbers”, *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 9, pp. 2693–2715, 2021.

[15] S. Qahtan, H. A. Alsattar, A. A., Zaidan, M. Deveci, D. Pamucar, D. Delen, “Performance assessment of sustainable transportation in the shipping industry using a q-rung orthopair fuzzy rough sets-based decision-making methodology”, *Expert Systems with Applications*, vol. 223, 119958, 2023, https://doi.org/10.1016/j.eswa.2023.119958.

[16] A. R. Mishra, P. Rani, D. Pamucar, V. Simic, “Evaluation and prioritization of sustainable enterprise resource planning in SMEs using q-rung orthopair fuzzy rough set-based decision support model”, *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 5, pp. 3260-3273, 2024.

[17] A. F. Alrasheedi, A. R. Mishra, P. Rani, E. K. Zavadskas, F. Cavallaro, “Multicriteria group decision making approach based on an improved distance measure, the SWARA method and the WASPAS method”, *Granular Computing*, vol. 8, pp. 1867–1885, 2023.

[18] S. Gogoi, B. Gohain, R. Chutia, “Distance measures on intuitionistic fuzzy sets based on cross-information dissimilarity and their diverse applications”, *Artificial Intelligence Review*, vol. 56, pp. 3471–3514, 2023.

[19] P. Rani, A. R. Mishra, F. Cavallaro, A. F. Alrasheedi, “Location selection for offshore wind power station using interval-valued intuitionistic fuzzy distance measure-RANCOM-WISP method”, *Scientific Reports*, vol. 14, no. 4706, 2024, https://doi.org/10.1038/s41598-024-54929-6.

[20] C. Wang, Y. Huang, M. Shao, X. Fan, “Fuzzy rough set-based attribute reduction using distance measures”, *Knowledge-Based Systems*, vol. 164, pp. 205-212, 2019.

[21] S. An, Q. Hu, C. Wang, “Probability granular distance-based fuzzy rough set model”, *Applied Soft Computing*, vol. 102, no. 107064, 2021, https://doi.org/10.1016/j.asoc.2020.107064.

[22] R. Sahu, S. R. Dash, S. Das, “Career selection of students using hybridized distance measure based on picture fuzzy set and rough set theory”, *Decision Making: Applications in Management and Engineering*, vol. 4, no. 1, pp. 104–126, 2021.

[23] A. Tiwari, Q. M. D. Lohani, “Interval-valued intuitionistic fuzzy rough set system over a novel conflict distance measure with application to decision-making”, *MethodsX*, vol. 10, no. 102012, 2023, https://doi.org/10.1016/j.mex.2023.102012.

[24] C. Wang, C. Wang, Y. Qian, Q. Leng, “Feature Selection Based on Weighted Fuzzy Rough Sets”, *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 7, pp. 4027-4037, 2024.

[25] S. Khan, M. Khan, M. S. A. Khan, S. Abdullah, F. Khan, “A novel approach toward q-rung orthopair fuzzy rough Dombi aggregation operators and their application to decision-making problems”, *IEEE Access*, vol. 11, pp. 35770-35783, 2023.

DOI: 10.15439/2024F905
ISSN 2300-5963 ACSIS, Vol. 40

Optimal Charge Scheduling and Navigation for Multiple EV Using Deep Reinforcement Learning and Whale Optimization

Aby N Raj
Department of EEE
Bharath Institute of Higher Education and Research
Chennai, India
abynraj@tataelxsi.co.in

Dr. K. Sakthivel
Department of EEE
Bharath Institute of Higher Education and Research
Chennai, India
sakthivelk.eee@bharathuniv.ac.in

***Abstract*—The demand for Electric Vehicles (EVs) is increasing exponentially in recent times because of its ability to minimize energy savings and carbon emission. However, the charging process and charging option increases the challenges for EV adoption. With the growing adaptability to EVs, the need for addressing the challenges related to limited range and the availability of charging infrastructure becomes crucial. This paper presents an optimized deep learning-based charge scheduling approach in EVs for intelligent transport systems. The study leverages the Deep Reinforcement Learning (DRL) for making real-time decisions. The DRL model is trained using various features such as Battery critical percentage (SOC), time slots, nearest charge station, and availability of charging station. The features are optimized using a nature inspired Whale Optimization Algorithm (WOA), which helps in obtaining optimal charge scheduling. The proposed approach is experimentally evaluated in terms of reducing the tow counts in the selected region. Results from the experimental analysis validate the efficacy of the proposed approach in achieving optimal charge scheduling and navigation for EVs which also improve energy efficiency and reduce charging costs and charging time.**

***Keywords*—Electric Vehicles, Charge Scheduling, Intelligent Transport System, Particle Swarm Optimization, Whale Optimization Algorithm, Deep Reinforcement Learning.**

I. Introduction

THE ADOPTION of Electric Vehicles (EVs) is growing immensely for achieving environmentally friendly transportation by reducing the use of fossil fuel thereby contributing to the zero emission of toxic greenhouse gas [1]. Despite the advantages, the adaptability of EVs is restricted due to the challenges associated with the charging process such as limited charging stations, availability of charging slots, dynamic charging patterns, and varying load demand [2]. These factors contribute to the increase in the peak demand, grid overload condition, voltage fluctuation etc., which affect the performance efficiency of EV echo system [3] [4]. This inefficiency results in traffic congestion near the charging stations that impacts traffic planning, and traffic order [5]. This problem can be addressed by the intelligent transportation system wherein the details about the availability of charging station and time slot can be obtained beforehand. Several studies have focused on developing optimized charge scheduling mechanisms [6] [7] [8]. These techniques intend to avoid overload on the charging station during peak hours. However, most of these techniques calculate the charging time when EVs are either parked at home or parking lots. In practical scenarios, EV users require charging stations while driving (both shorter and longer durations) due to the limited capacity of EV batteries [9]. In this context, in real time, an efficient navigation mechanism is required to suggest optimal route and availability of charging stations for charging EVs considering different aspects such as distance, charging rate, and waiting time [10][11][12]. On this basis, a novel optimized charging scheduling framework for supporting multiple EVs is designed, developed and evaluated in this work. The prominent aspects of this manuscript are as follows:

- A feature extraction technique is employed to extract the relevant features related to the charge scheduling and a DRL based feature selector known as (DRLFS) is implemented for finding suitable features to improve the charge scheduling process.
- The selected features are optimized using a nature inspired WOA to obtain an efficient optimal charge scheduling in EVs.
- This framework efficiency is evaluated checking the number of tow count i.e., number of vehicles dead on the road before reaching its destination.

The rest of the paper is organized as follows. Section II discusses charge scheduling techniques presented in other works. Section III discusses the proposed charge scheduling framework in EV charging navigation for intelligent transport systems. Section IV briefs the simulation results and Section V concludes the paper with prominent future study observations.

II. Literature Review

A. Utilizing Machine Learning

Several machine learning techniques have been proposed for improving the overall efficiency of EV Charging echo system. The work presented in this paper [13] compares and evaluated the effectiveness of various machine learning (ML) approaches for EV charging considering conventional charging, rapid charging, and vehicle-to-grid (V2G) technologies. The work presented in this paper [14] provides the insights on the usage of machine learning models for determining the optimal location of EV charging stations (EVCS) and its infrastructure. Although from the mentioned references it is evident about the range of ML algorithms used in the EV charging navigation area, there summarize the need for looking into more advanced ML techniques, which is Deep Learning.

B. Utilizing Deep Reinforcement Learning

A subset of machine learning (ML), Deep Reinforcement learning (DRL) has attracted a lot of attention in this subject. Table 1 provides a summary of the areas of EV Charging echo system in which DRL is used.

TABLE I. DRL USED STUDIES

Methodology & Focus	Reference
Deep reinforcement learning simulator to validate the feasibility of learning algorithms to be deployed	[15]
Deep-learning-based EV arrival rates calculated according to the historical data	[16]
Deep reinforcement learning for optimal scheduling of charging station according to the random behaviour characteristics of the EV charging arrival and departure times	[17]
Hybrid deep learning mechanism to assure safe and dependable charging operations that prevent the battery from being overcharged or discharged	[18]
Deep reinforcement learning based EV cluster scheduling strategy considering real-time electricity prices	[19]
Deep reinforcement learning to minimize the total charging time of EVs and maximal reduction in the origin-destination distance	[20]
Deep reinforcement learning based optimal charging strategy considering traffic conditions, user's behaviour, and the pricing	[21]
Deep reinforcement learning based evaluation of model-free coordination of EV	[22]
Deep reinforcement learning for EV charging navigation for single EV	[23]

C. Reseach Gaps

This research identifies some of the prominent research gaps from the existing works, which are outlined as follows:

In most of the existing techniques, the charging navigation estimations are not real time, which is required for making intelligent decisions for charge scheduling and navigation.

The most widely used deep learning used EV scheduling techniques has not ben verified for multiple EVs, which is required for measuring the efficiency of the EV charging echo system.

So, there is a need to investigate methods to enhance the adaptability of deep learning models to dynamic and uncertain environments, such as incorporating uncertainty estimation or developing RL techniques that handle real-time changes effectively.

III. PROPOSED RESEARCH METHODOLOGY

This research aims to develop an efficient DRL-based technique for charge scheduling and navigation in EVs to enhance overall power management and efficiency. In general reinforcement learning can learn from the actions and its continuous interactions from the external environment. This enables the reinforcement learning models to make fast decisions in a dynamic environment. In this research, the EV model itself is considered as the environment and the model learns the parameters of the EV such as the SoC of the battery, time slots, charging pattern and availability of charging station. The proposed charge scheduling strategy is designed to minimize the charging time, mitigate traffic congestion and improve the power management. In practical conditions, the driving cycle is more complex and DRL based control strategies help in finding the best solution for the complex problems aiming to achieve optimal charge scheduling.

The main goal of charge scheduling is to minimize the number of tow counts i.e., number of vehicles dead on the road before reaching its destination and also the power utilization of vehicles. Reduced number of tow counts is essential for achieving an optimal charge scheduling, which is also the aim of this research. In order to address the research gaps, the proposed DRL-WOA is designed which modifies the charging and scheduling process in real-time. This is achieved by optimizing model parameters which helps the system to dynamic environmental factors in an effective manner. In addition, the work also incorporates uncertainty estimation technique into the DRL framework to enhance the adaptability to uncertain environments. This integrated DRL-WOA approach enables continuous learning and adjustment based on real-time data, addressing the challenges of EV charging, scheduling, and navigation in dynamic settings.

The proposed work flow involved is shown in Fig. 1 and the same are explained in below subsections.

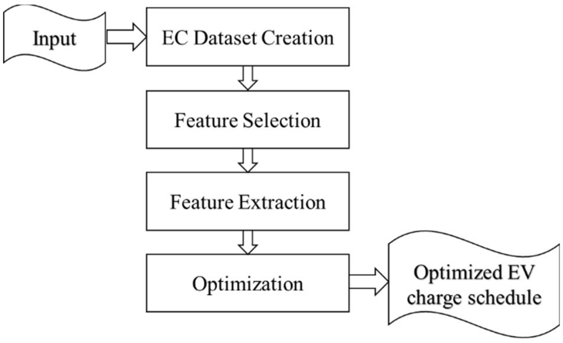


Fig. 1. Proposed workflow

A. EV Dataset Creation

In this study the data set is derived with random SOC, and at random coordinated within the city boundaries. To enable this, the following parameters are considered. Number of Vehicles, Number of Charging Stations, X Dimension of city (km), and Y Dimension of city. Based on these parameters, the EV Commute and Charging Station block set shall generate the dataset with the following data units such as Vehicle Number, Available battery Power, Needed Charging Time, Starting X-Position, Starting Y-Position, Destination X-Position and Destination Y-Position.

B. Feature Selction

In this step, feature selection is performed to select essential and relevant features from the dataset. It is important to perform feature selection in order to avoid the selection of irrelevant features. In this work, four important features are selected namely; (i) Time-Related Probabilities (timprob) (ii) Charging Probabilities (chprob) (iii) Battery Critical Level (batt_crit) and (iv) Charging Stations Locations as these are crucial for charge scheduling.

C. Feature Extraction

The DRL based Feature Selector (DRLFS) is implemented to identify an optimal subset containing relevant features. The architecture of the DRLFS is shown in Fig. 2. The DRLFS employs a reinforcement learning mechanism with an agent and an interactive environment. The agent in the DRL environment employs a learning-based policy for identifying and selecting the attributes for performing a specific task. Here, the features are selected based on actions and computes the rewards for every action. For effectively searching the feature subset the DRLGS employs a random policy along with two other search mechanism which helps in controlling the balance between exploration and exploitation. Entire searching process for feature selection is segmented into multiple smaller segments wherein each segment incorporates a continuous and sequential process. In this process, the features are selected and are grouped into a separate subset. This process is continued till the DRLFS reaches termination. In addition, several iterations are performed in each segmentation and during each iterative stage, the agent in the DRL environment identifies one feature and determines its actions, provides reward and stores the action for future learning. In particular, after every iteration the DRLFS generates a decision based on the selected features and acts accordingly.

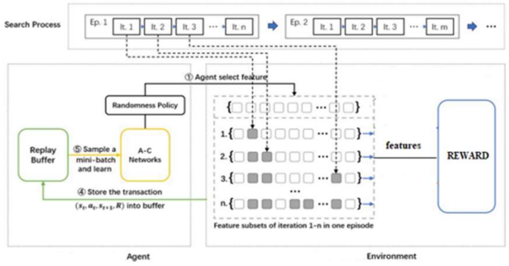


Fig. 2. Feature Selector

The proposed learning mechanism can be summarized as follows: Considering a search space ‘S’ wherein the feature set ‘F’ are randomly selected. In this work, the features are selected by computing a binary classification problem and hence is represented as |S| = 2|F|. Any change in the action of the DRL environment filters out certain important attributes and affect the correlation between them. In order to avoid this, each feature is reconfigured in the action space which in turn results in the formation of multiple discrete actions. However, it is a challenging and complex task to handle such a large action space since it affects the performance of the DRL in terms of making decisions about the feature selection process. In this context, this research considers a fine-grained action space in the continuous form as output. If the feature subset consists of only selected features then for every iteration, the DRL selects the features which belongs to the subset using a deterministic policy. After reaching termination, the feature subset obtained at the last iteration is considered as the final subset (Fe).

Let Error (F) and Error (Fe) represent the testing error for features F and final subset Fe, respectively. The objective is to obtain a minimum error and hence the reward function is computed as follows:

$$\mathrm{R} = \mathrm{Error}\ (\mathrm{F}) - \mathrm{Error}\ (\mathrm{Fe}) \tag{1}$$

For the obtained error function, the maximum reward can be obtained by computing the optimal deterministic policy μ:

$$\mu * = \arg\max \mu\ \mathrm{Error(F)} - \mathrm{Error(Fe)} \tag{2}$$

For generating an optimal policy for feature selection, the DRLFS uses a Deep Deterministic Policy Gradient (DDPG) which is an off-policy actor-critic DRL algorithm. The DDPG uses a train and error method along with a stable, fine-grained action for training the DRL to find an optimal feature subset.

Pseudocode of the DRLFS algorithm

```
Initialization:
Randomly initialize critic network Q (s, α|θQ) and actor μ(s|θμ)
Initialize target network Q’ and μ’
Initialize the Replay Buffer R
for episode = 0 to N do
    Initialize the feature subset F as an empty set
Initialize the state s0 as a zero vector with a length d
while st is not se do
According to the current actor policy, select the action at = μ (s|θμ)
    Randomize the action at by truncated normal distribution and decaying
    Transform st to st+1 and add newly selected feature into F by at
    Test the generalization error on the DRL algorithm and calculate the reward rt
    Store the transaction (st, at, rt, st+1) into R
    Sample a random minibatch from R
    Update critic Q(s, α|θQ) by minimizing the loss
    Update the actor policy μ(s|θμ) using the sampled policy gradient
    Update the target networks Q’ and μ’
    if ≠ features = limit then
    Set st = se
    end if
end while
end for
```

In the DRLFS algorithm, each iterative step follows a criteria to terminate the current step and begin with another one. In this process, the selected features in the subset are defined as the search depth. In this research a fixed depth search (FDS) is considered by selecting a limited number of features. In the FDS, the searching mechanism in each iterative step terminates at a fixed depth. Such a policy provides a stable search evenly for every depth, which means that all different depths are explored for the same number of times. This helps in exploring more depth space and selecting more features. Further, the selected features are optimized to optimize the charge scheduling process in EVs, which is discussed in the next section.

D. Optimization

A Whale Optimization Algorithm (WOA) is used for further optimization of charge scheduling. The WOA is a metaheuristic technique which mimics the hunting behavior of humpback whales. The algorithm is inspired by the bubble-net hunting strategy. The humpback whales prefer to hunt schools of krill or small fishes that are close to the surface. This is done by forming bubbles across a circular path with ‘upward-spirals’ and ‘double-loops’. Mathematically, the spiral bubble-net feeding maneuver is modelled in order to perform optimization.

WOA is mainly known for the hunting behaviour with the best search agent to chase the prey. The algorithm employs a spiral to simulate bubble-net attacking mechanisms of humpback whales. The stages involved in the algorithm are as follows:

Encircling Prey:

The algorithm assumes that the current best solution is close to target prey. Based on the obtained solutions, the position is further updated as shown in below given equations:

$$\vec{D}=|\vec{C}*\vec{X}\_best\ (t)-\vec{X}(t)| \tag{3}$$

$$\vec{X}(t+1)= \vec{X}\_best\ (t)-\vec{A}*\vec{D} \tag{4}$$

Where t defines the current iteration, A and C are the coefficient vectors, Xbest is the position vector of the best solution, and X indicates the position vector of the whales.

$$\vec{A}=2\vec{a}\,\vec{r}\_1-\vec{a} \quad (5)$$

$$\vec{C}=2\vec{r}\_2 \quad (6)$$

Where, vec{r1} , vec{r2} are random vectors in [0, 1].

Exploitation Stage:

This stage is also known as attacking mechanism of the Bubble net. This mechanism is mathematically modeled and involves two prominent mechanisms:

(i) Shrinking encircling mechanism: This behavior is achieved by decreasing the value of vec{a}, where a is decreased from 2 to 0 over the course of iterations.

(ii) Spiral updating position: In this process, the spiral position is updated with a random number that lies between -1 to 1.

Search for prey:

Humpback whales search randomly according to the position of each other

$$\vec{D}=|\vec{C}*\vec{X}\_rand\,(t)-\vec{X}(t)| \quad (7)$$

$$\vec{X}(t+1)=\vec{X}\_rand\,(t)-\vec{A}*\vec{D} \quad (8)$$

The pseudocode of the WOA is given below:

```
Initialization

Initialize the whale population Xi (i = 1, 2, …., n)
Calculate the fitness value of each search agent
      Xbest is the best search agent
while (t < maximum number of iterations)
         for each search agent
      update α, A, C, l and p
      if (p < 0.5):
      Update current agent using equation 3
else:
      Select the random agent Xrand
      Update current agent using equation 7
else:
      Update search agent using equation 4
end for
Check if the search agent crosses the search space
Calculate the fitness value of each search agent
Update Xbest if there is a better solution
t = t+1
end while
Return Xbest
End
```

The WOA leverages the selected features to generate an optimal charge scheduling plan, prioritizing charging sessions at times and locations where the distances to charging stations are minimal. This integrated approach ensures that EVs are charged in a manner that minimizes both time and distance, optimizing overall charging efficiency and user convenience. By optimizing the features, the optimal charge scheduling is generated, with a reduced number of tow counts, charging power, cost and time. The performance evaluation of this approach is discussed in the next section.

IV. RESULTS AND DISCUSSION

The proposed DRL based charge scheduling approach is experimentally evaluated with respect to different evaluation metrics.

A. *Experimental Setup*

Based on the defined city dimensions, a EV data set is created. This data set is included in Table 2.

TABLE II. BASE CONFIG DATA SET

Methodology & Focus	Reference
Number of Vehicles	20
Number of Charging Stations	10
X Dimension of city (km)	25
Y Dimension of city (km)	20

For navigation, this research identifies the details of the map locations which are tabulated in Table 3. Based on the map locations, the proposed approach simulates the map as shown in Fig. 3.

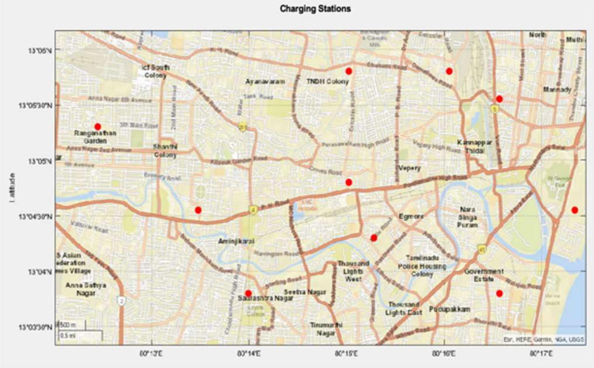


Fig. 3. Charging Station Locations

TABLE III. CHARGING LOCATION MAPPING

Locations	Simulation Map location (Latitude and Longitude)	Real Map locations (Latitude and Longitude)
1	(1, 1)	13.0859° N,80.2067° E
2	(3, 2)	12.9909° N, 80.2119° E
3	(4, 3)	13.0865° N, 80.2726° E
4	(-1, 2)	13.0698° N, 80.2245° E
5	(2, 2.5)	13.0853° N, 80.2607° E
6	(3.3, 4)	13.0938° N, 80.2891° E
7	(0.9, 3)	13.0629° N, 80.2314° E
8	(3, 5)	13.0732° N, 80.2609° E
9	(0, 3)	13.0696° N, 80.2728° E
10	(1.3, 5)	13.0806° N, 80.2876° E

With in this limits an EV charge scheduling dataset is created based on the features related to the charging station generator, charging station log, probability slabs, and EV generators the vehicle details with data required for validating the charging navigation which is included in Table 4.

TABLE IV. EV DATA SET

Vehicle Number	Available battery Power	Needed Charging Time	Starting X-Position	Starting Y-Position	Destination X-Position	Destination Y-Position
1	53	1.4	398	94	245	223
2	70	1.0	139	340	328	82
3	26	2.2	171	293	112	376
4	37	2.0	446	480	274	70
5	28	2.0	128	408	122	465
6	45	1.5	309	237	176	416
7	65	1.1	143	379	377	191
8	64	1.0	266	390	468	65
9	64	1.1	169	82	398	156
10	60	1.1	132	328	345	375
11	54	1.2	457	77	413	270
12	100	0.0	54	481	3	388
13	85	0.5	200	130	401	216
14	93	0.2	73	69	435	290
15	5	1.0	480	18	257	201
16	26	2.1	209	25	452	473
17	57	1.3	451	185	56	391
18	49	1.5	49	66	472	479
19	64	1.0	177	411	8	22
20	30	2.3	324	226	274	149

B. *Results*

The Tow count is the key output which is getting monitored between different stages. The stages which are getting monitored are represented in Fig. 4. For better charge scheduling the number Tow counts should be at minimal.

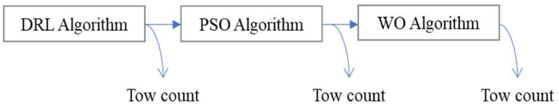


Fig. 4. Tow count computing stages

Using the generated data, initially, the performance of charge scheduling is determined without optimization. It was observed from the analysis that the total number of tow count vehicles without any optimization is 9. However, this number is too high and is not suitable for achieving appropriate charge scheduling. Hence the charge scheduling process is optimized using an existing particle swarm optimization (PSO) algorithm. Although PSO optimized results are better than the charge scheduling process without optimization, the number of Tow count vehicles is not completely minimized. The proposed WOA algorithm further minimize the Tow count vehicles. For the data set considered the Tow count got as zero. The result of the optimized charge scheduling is illustrated in Fig. 5.

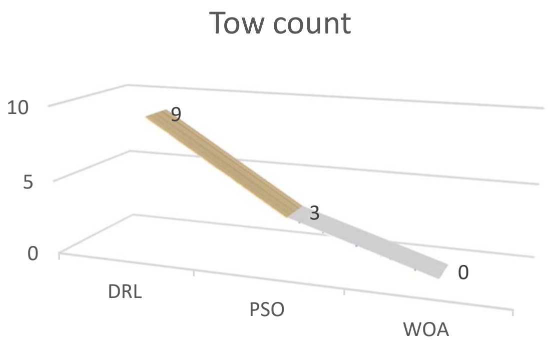


Fig. 5. Tow count values

As inferred from Fig. 6, the WOA algorithm achieves convergence in a lesser number of iterations and this shows that the features are optimized for the charge scheduling process

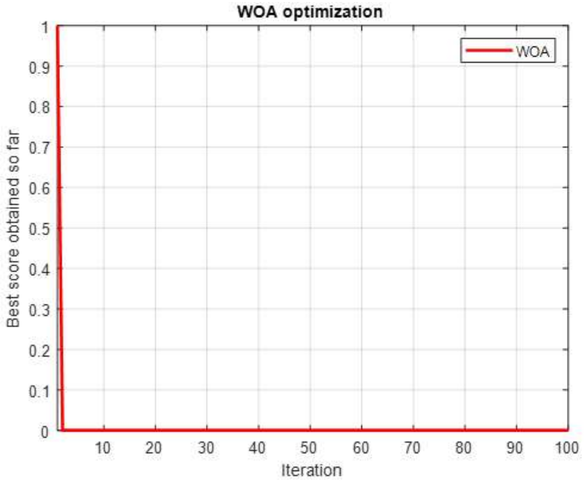


Fig. 6. WOA Iterations

Charge scheduling when the vehicle moves from one location to another with and without optimization is shown in below Fig. 7. In this its evident that the charge scheduling process is improved by also vehicle take the alternate routes. In addition, by improving the performance of the charge scheduling process the relative metrics such as average needed power, average charge cost, and average charging time also will be improved.

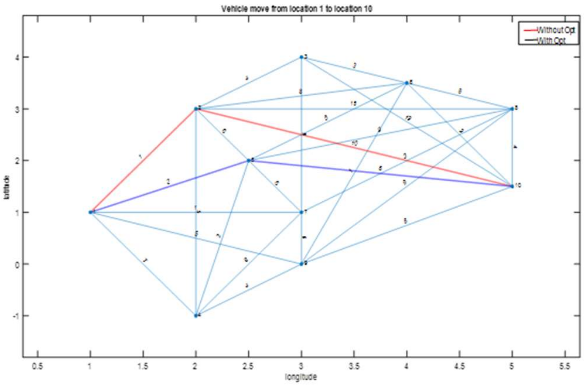


Fig. 7. Vehicle Navigation

Fig. 8 shows the implemented DRL based feature selector which selects the suitable features. This helps to reduce the overhead of analysing all features which inturn reduce the computational time.

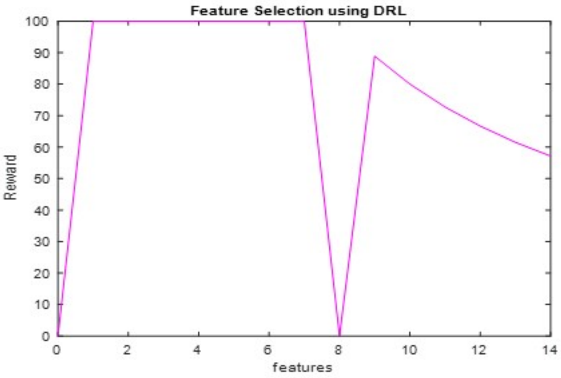


Fig. 8. Optimal Feature Selection

V. Conclusion And Future Work

The main aim of this research is to develop an efficient charge scheduling process for EVs to minimize the charging time and improve overall efficiency. An optimal EV charge scheduling approach is designed using an optimized DRL based framework. The proposed approach achieves an effective charge scheduling performance by selecting optimal features using the DRLFS technique. The efficacy of the DRLFS is evaluated in terms of effective feature selection which provides relevant features for optimizing the charge scheduling. The performance is evaluated without optimization and it was observed that the number of tow counts was high, which is not suitable for charge scheduling. Further, the features were optimized using the WOA, results show that the WOA reduces the number of tow count vehicles to zero in comparison to the existing PSO algorithm. In future, the study will be extended to evaluate the performance in a large city and with more vehicles.

Aligning to this, further exploration ought to focus on the below mentioned overlayered factors.

Integrate standardized protocols and interfaces: Evaluate the efficiency after integrating protocols and interfaces which are generally used with this frame work with diverse vehicle systems and components.

Safety and Validation: Evaluate this framework's capability to withstand the backdoor attacks which can be exploited to seriously harm the system components.

References

[1] Fescioglu-Unver, N., & Aktaş, M. Y. (2023). Electric vehicle charging service operations: A review of machine learning applications for infrastructure planning, control, pricing and routing. Renewable and Sustainable Energy Reviews, 188, 113873.

[2] Shahriar, S., Al-Ali, A. R., Osman, A. H., Dhou, S., & Nijim, M. (2020). Machine learning approaches for EV charging behavior: A review. IEEE Access, 8, 168980-168993.

[3] Yang, H., Deng, Y., Qiu, J., Li, M., Lai, M., & Dong, Z. Y. (2017). Electric vehicle route selection and charging navigation strategy based on crowd sensing. IEEE Transactions on Industrial Informatics, 13(5), 2214-2226.

[4] Zhang, X., Peng, L., Cao, Y., Liu, S., Zhou, H., & Huang, K. (2020). Towards holistic charging management for urban electric taxi via a hybrid deployment of battery charging and swap stations. Renewable Energy, 155, 703-716.

[5] Yan, L., Chen, X., Zhou, J., Chen, Y., & Wen, J. (2021). Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. IEEE Transactions on Smart Grid, 12(6), 5124-5134.

[6] Rezgui, J., & Cherkaoui, S. (2017, May). Smart charge scheduling for evs based on two-way communication. In 2017 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

[7] Nimalsiri, N. I., Ratnam, E. L., Smith, D. B., Mediwaththe, C. P., & Halgamuge, S. K. (2021). Coordinated charge and discharge scheduling of electric vehicles for load curve shaping. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7653-7665.

[8] Chung, H. M., Li, W. T., Yuen, C., Wen, C. K., & Crespi, N. (2018). Electric vehicle charge scheduling mechanism to maximize cost efficiency and user convenience. IEEE Transactions on Smart Grid, 10(3), 3020-3030.

[9] Lee, K. B., A. Ahmed, M., Kang, D. K., & Kim, Y. C. (2020). Deep reinforcement learning based optimal route and charging station selection. Energies, 13(23), 6255.

[10] Luo, L., Gu, W., Wu, Z., & Zhou, S. (2019). Joint planning of distributed generation and electric vehicle charging stations considering real-time charging navigation. Applied energy, 242, 1274-1284.

[11] Xia, F., Chen, H., Chen, L., & Qin, X. (2019). A hierarchical navigation strategy of EV fast charging based on dynamic scene. IEEE Access, 7, 29173-29184.

[12] Mo, W., Yang, C., Chen, X., Lin, K., & Duan, S. (2019). Optimal charging navigation strategy design for rapid charging electric vehicles. Energies, 12(6), 962.

[13] Mazhar, T., Asif, R. N., Malik, M. A., Nadeem, M. A., Haq, I., Iqbal, M., ... & Ashraf, S. (2023). Electric Vehicle Charging System in the Smart Grid Using Different Machine Learning Methods. Sustainability, 15(3), 2603.

[14] Panda, B., Rajabi, M. S., & Rajaee, A. (2022). Applications of Machine Learning in the Planning of Electric Vehicle Charging Stations and Charging Infrastructure: A Review. Handbook of Smart Energy Systems, 1-19.

[15] 15. Song, Y., Zhao, H., Luo, R., Huang, L., Zhang, Y., & Su, R. (2022). A sumo framework for deep reinforcement learning experiments solving electric vehicle charging dispatching problem. *arXiv preprint arXiv:2209.02921*.

[16] 16. Zhang, X., Chan, K. W., Li, H., Wang, H., Qiu, J., & Wang, G. (2020). Deep-learning-based probabilistic forecasting of electric vehicle charging load with a novel queuing model. *IEEE transactions on cybernetics*, *51*(6), 3157-3170.

[17] 17. Wang, R., Chen, Z., Xing, Q., Zhang, Z., & Zhang, T. (2022). A modified rainbow-based deep reinforcement learning method for optimal scheduling of charging station. *Sustainability*, *14*(3), 1884.

[18] 18. Venkitaraman, A. K., & Kosuru, V. S. R. (2023). Hybrid deep learning mechanism for charging control and management of Electric Vehicles. *European Journal of Electrical Engineering and Computer Science*, *7*(1), 38-46.

[19] 19. Wang, K., Wang, H., Yang, J., Feng, J., Li, Y., Zhang, S., & Okoye, M. O. (2022). Electric vehicle clusters scheduling strategy considering real-time electricity prices based on deep reinforcement learning. *Energy Reports*, *8*, 695-703.

[20] 20. Zhang, C., Liu, Y., Wu, F., Tang, B., & Fan, W. (2020). Effective charging planning based on deep reinforcement learning for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *22*(1), 542-554.

[21] 21. Wan, Z., Li, H., He, H., & Prokhorov, D. (2018). Model-free real-time EV charging scheduling based on deep reinforcement learning. *IEEE Transactions on Smart Grid*, *10*(5), 5246-5257.

[22] 22. Sadeghianpourhamami, N., Deleu, J., & Develder, C. (2019). Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning. *IEEE Transactions on Smart Grid*, *11*(1), 203-214.

[23] 23. Qian, T., Shao, C., Wang, X., & Shahidehpour, M. (2019). Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system. *IEEE transactions on smart grid*, *11*(2), 1714-1723.

DOI: 10.15439/2024F9093
ISSN 2300-5963 ACSIS, Vol. 40

Conceptional Framework for the Objective Work-Related Quality of Life Measurement Through Multimodal Data Integration from Wearables and Digital Interaction

□

Jenny Voigt, Jakob Hohn
0009-0003-1881-7422
0009-0000-3055-2357
4K Analytics GmbH
Email: {jenny.voigt, jakob.hohn}@4k-analytics.de

Ekaterina Mut, Christian Hrach, Ulf-Dietrich Brauman
0009-0000-1220-3305
0000-0002-1643-188X
0000-0002-0987-4498
Institute for Applied Informatics (InfAI e. V.)
Email: {mut, hrach, braumann}@infai.org

Celine Schreiber, Carsta Militzer-Horstmann
0009-0009-4130-4039
0000-0003-4566-9755
University of Leipzig, Health Economics and Management
Email: celine.schreiber@uni-leipzig.de; Email: militzer-horstmann@wifa.uni-leipzig.de

Sophia Mareike Geisler, Pauline Sophia Pinta, Alisa Hamm, Franziska Stutzer
0000-0001-6300-4349
0009-0004-9522-0897
0009-0006-9574-2964
0009-0003-1850-8750
Scientific Institute for Health Economics and Health System Research (WIG2 GmbH)
Email: {mareike.geisler, pauline.pinta, alisa.hamm, franziska.stutzer}@wig2.de

Hamlet Kosakyan
0009-0000-5103-6677
Appsfactory GmbH
Email: hamlet.kosakyan@appsfactory.de

Hubert Österle
0009-0002-0084-998X
University of St. Gallen, Business Engineering (Professor Emeritus),
Email: hubert@oesterle.ch

Juliette-Michelle Burkhardt, Bogdan Franczyk
0009-0001-1363-6716
0000-0002-5740-2946
University of Leipzig, Information Systems
Email: juliette-michelle.burkhardt@mailbox.tu-dresden.de; Email: franczyk@wifa.uni-leipzig.de

***Abstract*—In the evolving domain of occupational health, assessment of Work-related Quality of Life (WrQoL) has gained critical importance, particularly with recent expedited developments of decentralized and digital work. Conventional methods relying on subjective questionnaires are limited by high drop-out rates and potential biases. This paper introduces a novel approach to evaluating WrQoL by leveraging data generated from digital office environments, wearable devices, and smartphone applications. Our methodology includes the collection of physiological data, analysis of digital interactions, and prosody analysis to construct a comprehensive model of WrQoL influences. Initial and weekly questionnaires as well as multiple daily self-reports of valence and arousal levels will serve to initially validate this model. Prospectively utilizing machine learning, we aim to predict WrQoL scores from aggregated data. This method presents a non-invasive alternative for assessing WrQoL, providing significant implications for both research and industry with the potential to enhance workplace conditions and employee well-being.**

***Index Terms*—job satisfaction, machine learning, Occupational Health, valence, sensors, multimodal data integration, Organizational studies-Behavior**

I. Introduction

A. Background: (work-related) quality of life

IN THE contemporary landscape of occupational health and well-being, the concept of Quality of Life (QoL) and, more specifically, Work-related Quality of Life (WrQoL) has gained paramount importance. While recent expedited developments of decentralized work due to the COVID-19 pandemic entailed the promise of a more seamless integration of work and private life, it has simultaneously fractured the traditional work-leisure divide, necessitating a nuanced examination of the impacts of this new reality [1]. These new work constructs not only encapsulate the general well-being of individuals but also highlight the critical interplay between their professional environments and their life satisfaction. Understanding and improving WrQoL is essential for fostering productive, healthy, and sustainable workplaces. Research shows

□□□□This project is co-funded by the European Union and is co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament (ZX66X3CWB).

Freistaat Sachsen
This project is co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

Co-funded by the European Union

that high workplace stress severely impacts employees' mental and physical health. For instance, a Harvard University study found that insecure work environments increase poor health risks by about 50%, high job demands raise illness likelihood by 35%, and long working hours elevate mortality rates by nearly 20% [2]. This position paper presents the current research endeavor focused on the objective measurement of WrQoL, proposing a methodological advancement in the assessment and optimization of employee well-being beyond conventional measurement approaches.

QoL is a comprehensive concept encompassing overall well-being, reflecting both positive and negative life dimensions. It is inherently multidimensional, covering emotional, physical, material, and social aspects along with subdomains, such as income and wealth, jobs and earnings, housing, health status and work-life balance [3]–[5]. The World Health Organization (WHO) defines QoL as "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns," highlighting its subjective nature and measurement challenges (WHOQOL-BREF [3]).

Given the significant time spent at work, WrQoL is a crucial indicator of overall well-being and a core component and subdomain of QoL. WrQoL, which focuses on how the work environment affects an individual's overall QoL, is defined as "a multidimensional and dynamic psychological construct, directly related to individual and situational characteristics, which encompasses a set of worker characteristics and specific aspects of the organizational context" [6]. This definition highlights WrQoL's complexity, rooted in both personal attributes and workplace conditions. High WrQoL enhances job satisfaction, mental health, and personal fulfillment benefiting organizations with enhanced productivity, better employee retention, and reduced absenteeism [7]–[10]. Traditional WrQoL measurements rely on subjective self-report surveys, limited by respondent biases, mood fluctuations, and low response rates [11], [12]. This hampers both immediate data collection and longitudinal tracking. Hence, there is growing interest in developing more objective, reliable, and nuanced measurement tools to better understand WrQoL. The "Machine intelligence to objectively measure individual quality of life" (MI-LQ) project aims to objectively measure key indicators of WrQoL utilizing physiological data, digital interaction, and prosodic data, that will be incorporated into machine learning models. Data will be gathered in real office and residential environments as part of a pilot trial involving office workers. A mobile phone app prototype is developed serving as a user interface and a data relay system to a cloud-based platform for offline analysis of WrQoL metrics.

B. State of the art

Current research in the field of capturing and analyzing WrQoL includes a variety of methods that integrate both subjective and objective measurements. Subjective approaches such as questionnaires, interviews, and focus groups allow researchers to explore individual perceptions and interpretations of emotions in the workplace [13]. These predominantly qualitative methods provide important insights into the subjective experiences of employees and help to deepen our understanding of the complex dynamics of the work environment. The important insights include, for example, high workload, lack of support from superiors, job insecurity, inadequate work-life-balance or interpersonal conflicts [14]. Objective measurement approaches, however, rely on advanced technologies to capture and analyze physiological signals, behavioral data, and environmental factors. These advanced technologies can be organized based on the data sources they use, such as mouse, keyboard data, biosensors and mobile phone data.

Mouse and keyboard data

Behavioral data, such as keystrokes, mouse movements [15], and mobile phone activity [16], are used to identify behavioral patterns and derive emotional states. Recent studies by Naegelin *et al.* [17] and Shinde *et al.* [18] illustrate that the merging of physiological and behavioral data using machine learning models can lead to improved detection of workplace stress. Naegelin *et al.* [17] developed a machine learning method for stress detection based on multimodal data (mouse, keyboard, and cardiac data) and tested it in a simulated group office environment. They found that mouse and keyboard data detect stress in the office context better than cardiac data and that certain mouse movements and typing behavior are characteristic for specific stress predictions [17].

Biosensors

Wearable biosensors, such as heart rate monitors and skin conductance monitors, enable continuous data collection on workers' physical responses in different work situations [19]. Studies such as those by Shaffer and Ginsberg [20] and Ernst [21] provide a thorough analysis of heart rate variability metrics and their association with emotional states and work performance factors. Saganowski [22], on the other hand, discusses in particular the commercially available sensors for recording physiological data, signal processing techniques and deep learning architectures for the classification of emotions and the integration of emotion recognition technologies into everyday working life.

Mobile phone data

Furthermore, studies such as those by Burns *et al.* [23] and Hart *et al.* [24] show the usefulness of smartphone sensor technology for detecting depression and assessing well-being in the work context. For example, Burns *et al.* [23] developed a mobile phone application and supporting architecture like the cloud system and programming environment. This enabled the machine learning models to predict patients' mood, emotions, cognitive/motivational state, activities, environmental context, and social context based on at least 38 concurrent sensor readings from the phone (e.g., global positioning system, ambient light, recent calls) [23]. Contrarily, Hart *et al.* [24] investigated whether sparse motion-related sensor data can be used to train machine learning models capable of inferring individuals' states of work-related rumination, fatigue, mood, arousal, life engagement, and sleep quality. The participants' sensor data was collected via questionnaires on their smartphones [24].

Additionally, the objective measurement approaches can be categorized by target outcomes, such as:

Work-Life Balance

Research findings by Pawlicka *et al.* [25] and Gamage and Askana [26] contribute to the prediction of work-life balance and the detection of mental stress in IT work environments. For example, Pawlicka *et al.* [25] examined a machine learning tool to investigate correlations between employee-specific and job-related factors and the subjective feeling of work-life balance. They concluded that the relationship between the feeling of work-life balance and actual working hours was the most significant [25]. Moreover, Gamage and Asanka [26] have worked on a concept for a screening system that can predict mental health problems based on people's external characteristics. Supervised machine learning is used to identify workers at risk and refer them to professional help at an early stage [26].

Stress and emotion detection

Shinde *et al.* [18] developed the Real Time Employee Emotion Detection System (RTEED), which captures facial data in real time and uses machine learning to recognize the emotions of happiness, sadness, surprise, fear and disgust. The system helps companies to monitor the well-being of their employees and sends recognized emotions to the relevant employees to improve their work performance and lifestyle [18].

Moreover, Artificial Intelligence (AI) in speech analysis has the potential to become a crucial tool for measuring workplace stress as this technology can detect and evaluate stress-related strains in real time [27]. The work of Bromuri *et al.* [27] shows that a deep neural network trained for emotion recognition based on speech data can predict stress in call center employees with an accuracy of 80% in real time. This approach enables continuous and unobtrusive monitoring, which can lead to early warning systems and personalized training programs. The study by Baird *et al.* [28] investigates how language features can predict physiological stress markers. By using Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) and three German speech corpora, the study shows that speech features can effectively predict stress indicators such as cortisol levels, heart rate and respiration, opening up new possibilities for real-time, non-invasive stress monitoring.

These research findings contribute significantly to deepening our understanding of the physiological underpinnings of different aspects of WrQoL and support the development of more accurate models for predicting and assessing well-being in the workplace.

C. MI-LQ project: base model

Building upon our conceptual understanding of overall QoL, our focus is directed towards the nuanced factors shaping WrQoL, recognizing the pivotal role of workplace environment, job satisfaction, and occupational stress within this framework. The MI-LQ project is based on identifying key factors that influence WrQoL and translating them into a base model, which comprises the following indicators: 1) workload, 2) overtime, 3) workspace (office/home office), and 4) commute. All of these factors influence WrQoL through emotional experiences measured in two dimensions: 5) valence and arousal [29]. In addition, because of their strong direct influence on WrQoL, 6) spatial autonomy, 7) task autonomy, and 8) temporal autonomy are core components of the model (Fig. 1). The base model is conceptualized as an initial framework, established within office and residential settings to leverage digital behavioral data collection, crucial for refining and validating WrQoL indicators objectively. This conceptual framework guides the application of a digital assessment approach in a pilot study involving office workers for a minimum duration of four weeks, facilitating the initial validation and refinement of WrQoL assessment methods.

II. CONCEPTUAL FRAMEWORK

A. Data sources and model of data integration

The WrQoL indicators of the base model are measured by different objective measurement approaches. 1) Workload is assessed using calendar system-related data (e.g., Outlook) such as meeting overlap and meeting to work ratio. 2) Overtime is calculated as the difference between absolute hours worked, derived from computerized behavioral data, and self-reported contractual hours worked per week. 3) Workspace is intended to be assessed via workplace booking systems and 4) Commute via GPS. 5) Valence and arousal are determined by several approaches: a) work-related data from mouse and keyboard as well as software (e.g., Outlook, calendar) utilization b) physiological data application such as heart rate variability (HRV) from wrist-worn, c) stress index data based on plethysmography method obtained by ShenHealth application, and d) prosody analysis serving as speech-based emotion classification. 6–8) Autonomy data are mainly gathered by questionnaires and additionally by workplace booking systems (6), task management software, e.g., Trello (7) and working arrangement (8). Currently, the project is focused on gathering behavioral (5a), physiological (5b) and prosodic data (5d).

Features describing movements and actions of the mouse, speed of typing, and error correction of typed words have previously been studied as approach to predict work-related valence and arousal. Investigations have mostly been conducted in laboratory settings [15], [17], [30] and less in real office environments [31], with non-publicly available software.

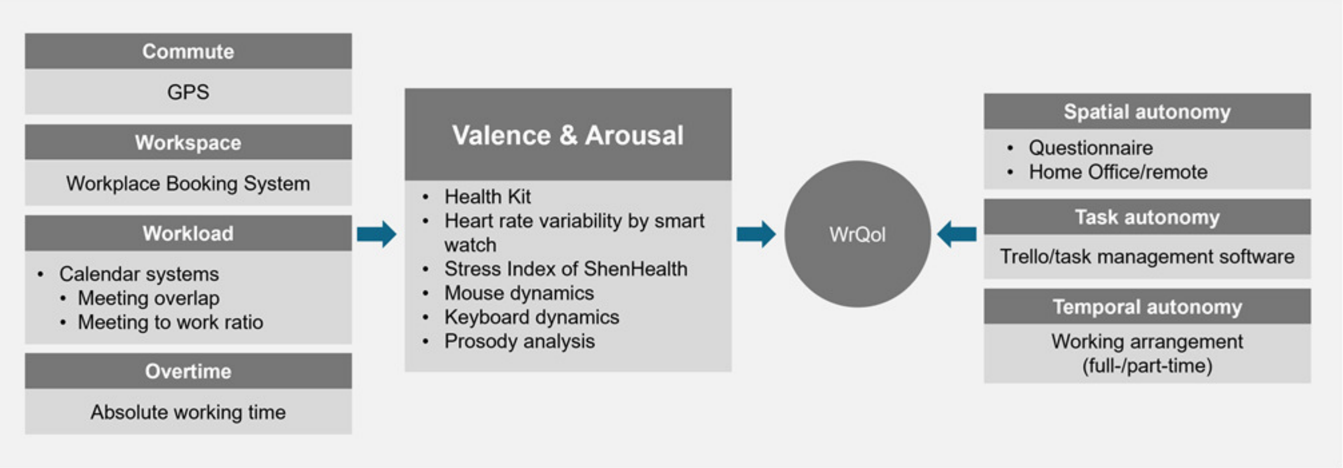


Fig. 1. Base model of WrQoL indicators and their objectifiability with different digital approaches

Therefore, a software prototype was developed to continuously monitor the mouse and keyboard dynamics by the office worker under real office conditions. For each keystroke, the time of press, the time of release and the type of the key pressed are documented, without content of typing. Privacy is ensured by distinguishing between two types of keys: delete keys (Delete, Backspace) and general use keys. Mouse dynamics include recording of mouse operation (move, scroll, click), time of movement, cursor speed, screen size in pixels, and x/y-coordinates for each mouse movement. The raw data collected are used to extract 25 mouse and 11 keyboard features per minute (Table I), representing the aggregated features with the highest predictive value [17]. In addition to the Naegelin *et al.* [17] metrics, two new metrics have been implemented: mean dwell time and SD dwell time to improve accuracy. The software prototype also monitors the applications and types of software being used by the office worker on a minute basis to extract the number of open windows without content analysis. Tracking and data collection is initiated after the user has launched the software and agreed to data collection upon every launch. To relate the aggregated mouse and keyboard data to the individual stress levels of office workers, respondents are asked to rate their current mood on a two-dimensional 5-point Likert scale several times a day via a software-integrated questionnaire using Self-Assessment Manikins (SAM) — non-verbal pictorial emotion manikins [31], [32]. The two dimensions describe valence and arousal, allowing any emotion to be described using the Circumplex model [29]. The Job-Related Well-Being Scale adapted this model to the work context [33].

B. *Questionnaires*

An initial questionnaire, based on existing validated WrQoL questionnaires, will be developed to assess baseline characteristics and self-reported WrQoL of office workers. Follow-up questionnaires will be administered throughout the study period to subjectively assess workload and emotional states for cross-checking with objective measurement data.

TABLE I.
EXAMPLES OF AGGREGATED DATA SOURCES OBTAINED BY MOUSE AND KEYBOARD INTERACTION

Mouse	**Keyboard**	**Software/ Application**
• Number of mouse movements per minute • Direct distance of a movement • Number of mouse pauses per minute • Mean/SD duration of a mouse pause • Mean/SD time between two clicks • Mean/SD Euclidean distance of a mouse movement • Mean/SD real distance of mouse movement • Mean/SD time duration of a mouse movement • Mean/SD average speed of a mouse movement per min • Mean/SD average angle of all angles in a movement • Mean/SD average distance of the real and straight line of a movement • Mean/SD sum of the difference between the real and straight line • Mean/SD number of direction changes in a mouse movement	• Number of pressed keys per minute • Error count • Mean/SD dwell time • Mean/SD digraph duration • Typing time • Mean/SD pause in typing per min • Keyboard pause count	• Average window count • Category of software (e.g., calendar, E-mail, presentation editor, text editor, messenger, programming environment etc.)

We evaluated four German validated self-report questionnaires and combined them into a comprehensive set of indicators of the base model and beyond, creating an initial and a shorter weekly questionnaire to be conducted in the MI-LQ app. The initial baseline questionnaire is designed to provide detailed information on participants' job satisfaction, autonomy dimensions, and work preferences. The shorter weekly check-in surveys via the MI-LQ app will be administered to capture changes over time on workload. The state of valence and arousal is asked using SAM at hourly intervals at the computer.

C. User interface

Our project's mobile phone app prototype (MI-LQ app) serves as both user interface and a data relay system to a cloud-based platform for offline analysis for WrQoL metrics. To enhance data collection, future iterations of the app will incorporate additional sensor categories, including work-related schedules and external influences. Ultimately, our aim is to provide participants with individualized WrQoL indicator scores and potential resources based on their analyzed data. The central aim of the application is to become a transparent point of data collection and data analysis presentation for the user, illustrated in Fig. 2.

D. Sensors and pulse data extraction and aggregation

After evaluation of various wearables from several manufacturers (Polar, Xiaomi, Apple, Samsung, Garmin, Fitbit), we decided to focus on Polar optical heart rate sensors: OH1 and Verity Sense. Those devices provide pulse-to-pulse intervals (PPI) extracted from photoplethysmography (PPG) signals. The Polar SDK for the iOS platform allows us to connect with a Polar device and perform live streaming measurements. The SDK provide pulse-to-pulse interval data in the following structure: 1) a PPI integer value which represents the interval between two pulses in milliseconds, 2) a heart rate (HR) value as calculated based on the PPI, 3) an error estimate integer value which represents an estimate of the expected absolute error of the PPI in milliseconds, 4) a block bit value which is set to 0 if PPI is considered valid, and otherwise set to 1, e.g., due to strong movement, and 5) a binary skin contact value which is 0 if there is no contact of the wearable to the skin detected, and 1 otherwise. Such additional information is important for a high-quality HRV analysis. Upon receipt from the sensor each PPI measurement is provided with a timestamp by the MI-LQ application and stored locally on the mobile phone in the following structure: Unix timestamp, PPI value, error estimate, block bit, skin contact bit. These data are accessible exclusively through the MI-LQ application, ensuring data security and privacy. The data is then processed and aggregated for analysis.

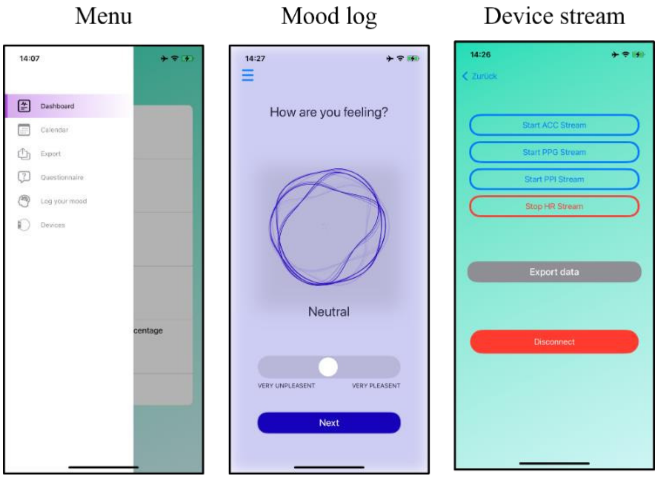

Menu Mood log Device stream

Fig. 2. Screenshots of the mobile app prototype

Basically, we can also integrate devices such as wearables into our application that can measure and provide health-related information. The workflow for those devices is slightly different:

1) Use device-specific application provided by manufacturers.

2) Synchronize health-related data from the manufactory's app into HealthKit (Apple, Cupertino, California, USA)

3) Grant permission for MI-LQ application to read data from HealthKit.

4) Prepare the received data for further aggregation and analysis.

E. Calculation of HRV metrics

HRV parameters have long been valued for their objective assessment of physical and mental status [34]. Within our preparative work we have seen more than 30 aggregated HRV-related indices based on Polar PPG sensor-derived PPI values. Although the manufacturers of so-called fitness trackers or pulse watches clearly have noticed the huge potential of the HRV framework for their consumer products, we rarely see reliable ready-to-use HRV-based measures, and some solutions applying a stress scale (from 0 to 100 percent) are presently not convincing: Calculations from different manufacturers are not transparent and cannot be verified; and each manufacturer uses its own scale, which is not comparable and not reproducible. Currently, we consider seven HRV-related indices (either statistical ones or combined ones according to Baevsky and Chernikova [34]) to be significant and meaningful as well as computationally effective:

1) Mean HR: mean of Heart Rate, i.e., average number of heart beats per minute

2) Mean NNI: mean of PPG-based PPIs

3) CV: Coefficient of Variation, i.e., mean of standard deviation of PPIs divided by the respective mean PPI (mean NNI)

4) RMSSD: Root Mean Square of Successive Differences of PPIs

5) SI: Stress Index according to Baevsky and Chernikova [34], i.e., ratio of amplitude of the modal value in the PPI histogram and the doubled product of modal value itself weighted by the min-max-span of PPIs

6) IVR: Index of Vegetative Regime is characterizing the ratio between sympathetic and parasympathetic influences on the heart rhythm, i.e., amplitude of the modal value in the PPI histogram divided by the standard deviation of PPIs

7) PAPR: Parameter of Appropriateness of Processes of heart Regulation, i.e., amplitude of the modal value in the PPI histogram divided by the respective modal value

These HRV parameters provide information about the autonomic nervous system during working hours and they allow to assess the effects of physiological responses to work stress.

F. AI-driven analysis of speech for emotion classification

In future work, we aim to develop a proprietary dataset for valence and arousal speech analysis based on the SAM scale [31], [32]. One of the objectives is to minimize aleatoric uncertainty, and therefore, we aim to achieve high entropy between the distinct classes. The available datasets do not meet our study's specific requirements, necessitating the decision to collect our own data. They lack German language compatibility, essential annotations for valence, and SAM, and access to datasets solely annotated with valence and arousal ratings has been denied despite requests. Moreover, the accessible dataset contains machine-generated annotations and an insufficient sample size (less than 200). Therefore, to ensure the completeness and accuracy of our research, independent data collection is imperative. Questions are asked under artificial induction of emotions according to Almazrouei *et al.* [35].

Once the dataset is created, we will develop separate classification models for valence and arousal respectively. These models will leverage prosodic features extracted from the speech data, such as pitch, shimmer, jitter (which play a crucial role when it comes to predicting stress out of the voice [36]), and Mel-Frequency Cepstral Coefficients (MFCC). According to the findings of Li *et al.* [36] these features play a critical role in the accurate classification of emotional states. For example, does a higher pitch directly correlate with emotions like anger.

Our approach will use these prosodic features to make predictions about the SAM ratings, providing a detailed analysis of how these features correlate with subjective emotional assessments. We will implement cross-validation techniques to evaluate the performance of the valence and arousal speech models, ensuring that they are rigorously tested and validated against diverse speech data collected through a website set up for this purpose.

This future work aims to contribute to the field of affective computing by providing a robust dataset and a validated methodology for speech-based stress classification, which could have wide-ranging applications in areas such as human-computer interaction and mental health monitoring.

G. Machine learning concept

Being time series data, PPI data as well as mouse and keyboard dynamics data allow tracking changes over time. Therefore, we use Long Short-Term Memory (LSTM) models based on the Keras framework to predict hourly valence and arousal class values, respectively. LSTM is a deep learning, sequential neural network that can learn hidden patterns in temporal sequences and retain information from previous time points [37]. The training data includes minute-by-minute data with up to 60 samples of HRV, mouse and keyboard features, along with information on defined active windows per target. These models are individually trained to generalize across different individuals, considering variations in physiological responses and interaction patterns, and are evaluated using classification report metrics to predict valence and arousal classes while capturing human emotional states over time.

III. Next Steps

Our research develops instruments to objectively assess workload through valence and arousal using smartphones and wearables in tandem with a software prototype for keyboard and mouse tracking. In the developmental stage of the technology, questionnaire responses from study participants are used to annotate and check data but will later be phased out when validity is reached.

Next steps in the MI-LQ project cover the establishment and implementation of the AI-driven analysis of speech-based emotion classification, an additional and supporting approach to detect and annotate stress-related strain at work to physiological and computer interaction-derived data. In addition, data derived from outlook, calendar, and project management system will be integrated into the MI-LQ app. Following this, the pilot study we aim to conduct will assess the reliability, validity, and feasibility of the digital framework for objective WrQoL assessment, involving office workers for a minimum of four weeks. This study will also enable the collection of work-related data within an authentic office environment, facilitating the assembly of a dataset sufficient for machine learning annotation.

IV. Limitations

While this study provides valuable insights into WrQoL and new approaches on how to measure it more objectively, several limitations should be acknowledged. First, due to the limited standards for wearable electronics, MI-LQ currently uses well-defined, brand-specific digital devices with proprietary hardware and software interfaces such as Polar optical sensors combined with iOS platform. Currently, there is no easy way to broadly integrate fitness trackers and smartwatches to reliably quantify QoL, as they lack comprehensive monitoring capabilities due to their lifestyle focus, limiting the number of potential users. HRV-related indices can only be calculated using PPI data, but most wearables only provide heart rate data and not PPI. This is not the case with Polar's optical sensors, which offer reliable measurement and data quality, and the great advantage of live streaming measurements when combined with the Polar iOS-SDK. As MI-LQ evolves, the integration of a wider range of digital devices and services will be considered, as well as the expansion to the Android platform. Second, in the initial phase, the MI-LQ project focused on assessing WrQoL specifically within the office context, recognizing the necessity to start with a defined population subset. Importantly, the modular design of the base model enables the seamless incorporation of new dimensions, indicators, and technologies as the study progresses. However, it's crucial to acknowledge the limitation of generalizability beyond the office context, necessitating further validation and demonstration of transferability in subsequent projects. This highlights the project's iterative nature and the ongoing refinement required to extend its applicability to diverse occupational settings such as construction workers, laboratory technicians, and beyond. Third, the data annotation process and subsequently the applied ML model depends on the provision of numerous reliable and continuous individual physiological and computational data and is therefore susceptible to patient dropouts. One countermeasure to avoid critical attrition rates that will be implemented in the pilot is incentives in the form of gift cards or expense reimbursements. Fourth, Likert scales, which will be part of the initial and weekly questionnaires need to be implemented carefully in terms of the number of items and a neutral position. Survey precision requires a careful balance; too few items risk imprecision, while an excess can hinder responses. Pre-test observations revealed a tendency to avoid extreme positions on the Likert scale.

V.Conclusion

WrQoL has emerged as a powerful indicator for industry and research into workplace conditions and employee well-being, revealing areas for improvement as well as levels of employee stress and burnout. As realized in the innovative MI-LQ approach, leveraging and linking the large amounts of data generated by wearable biosensors and computer interaction offers the opportunity to make WrQoL objectively measurable. In a base model, important physiological and WrQoL indicators have been considered and linked to measurement variables and instruments, set up in a modular way, with the possibility to expand the base model to more indicators and variables. Brand-specific digital devices and applications, ML and a mobile app architecture were found to be suitable tools as a basic framework for data input, output, processing, and prediction aimed at WrQoL assessment. Two wearable biosensors with satisfactory data granularity and quality were identified to provide the PPI needed to calculate the identified seven meaningful and computationally effective HRV indices. HRV data combined with mouse and keyboard dynamics and mood tracking data will be used to train LSTM models to predict particularly stressful working periods and emotional states. Opening to a greater variety of digital devices within the current model as well as expansion to other indicators or dimensions of WrQoL and/or QoL will be considered as MI-LQ evolves.

References

[1] J. H. Greenhaus, K. M. Collins, and J. D. Shaw, "The relation between work–family balance and quality of life," *JOURNAL OF VOCATIONAL BEHAVIOR*, vol. 63, no. 3, pp. 510–531, 2003, doi: 10.1016/S0001-8791(02)00042-8.

[2] J. Goh, J. Pfeffer, and S. A. Zenios, "Workplace Stressors & Health Outcomes: Health Policy for the Workplace," *Behavioral Science & Policy*, no. 1, pp. 43–52, 2015.

[3] WHOQOL Group, "Development of the World Health Organization WHOQOL-BREF quality of life assessment. The WHOQOL Group," *Psychological medicine*, vol. 28, no. 3, pp. 551–558, 1998, doi: 10.1017/s0033291798006667.

[4] Eurostat, "Final report of the expert group on quality of life indicators," 2017. Accessed: Mar. 1 2024. [Online]. Available: https://ec.europa.eu/eurostat/documents/7870049/7960327/KS-FT-17-004-EN-N.pdf/f29171db-e1a9-4af6-9e96-730e7e11e02f

[5] M. Durand, "The OECD Better Life Initiative: How's Life? and the Measurement of Well-Being," *Review of Income and Wealth*, vol. 61, no. 1, pp. 4–17, 2015, doi: 10.1111/roiw.12156.

[6] C. Mendes and H. Pereira, "Assessing the Impact of COVID-19 on Work-Related Quality of Life through the Lens of Sexual Orientation," *Behavioral sciences (Basel, Switzerland)*, vol. 11, no. 5, 2021, doi: 10.3390/bs11050058.

[7] A. Jaiswal and A. Mahila, "Quality of work life," *Journal of Business Management & Social Sciences Research*, 02.2014, pp. 83–87, 2014. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2cfb960da043838669f9f75bda049c193e60e3a8

[8] J. Leitão, D. Pereira, and Â. Gonçalves, "Quality of Work Life and Contribution to Productivity: Assessing the Moderator Effects of Burnout Syndrome," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, 2021, doi: 10.3390/ijerph18052425.

[9] S. Wang, D. Kamerade, B. Burchell, A. Coutts, and S. U. Balderson, "What matters more for employees' mental health: Job quality or job quantity?," *Camb. J. Econ.*, vol. 46, no. 2, pp. 251–274, 2022, doi: 10.1093/cje/beab054.

[10] D. J. Horst, E. E. Broday, R. Bondarick, L. F. Serpe, and L. A. Pilatti, "Quality of Working Life and Productivity: An Overview of the Conceptual Framework," *International Journal of Managerial Studies and Research*, 07.2014, pp. 87–98, 2014.

[11] C. Seale, Ed., *Researching society and culture*, 3rd ed. Los Angeles, Calif.: Sage, 2011.

[12] M. L. Patten, *Questionnaire Research: A Practical Guide*, 4th ed. Los Angeles: Taylor and Francis, 2014. [Online]. Available: http://gbv.eblib.com/patron/FullRecord.aspx?p=4710457

[13] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of biomedical informatics*, vol. 59, pp. 49–75, 2016, doi: 10.1016/j.jbi.2015.11.007.

[14] N. Nappo, "Job stress and interpersonal relationships cross country evidence from the EU15: a correlation

analysis," *BMC PUBLIC HEALTH*, vol. 20, no. 1, p. 1143, 2020, doi: 10.1186/s12889-020-09253-9.

[15] T. Androutsou, S. Angelopoulos, E. Hristoforou, G. K. Matsopoulos, and D. D. Koutsouris, "Automated Multimodal Stress Detection in Computer Office Workspace," *Electronics*, vol. 12, no. 11, p. 2528, 2023, doi: 10.3390/electronics12112528.

[16] 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, e24465, 2021, doi: 10.2196/24465.

[17] M. Naegelin *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *Journal of biomedical informatics*, vol. 139, p. 104299, 2023, doi: 10.1016/j.jbi.2023.104299.

[18] V. R. Shinde, K. R. Sonawane, A. D. Bhawar, and H. D. Walam, Eds., *Real time-Employee Emotion Detection system (RTEED) using Machine Learning*. Zenodo, 2023.

[19] S. Majumder, T. Mondal, and M. J. Deen, "Wearable Sensors for Remote Health Monitoring," *Sensors (Basel, Switzerland)*, vol. 17, no. 1, 2017, doi: 10.3390/s17010130.

[20] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Health*, vol. 5, p. 258, 2017, doi: 10.3389/fpubh.2017.00258.

[21] G. Ernst, "Hidden Signals-The History and Methods of Heart Rate Variability," *Front. Public Health*, vol. 5, p. 265, 2017, doi: 10.3389/fpubh.2017.00265.

[22] S. Saganowski, "Bringing Emotion Recognition Out of the Lab into Real Life: Recent Advances in Sensors and Machine Learning," *Electronics*, vol. 11, no. 3, p. 496, 2022, doi: 10.3390/electronics11030496.

[23] M. N. Burns *et al.*, "Harnessing context sensing to develop a mobile intervention for depression," *Journal of medical Internet research*, vol. 13, no. 3, e55, 2011, doi: 10.2196/jmir.1838.

[24] A. Hart, D. Reis, E. Prestele, and N. C. Jacobson, "Using Smartphone Sensor Paradata and Personalized Machine Learning Models to Infer Participants' Well-being: Ecological Momentary Assessment," *Journal of medical Internet research*, vol. 24, no. 4, e34015, 2022, doi: 10.2196/34015.

[25] A. Pawlicka, M. Pawlicki, R. Tomaszewska, M. Choraś, and R. Gerlach, "Innovative machine learning approach and evaluation campaign for predicting the subjective feeling of work-life balance among employees," *PloS one*, vol. 15, no. 5, e0232771, 2020, doi: 10.1371/journal.pone.0232771.

[26] N. S. Gamage and D. P. Asanka, "Machine Learning Approach to Predict Mental Distress of IT Workforce in Remote Working Environments," in *2022 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka, 2022, pp. 211–216.

[27] S. Bromuri, A. P. Henkel, D. Iren, and V. Urovi, "Using AI to predict service agent stress from emotion patterns in service interactions," *J. Serv. Manage.*, vol. 32, no. 4, pp. 581–611, 2021, doi: 10.1108/JOSM-06-2019-0163.

[28] A. Baird *et al.*, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Front. Comput. Sci.*, vol. 3, 2021, doi: 10.3389/fcomp.2021.750284.

[29] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[30] L. Pepa, A. Sabatelli, L. Ciabattoni, A. Monteriu, F. Lamberti, and L. Morra, "Stress Detection in Computer Users From Keyboard and Mouse Dynamics," *IEEE Trans. Consumer Electron.*, vol. 67, no. 1, pp. 12–19, 2021, doi: 10.1109/TCE.2020.3045228.

[31] N. Banholzer, S. Feuerriegel, E. Fleisch, G. F. Bauer, and T. Kowatsch, "Computer Mouse Movements as an Indicator of Work Stress: Longitudinal Observational Field Study," *Journal of medical Internet research*, vol. 23, no. 4, e27121, 2021, doi: 10.2196/27121.

[32] M. M. Bradley and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994, doi: 10.1016/0005-7916(94)90063-9.

[33] P. T. van Katwyk, S. Fox, P. E. Spector, and E. K. Kelloway, "Using the Job-Related Affective Well-Being Scale (JAWS) to investigate affective responses to work stressors," *J. Occup. Health Psychol.*, 1939-1307(Electronic),1076-8998(Print), pp. 219–230, 2000, doi: 10.1037/1076-8998.5.2.219.

[34] R. M. Baevsky and A. G. Chernikova, "Heart rate variability analysis: physiological foundations and main methods," *Cardiometry*, no. 10, pp. 66–76, 2017, doi: 10.12710/cardiometry.2017.10.6676.

[35] M. A. Almazrouei, R. M. Morgan, and I. E. Dror, "A method to induce stress in human subjects in online research environments," *BEHAVIOR RESEARCH METHODS*, vol. 55, no. 5, pp. 2575–2582, 2023, doi: 10.3758/s13428-022-01915-3.

[36] X. Li *et al.*, "Stress and Emotion Classification using Jitter and Shimmer Features," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing: ICASSP 2007] ; Honolulu, HI, 16 [i.e. 15] - 20 April 2007*, Honolulu, HI, 2007, IV-1081-IV-1084.

[37] D. M. Ahmed, M. M. Hassan, and R. J. Mstafa, "A Review on Deep Sequential Models for Forecasting Time Series Data," *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–19, 2022, doi: 10.1155/2022/6596397.

DOI: 10.15439/2024F3588
ISSN 2300-5963 ACSIS, Vol. 40

A generic method of pose generation in molecular docking via quadratic unconstrained binary optimization

Pei-Kun Yang
Research Center for Applied Sciences,
Academia Sinica
Taipei, Taiwan
peikun@gate.sinica.edu.tw

Jung-Hsin Lin
Biomedical Translation Research Center, Research Center for Applied Sciences,
Academia Sinica
Taipei, Taiwan
jhlin@gate.sinica.edu.tw

***Abstract*—Docking of a ligand onto the binding pocket of its protein target, designated as the molecular docking problem, is a very important method for structure-based drug design. We have implemented a generic pose generation method for molecular docking by solving the quadratic unconstrained binary optimization (QUBO) problem with the Fujitsu digital annealer. In combination with the AutoDock 4 scoring function, the success rate for predicting the binding poses to be sufficiently close to their experimental binding poses, namely, with the root mean squared deviation (RMSD) less than 2 Å, was 84.3 %, when benchmarking against part of the PDBbind core set (242 protein-ligand complexes). To our best knowledge, this is the first implementation of molecular docking that conforms with the QUBO formalism demonstrating a performance comparable with the conventional methods.**

***Index Terms*—molecular docking, quadratic unconstrained binary optimization, QUBO, pose generation, AutoDock**

I. Introduction

MOLECULAR docking is an essential method for structure-based drug design and virtual screening of chemical libraries for finding chemical skeletons for creating novel chemical entities. Docking of a ligand onto the binding pocket of its protein target, generally consists of two parts: pose generation and binding affinity evaluation. In the first step, a myriad of ligand conformations at the protein surface (usually at the binding pocket) need to be generated, and these conformations should include the ones that are very close to the experimentally determined binding poses. Typical experimental methods are protein X-ray crystallography, nuclear magnetic resonances, and cryogenic electron microscopy. In the second step, the binding affinities of these ligand conformations at the protein binding pocket will be evaluated with a scoring function (or a free energy functional), and the poses with best binding affinities should be very close the experimental binding poses. It is considered as a successful molecular docking when the second step can be achieved, i.e., the root mean squared deviation (RMSD) of the predicted binding pose with the best score (binding affinity) from the experimental binding pose is less than, e.g., 2 Å.

Quadratic unconstrained binary optimization (QUBO), sometimes also known as unconstrained binary quadratic programming (UBQP), is a class of combinatorial optimization problems with a huge variety of applications. QUBO is known as an NP hard problem. Many classical problems from theoretical computer science, e.g., maximum cut, graph coloring and the partition problem, have been formulated into QUBO. Due to its close connection to Ising models, QUBO constitutes a major class of computational problems for adiabatic quantum computation, where it can be solved through a physical process named quantum annealing.

D-Wave and Fujitsu are two well-known companies that strive to develop computers to efficiently solve the QUBO problems with quantum annealer and quantum-inspired (or physics-inspired) annealers, respectively. Although many important applications have been embedded into the QUBO formulism, molecular docking is still not yet implemented and it is not clear whether such an implementation can indeed lead to practically useful applications.

II. Methods

Many important problems in molecular biology, including protein folding, protein-protein binding, protein-DNA (or RNA) binding, and protein-ligand binding, are problems of searching for free energy minimum, from the perspective of statistical thermodynamics. The problem of finding the minimum value in one dimension can be easily solved by, e.g., the Newton-Raphson method, etc. The difficulty of finding the global minimum exponentially escalates as the dimensionality increases. Compared with algorithms that have been developed for decades, the emerging hardware such as those developed by D-Wave and Fujitsu have the opportunity to find solutions with lower function values at high dimensions with dramatically less time than the conventional methods. The QUBO formulism generally reads:

$$F(\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{J}\mathbf{X} + \mathbf{H}\mathbf{X} \tag{1}$$

In Equation (1), **X** has n binary variables, and **J**, **H** and **X** are $(n \times n)$, $(1 \times n)$ and $(\mathrm{n} \times 1)$ matrices respectively. Given the element values J_{ij} and H_{i} of the **J** and **H** matrices, QUBO solvers can be used to find a set of solutions to X that minimize the value of F(**X**).

In this work we construct a QUBO model for finding binding poses in molecular docking. In a 3-D lattice covering the binding pocket, QUBO solvers should obtain a solution X_i of 1, to indicate a ligand atom will locate as this lattice point. Compare the distribution of ligand atoms and X_i to obtain the possible binding positions of ligand. In addition to using Fujitsu DAU3 as QUBO solvers, we can also PyTorch to solve the QUBO model.

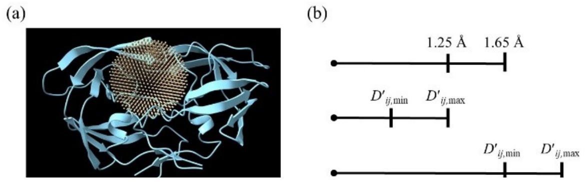


Figure 1: (a) Lattice points near to the binding pocket of the protein in the PDB ID 1a30. The protein conformation was shown in ribbon representation, colored in cyan. In the region of the protein binding pocket, the grey round points are the lattice points X_i. (b) (Up) The distance between the non-hydrogen covalent bonds are between 1.25 Å and 1.65 Å. (Middle) The maximum distance between two lattice points $D'_{ij,\max}$ is less than 1.25 Å. Lattice points *i* and *j* cannot co-exist inside an atom. (Down) The minimum distance between two lattice points $D'_{ij,\min}$ is larger than 1.65 Å. Lattice points *i* and *j* cannot form the covalent bond.

III. Results

Our major results can be seen from Figure 2 and Figure 3.

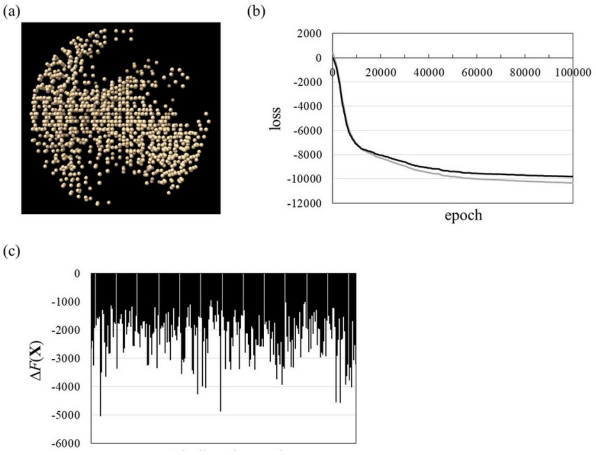


Figure 2: (a) Lattice points with $X_i = 1$, as determined by Fujitsu DAU3. Represented as small round spheres. (b) The grey line shows that $F(\mathbf{X}')$ decrease as the number of epoch increases. The black line shows the values of $F(\mathbf{X})$ in the course of epoch. (c) The difference between the minimum values determined by Fujitsu DAU3 and PyTorch.

IV. Discussion

From our results it indicates that with proper implementation and suitable parameters, it is possible to embed the molecular docking problem into the QUBO format. Currently our implementation is largely guided by our intuition and our prior understandings of the molecular docking problem, and these background and experiences greatly accelerate the progress of this work. It may be possible to employ some machine learning approaches or large language models, such as newer generation of ChatGPT, to translate the problem of interest into the QUBO formulism. However, it may take some more time to witness this to be a reality.

V. Conclusion

Molecular docking is an essential workhorse for structure-based drug design and virtual screening of chemical libraries for finding chemical skeletons for creating novel chemical entities. Emerging hardware such as those developed by D-Wave and Fujitsu have a great opportunity to find solutions with lower function values at high dimensions with dramatically less time than the conventional methods. To our best knowledge, this is the first implementation of molecular docking that conforms with the QUBO formalism demonstrating a performance comparable with the conventional methods. It can be envisioned that such an approach could evolve to become to more efficient and more accurate method for molecular docking and thereby accelerate the drug discovery process.

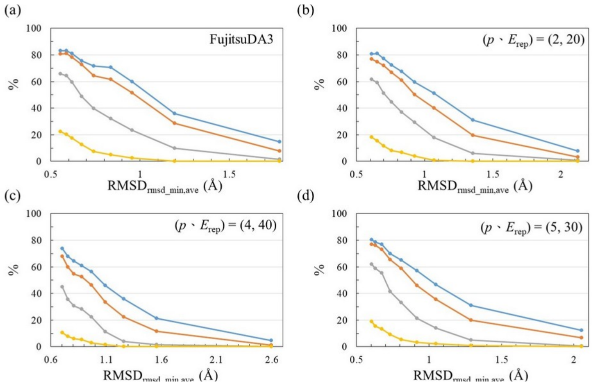


Figure 3: Resulting using (a) FujitsuDAU3, (b) PyTorch (p, E_{rep}) = (2, 20), (c) (4, 40) and (d) (5, 30) for solving QUBO. The fraction of RMSD $_{\text{bfe\_min}}$ < 2.0 Å (blue line), < 1.5 Å (red line)、 < 1.0 Å (grey line) and < 0.6 Å (yellow line)。

Acknowledgment

JHL would like to acknowledge the financial support of National Science of Technology Council (NSTC), under the grants NSTC 111-2112-M-001-053-MY3 and NSTC 112-2119-M-006-004. JHL would like to also knowledge the financial support and Research Center for Applied Sciences (RCAS), Biomedical Translational Research Center (BioTReC), both are research institutes of Academia Sinica (AS).

References

[1] Maia, E. H. B.; Assis, L. C.; De Oliveira, T. A.; Da Silva, A. M.; Taranto, A. G., Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in chemistry* **2020,** *8*, 343.

[2] Varela-Rial, A.; Majewski, M.; De Fabritiis, G., Structure based virtual screening: Fast and slow. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022,** *12* (2), e1544.

[3] Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M., RCSB Protein Data Bank (RCSB. org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research* **2023,** *51* (D1), D488-D508.

[4] Tingle, B.; Tang, K.; Castanon, J.; Gutierrez, J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y.; Irwin, J., ZINC-22-A free multi-billion-scale database of tangible compounds for ligand discovery. **2022**.

[5] Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L., Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling* **2012,** *52* (11), 2864-2875.

[6] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016,** *72* (2), 171-179R.

[7] Fu, H.; Chen, H.; Blazhynska, M.; Goulard Coderc de Lacam, E.; Szczepaniak, F.; Pavlova, A.; Shao, X.; Gumbart, J. C.; Dehez, F.; Roux, B., Accurate determination of protein: ligand standard binding free energies from molecular dynamics simulations. *Nature Protocols* **2022,** *17* (4), 1114-1141..

[8] Heinzelmann, G.; Gilson, M. K., Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation. *Scientific reports* **2021,** *11* (1), 1-18.

[9] Wang, E.; Fu, W.; Jiang, D.; Sun, H.; Wang, J.; Zhang, X.; Weng, G.; Liu, H.; Tao, P.; Hou, T., VAD-MM/GBSA: A Variable Atomic Dielectric MM/GBSA Model for Improved Accuracy in Protein–Ligand Binding Free Energy Calculations. *Journal of Chemical Information and Modeling* **2021**.

[10] Dittrich, J.; Schmidt, D.; Pfleger, C.; Gohlke, H., Converging a knowledge-based scoring function: DrugScore2018. *Journal of chemical information and modeling* **2018,** *59* (1), 509-521.

[11] Bao, J.; He, X.; Zhang, J. Z., Development of a New Scoring Function for Virtual Screening: APBScore. *Journal of Chemical Information and Modeling* **2020,** *60* (12), 6355-6365.

[12] Cavasotto, C. N.; Aucar, M. G., High-throughput docking using quantum mechanical scoring. *Frontiers in chemistry* **2020,** *8*, 246.

[13] Kadukova, M.; Machado, K. d. S.; Chacón, P.; Grudinin, S., KORP-PL: a coarse-grained knowledge-based scoring function for protein–ligand interactions. *Bioinformatics* **2021,** *37* (7), 943-950.

[14] Yang, C.; Zhang, Y., Lin_F9: A Linear Empirical Scoring Function for Protein–Ligand Docking. *Journal of Chemical Information and Modeling* **2021**.

[15] Schneider, C.; Buchanan, A.; Taddese, B.; Deane, C. M., DLAB: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* **2022,** *38* (2), 377-383.

[16] Li, H.; Sze, K. H.; Lu, G.; Ballester, P. J., Machine-learning scoring functions for structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021,** *11* (1), e1478.

[17] Ricci-Lopez, J.; Aguila, S. A.; Gilson, M. K.; Brizuela, C. A., Improving structure-based virtual screening with ensemble docking and machine learning. *Journal of Chemical Information and Modeling* **2021,** *61* (11), 5362-5376.

[18] McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R., GNINA 1.0: molecular docking with deep learning. *Journal of cheminformatics* **2021,** *13* (1), 1-20.

[19] Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S., AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling* **2021,** *61* (8), 3891-3898.

[20] Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C., DOCK 6: Impact of new features and current docking performance. *Journal of computational chemistry* **2015,** *36* (15), 1132-1156.

[21] Ypma, T. J., Historical development of the Newton–Raphson method. *SIAM review* **1995,** *37* (4), 531-551.

[22] Willsch, D.; Willsch, M.; Gonzalez Calaza, C. D.; Jin, F.; De Raedt, H.; Svensson, M.; Michielsen, K., Benchmarking Advantage and D-Wave 2000Q quantum annealers with exact cover problems. *Quantum Information Processing* **2022,** *21* (4), 141.

[23] Şeker, O.; Tanoumand, N.; Bodur, M., Digital annealer for quadratic unconstrained binary optimization: a comparative performance analysis. *Applied Soft Computing* **2022,** *127*, 109367.

[24] Woods, B. D.; Kochenberger, G.; Punnen, A. P., QUBO Software. In *The Quadratic Unconstrained Binary Optimization Problem*, Springer: 2022; pp 301-311

[25] Zaman, M.; Tanahashi, K.; Tanaka, S., PyQUBO: Python library for mapping combinatorial optimization problems to QUBO form. *IEEE Transactions on Computers* **2021,** *71* (4), 838-850.

[26] Kochenberger, G.; Hao, J.-K.; Glover, F.; Lewis, M.; Lü, Z.; Wang, H.; Wang, Y., The unconstrained binary quadratic programming problem: a survey. *Journal of combinatorial optimization* **2014,** *28* (1), 58-81.

[27] Tavares, G., New algorithms for Quadratic Unconstrained Binary Optimization (QUBO) with applications in engineering and social sciences. Rutgers The State University of New Jersey-New Brunswick: 2008.

[28] Koh, Y. W.; Nishimori, H., Quantum and classical annealing in a continuous space with multiple local minima. *Physical Review A* **2022,** *105* (6), 062435.

[29] Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S., The PDBbind database: methodologies and updates. *Journal of medicinal chemistry* **2 005,** *48* (12), 4111-4119.

[30] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019,** *32*.

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 73–78 DOI: 10.15439/2024F8800 ISSN 2300-5963 ACSIS, Vol. 40

Resource efficient Internet-of-Things intrusion detection with spiking neural networks

Miloš Živadinović
0000-0002-0342-340X
Faculty of Organizational Sciences
University of Belgrade
Jove Ilića 154, 11000 Beograd, Republic of Serbia
Email: mzdv@protonmail.com

Dejan Simić
0000-0002-0744-5411
Faculty of Organizational Sciences
University of Belgrade
Jove Ilića 154, 11000 Beograd, Republic of Serbia
Email: dsimic@fon.bg.ac.rs

***Abstract*—Spiking neural networks are a novel implementation of artificial neural networks closely based on neurobiology. Our goal is to analyze and see the plausibility of spiking neural networks as intrusion detection models based on the BoT-IoT dataset under a limited set of circumstances. We created a spiking neural network classifier in PyTorch and snn-torch based on Leaky Integrate-and-Fire neurons that managed to get an F1 score of 0.957 on 10 000 samples of the BoT-IoT dataset and 240 hidden spiking neurons. We performed training on the CPU for 300 epochs and 10 simulation steps per epoch, utilizing Adam optimizer, cross-entropy loss and backpropagation as a learning algorithm. Lowering hidden spiking neuron count from 240 to 72 and sample size from 10 000 to 1 000, we were able to optimize training time by 84% and testing time by 57% while having an F1 score of 0.944. We present Loss, Receiver Operating Characteristics, and Precision-Recall curves for the two experiments and summarized data for additional experiments performed with different sample sizes and neuron counts. We conclude that spiking neural networks for intrusion detection represents a viable solution for training and classification on resource-constrained devices with limited samples. Further research steps are presented to improve performance.**

***Index Terms*—spiking neural networks, artificial intelligence, intrusion detection systems, internet of things**

I. Introduction

EVEN though a strict definition does not exist, Internet-of-things (IoT) can be defined as a set of connected devices with the goal of exchanging sensor and communication data between themselves to provide joint computing capabilities [1]. These devices often function in environments with constrained resources, such as electricity, processing speed, and memory. In 2023 there were 15.14 billion IoT devices active, with the trend increasing to more than 29 billion devices by the decade's end [2]. allowing potential malicious actors to launch attacks across different domains and infrastructure.

With the global number of IoT devices available, they present a formidable attack surface for targeted cyber attacks. The number of cyber attacks launched against IoT devices surpassed one hundred twelve million potential intrusions in 2022. [3], emphasizing a need for IoT device intrusion detection systems. Ideally, edge devices should handle intrusion detection for quick recognition and adequate accuracy.

Dorothy Denning defined the first known occurrence of intrusion detection systems (IDS) [4] in 1987. She defined intrusion detection systems as expert systems consisting of a tuple of six elements (Subjects, Objects, Audit records, Profiles, Anomaly records, and Activity rules with the role of detecting anomalies in computer system access. Expert system methodology was gradually improved and replaced by statistical analysis [5] and pattern-oriented intrusion detection [6], as well as machine learning and artificial intelligence methods [7]. Applying statistics, machine learning, and artificial intelligence allows greater autonomy for IDS systems and improves response times and detection rates with unknown intrusions.

Deploying IDS solutions to IoT devices has always been challenging due to the constraints of resources mentioned above, most notably available memory. Machine learning and statistics powered IDS solutions provide potential improvements for deployment on IoT devices, but the device's computing power often constrains them.

The concept of neuromorphic computing promises to improve upon these limitations. Neuromorphic computing is defined as the development of computer systems based on biological characteristics of neurons and nerve systems [8]. One of important neuromorphic computing concepts are spiking neural networks, which are more energy and efficient than their traditional counterparts, as shown in the Izhikevich model [9] which is used as the baseline for biological neuron simulation due to its biological plausibility.

Spiking neural networks can be traced back to the original discovery of neural spiking by Hodgkin and Huxley [10], which was later abstracted by Bonhoeffer [11] into a more general purpose model that handles different kinds of neuronal behavior. This model is known as the Bonhoeffer - van der Pol model, inspired by models of the human heart [12]. In 1982. Hopfield [13] defined the concept of a "Hopfield network," which represents the first artificial spiking neural network.

The key feature that provides efficiency is the concept of spiking (or action potentials) [10] which allow efficient signaling of changes between neurons. Each neuron has an activation function operating on the electrochemical interactions between the neuron and its environment. When the neuron reaches the activation potential threshold, it performs an electrical discharge to other connected neurons.

Spiking neural networks on dedicated chips can utilize a maximum 65mW of power usage per 1 000 000 spiking neurons [14]. Asghar et al. [15] created a neuromorphic chip based on spiking neural networks that are even more resource efficient, consuming, on average, 1.06 mW of power. The Spiking-YOLO model has achieved similar resource efficiency [16] utilizing spiking neural networks for object detection.

Another key difference between spiking neural networks and traditional artificial neural networks is the existence of time as a component used for neuronal learning. Spike timing dependant plasticity (STDP) [17] is based on the duration before pre-synaptic and post-synaptic spikes and represents one of the key learning mechanisms in artificial and biological spiking neural networks. Time component utilization represents the side-effect spiking neural network models, which are based on systems of differential equations compared to other common artificial intelligence models. It is worth mentioning that, like modern artificial neural network learning, Lillicrap et al. [18] observed a variant of backpropagation as a learning mechanism for biological neurons.

With current knowledge, applying spiking neural networks to the domain of intrusion detection would allow us equal or better performance than current state-of-the-art solutions. The added benefit would be significantly less usage of power and computing resources, per current literature, making them suitable for IoT devices.

II. Current state of spiking neural networks and intrusion detection

The application of spiking neural networks in intrusion detection systems is a relatively new concept. One of the first usages dates back to 2014. spiking neural network concepts managed to get a 99.78% success rate when detecting failure rates in power systems [19].

Alom and Taha defined a way for autoncoders to be transformed into spiking neural networks to gain the benefits of neuromorphic computing [20]. Utilizing IBM TrueNorth [14], it was possible to have an intrusion detection system with 90.12% accuracy consuming less than 50 mW of power. This research also presents the method of converting traditional artificial neural networks to spiking neural networks, allowing potential performance improvements without retraining.

Zhou et al. introduced the first complete spiking neural network implementation for intrusion in 2020 [21]. The system in question consists of three layers with a total of 205 neurons, and it managed to reach 98.98% accuracy for intrusion detection.

Zarzoor et al. have applied spiking neural networks with decision trees for Internet-of-things attack classification [22] with 95% accuracy on attack classification from the IoT Botnet 2020 dataset.

Besides spiking neural networks, Hassini et al. [23] provide a solution based on deep learning for intrusion detection that reached 99.96% accuracy across 15 classes for edge IoT devices. Although unrelated to spiking neural networks, this solution represents one of the most advanced state-of-the-art approaches.

Encountered research so far focuses on spiking neural networks that were used for classification outside of Internet-of-things devices. Even though this approach proves their applicability, it does not factor in the possibility of IoT devices performing attack classification independently of a centralized classifier.

Table I contains summarized findings with F1 scores and accuracy, whichever metrics are available due to the quality of work.

TABLE I
Summarized findings from current state

Paper	Accuracy
Wang et al. [19]	99.78%
Alom and Taha [20]	90.12%
Zhou et al. [21]	98.98%
Zarzoor et al. [22]	95%
Hassini et al. [23]	99.96%

III. Experiment setup

We have used the BoT-IoT [24]–[27], [27], [28] dataset for our research due to its subject area of network traffic belonging to IoT devices. Due to resource constraints regarding computing power used for training, we have used the already prepared 5% subset of the BoT-IoT dataset to perform training and testing of the model. The dataset contains 3 668 522 entries, with 477 entries marked as non-malicious and the remaining marked as malicious, randomized and split into 80% of the dataset used for training and 20% of the dataset used for testing before further subsampling for our experiments. We randomized data before subsampling according to experiment requirements and applied mini-batching with a size of 100 samples per mini-batch as a way to optimize limited experiment resources.

We have not used spike trains for our experiments and have decided to use the numeric representation of data in order to simplify the experiment.

The BoT-IoT dataset contains 46 different features. In order to simplify handling, we have selected the subset of features shown in Table II:

TABLE II
Selected BoT-IoT features

Feature name	Feature description	Type
proto	Network traffic protocol	String
spkts	Source-to-destination packet count	Numeric
dpkts	Destination-to-source packet count	Numeric
srate	Source-to-destination packets per second	Numeric
drate	Destination-to-source packets per second	Numeric
state	Transaction state	String

We have performed one-hot encoding of the *proto* and *state* features for easier training, increasing the total input features to 18 presented in Table III. The final list of utilized features is presented below:

TABLE III
MODIFIED SELECTED BOT-IOT FEATURES

Feature name	**Feature description**	**Type**
arp	ARP protocol detected	Boolean
icmp	ICMP protocol detected	Boolean
tcp	TCP protocol detected	Boolean
udp	UDP protocol detected	Boolean
spkts	Source-to-destination packet count	Numeric
dpkts	Destination-to-source packet count	Numeric
srate	Source-to-destination packets per second	Numeric
drate	Destination-to-source packets per second	Numeric
acc	ACC state	Boolean
con	CON state	Boolean
eco	ECO state	Boolean
fin	FIN state	Boolean
int	INT state	Boolean
mas	MAS state	Boolean
req	REQ state	Boolean
rst	RST state	Boolean
tst	TST state	Boolean
urp	URP state	Boolean

The goal of our research was to classify attack subcategories according to the above features. There was a total of four attack subcategories identified in the BoT-IoT dataset:

- Denial of Service (DoS)
- Distributed Denial of Service (DDoS)
- Reconnaissance
- Theft

Due to the uneven distribution of attack subcategories, we have removed the Theft attack subcategory from the classification since its number of occurrences is much lower than that of other attack subcategories. Theft attack subcategory has the potential to skew results since we need more data to train for Theft subcategories (in the 5% dataset, the Theft subcategory appears in less than 0.01% due to rounding). We applied one-hot encoding to the remaining subcategories. Table IV shows the number of occurrences for attack subcategories:

TABLE IV
ATTACK SUBCATEGORY OCCURENCE

Attack subcategory	**Number of occurrences**	**Percentage of dataset**
DoS	1 650 260	44.99%
DDoS	1 926 624	52.52%
Reconnaissance	91 082	2.48%
Theft	79	Less than 0.01%

The classifier consists of two main Leaky Integrate-and-Fire layers (with decay rate $\beta = 0.95$ and 120 fully connected neurons per layer) joined with three linear transformation layers used for reshaping data. Although we have introduced the Izhikevich model previously in this paper, we have decided to use the Leaky Integrate-and-Fire spiking neuron model, which is easier to implement, albeit with less biological plausibility. This decision is because we do not require full biological plausibility for our use case, and the Leaky Integrate-and-Fire spiking neuron model is less complex than the Izhikevich model.

We applied backpropagation with temporal characteristics to update the weights after each epoch of the experiment. Biological spiking neurons operate with different learning methods, such as spike timing dependant plasticity (STDP) and its variants. This field is under constant research to determine how learning is performed between neurons; hence, there is no correct answer for using the learning method.

We applied surrogate gradients to improve backpropagation performance. In our case, we used the fast sigmoid surrogate gradient applied to the spiking neural network layers. We chose the Adam optimization algorithm [29] with $learning\_rate = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$ and cross-entropy loss [30] for calculating the value of loss.

The training lasted for 300 epochs, each epoch containing 10 discrete time steps as the temporal component. There were no optimizations regarding stop-loss values where training would halt. Training was performed by Intel i7-3630QM laptop CPU, without GPU usage. PyTorch version used was 2.1.0+cpu, and snntorch version used was 0.7.0.

IV. EXPERIMENT RESULTS

We executed the original experiment with 240 hidden neurons and 10 000 samples from the 5% BoT-IoT dataset with applied one-hot encoding transformations, resulting in an F1 score of 0.957. The loss curve is shown in Fig. 1. The noisiness shown inside loss curves is characteristic because the training was performed per epoch per set of discrete time steps. The loss gradient was calculated after each epoch, thus causing the noisiness inside every epoch.

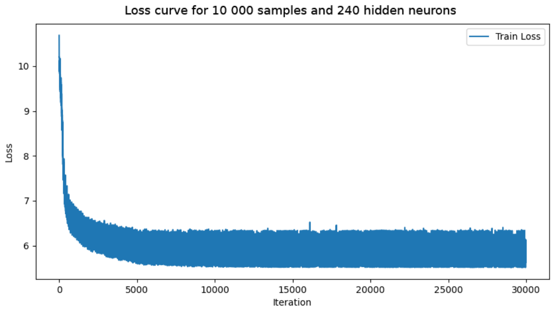


Fig. 1. Loss curve for 10 000 samples and 240 hidden neurons (two hidden layers of 120 spiking neurons)

ROC and PRC curves shown as Fig. 2. and Fig. 3 show additional experiment performance results. The area under Curve values is at the bottom for every attacking category.

Reconnaissance as the attack category was abbreviated to Recon due to brevity.

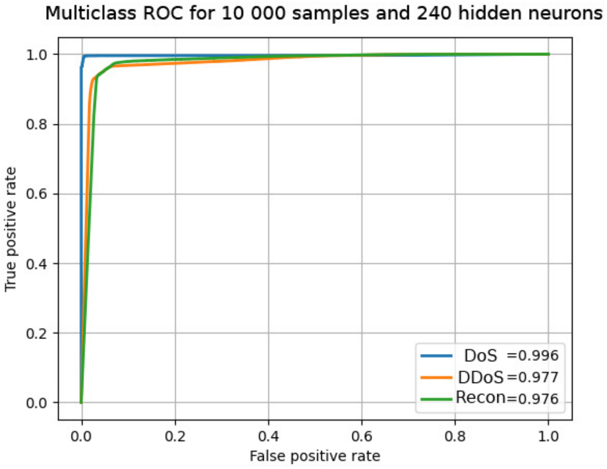


Fig. 2. Multiclass ROC curves for 10 000 samples and 240 hidden neurons (two hidden layers of 120 spiking neurons)

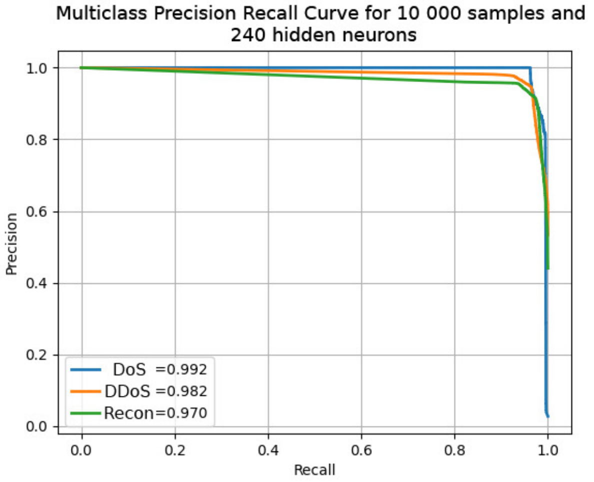


Fig. 3. Multiclass Precision-Recall curves for 10 000 samples and 240 hidden neurons (two hidden layers of 120 spiking neurons)

We have repeated the experiment with 1 000 samples and 72 spiking neurons to improve training and testing time. Our experiment yielded an F1 score value of 0.944.

The loss curve shown in Fig. 4 has a longer drop than the previous one with 10 000 samples. ROC (Fig. 5) and PRC (Fig. 6) curves show the impact of fewer samples and fewer neurons used for training, but showing similar results.

After promising results from experiments done with 1 000 and 10 000 samples and 72 and 240 hidden neurons, we have decided to start lowering the experiment variables, most notably sample size and hidden neuron count. Table V shows the results of previous experiments and additional ones performed with varying sample sizes and neuron counts.

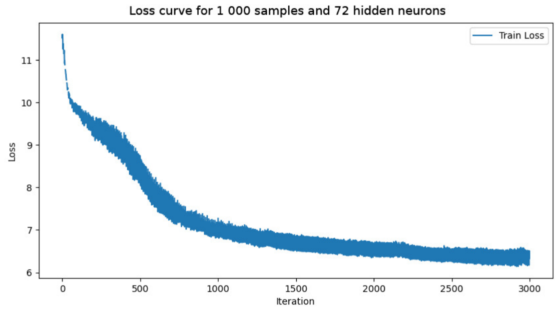


Fig. 4. Loss curve for 1 000 samples and 72 hidden neurons (two hidden layers of 36 spiking neurons)

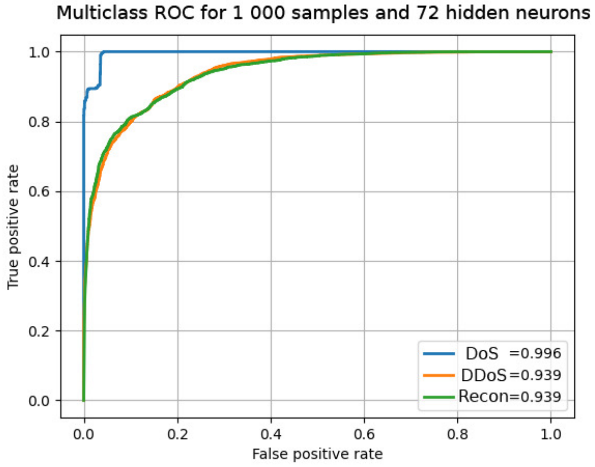


Fig. 5. Multiclass ROC curves for 1 000 samples and 72 hidden neurons (two hidden layers of 36 spiking neurons)

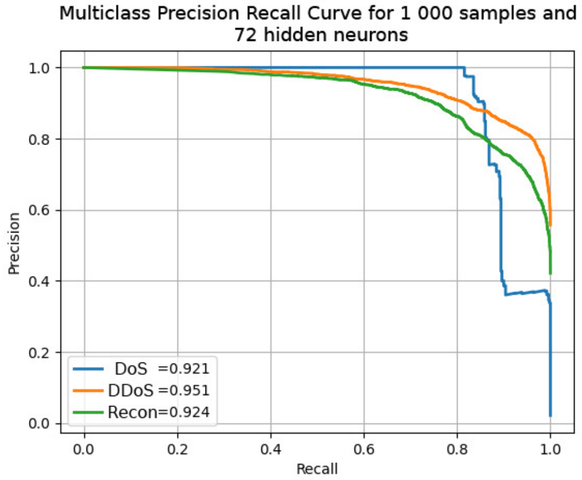


Fig. 6. Multiclass Precision-Recall curves for 1 000 samples and 72 hidden neurons (two hidden layers of 36 spiking neurons)

Sample sizes and neuron counts were randomly selected to see how the model would perform under new experimental conditions.

We can see that we get similar performance as state-of-the-art solutions with fewer samples and neurons that might apply to constrained systems. Previously described experiments with loss, ROC, and PRC charts are in bold as a baseline for other performed experiments that weren't presented in more detail.

TABLE V
SUMMARIZED EXTENDED EXPERIMENT DATA

Number of samples	Hidden neurons	Training time (seconds)	Testing time (seconds)	F1 score
500	20	58.730	0.091	0.910
500	240	103.546	0.149	0.949
1 000	20	145.789	0.085	0.915
1 000	**72**	**196.461**	**0.128**	**0.944**
1 000	240	231.894	0.159	0.959
10 000	72	1844.896	0.170	0.953
10 000	**240**	**2339.638**	**0.210**	**0.957**

Table VI below aggregates our experiment data with previously presented state-of-the-art data for comparison purposes.

TABLE VI
SUMMARIZED FINDINGS FROM CURRENT STATE

Paper	Accuracy
Wang et al. [19]	99.78%
Alom and Taha [20]	90.12%
Zhou et al. [21]	98.98%
Zarzoor et al. [22]	95%
Hassini et al. [23]	99.96%
1000 samples and 72 hidden neurons	95.99%
10 000 samples and 240 hidden neurons	96.33%

Following the experiment data and comparison with discovered state-of-the-art solutions, we can see that our two experimental models are comparative with state-of-the-art models, surpassing accuracy on some models, especially since our model is based on common hardware.

V. Discussion and future work

Experimental results from the previous section offer exciting insight into the applicability of spiking neural networks for intrusion detection systems and their behavior when the number of samples and neurons vary.

From our starting setup with 10 000 samples and 240 hidden neurons, we were able to lower the number of samples and neurons to 1000 samples and 72 neurons with similar F1 scores (0.957 versus 0.944, respectively). We performed additional experiments with even lower number of samples and hidden neurons that yielded F1 scores above 0.91 for all setups described in the tables above.

The first step should be to improve the computing resources used for experimentation and define stricter experimental criteria. Improvements would give us more detailed information on how the model performs with unconstrained sample sizes and diverse features and the inclusion of the Theft attack subcategory. Another task regarding computing resources would be accurate measurements of the spiking neural network to confirm resource efficiency, both on the original experiment environment and IoT devices. Experimental criteria should be stricter, and detailed experiment analysis should be performed on the behavior of false positives and false negatives, which can impact IoT devices differently due to their nature of being edge devices and with limited resources.

Besides resource increase, further research should be performed on different kinds of spiking neurons and different learning algorithms (such as STDP), as well as models with more than two hidden layers of spiking neurons.

Another interesting topic is the application of neuromorphic hardware such as Intel Loihi [31] or Graphcore's IPU processors [32] as integral components of the IoT device for intrusion detection and potential other artificial intelligence tasks that could be performed on-device. Utilizing neuromorphic hardware would improve total performance regarding accuracy and training times while allowing easier on-device intrusion detection.

VI. Conclusion

Internet-of-Things as a platform presents a new attack vector for malicious actors. Due to their decentralized and resource-constrained nature, performing adequate cyber attack detection and prevention without a centralized or significantly powerful device can be difficult.

We have introduced a method using spiking neural networks to enable intrusion detection classification in resource-constrained environments with cutting-edge performance. We performed testing using the BoT-IoT dataset, which includes a range of typical attacks found in Internet-of-Things environments and varying numbers of samples and hidden neurons. This allowed us to examine how F1 scores change as the number of samples and hidden neurons change, further optimizing performance with an F1 trade-off.

Even though we have reached state-of-the-art performance, we have presented additional steps in our research that can potentially improve performance by experimenting with different kinds of spiking neurons, more layers, and different learning algorithms.

References

[1] S. Li, L. D. Xu, and S. Zhao, "The internet of things: a survey," vol. 17, no. 2, pp. 243–259.

[2] L. S. Vailshery, "IoT connected devices worldwide 2022-2033."

[3] "Annual number of IoT attacks global 2022."

[4] D. Denning, "An intrusion-detection model," vol. SE-13, no. 2, pp. 222–232.

[5] T. Lunt, "Real-time intrusion detection," pp. 348–353. Conference Name: Digest of Papers. COMPCON Spring 89. Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage ISBN: 9780818619090 Place: San Francisco, CA, USA Publisher: IEEE Comput. Soc. Press.

[6] S. Shieh and V. Gligor, "A pattern-oriented intrusion-detection model and its applications," pp. 327–342. Conference Name: 1991 IEEE Computer Society Symposium on Research in Security and Privacy ISBN: 9780818621680 Place: Oakland, CA, USA Publisher: IEEE Comput. Soc. Press.

[7] Wenke Lee, S. Stolfo, and K. Mok, "A data mining framework for building intrusion detection models," pp. 120–132. Conference Name: 1999 IEEE Symposium on Security and Privacy ISBN: 9780769501765 Place: Oakland, CA, USA Publisher: IEEE Comput. Soc.

[8] C. A. Mead, "Neuromorphic electronic systems," vol. 78, no. 10, pp. 1629–1636. MAG ID: 2163630896 S2ID: 459c554583c9a2f70dd36e84149989fde1e9f833.

[9] E. Izhikevich, "Simple model of spiking neurons," vol. 14, no. 6, pp. 1569–1572.

[10] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," vol. 117, no. 4, pp. 500–544. MAG ID: 1985940938.

[11] K. F. Bonhoeffer, "Activation of passive iron as a model for the excitation of nerve.," vol. 32, no. 1, pp. 69–91. MAG ID: 2019082810 S2ID: d4a34157d41b45efe5622fb11d29995ed0f26b82.

[12] B. van der Pol Jun Docts. Sc. and J. van der Mark, "LXXII. the heartbeat considered as a relaxation oscillation, and an electrical model of the heart," vol. 6, no. 38, pp. 763–775. MAG ID: 2071313546 S2ID: 9de0580210474428aee2312db59263647c568337.

[13] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," vol. 79, no. 8, pp. 2554–2558. MAG ID: 2128084896.

[14] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," vol. 34, no. 10, pp. 1537–1557. Conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[15] M. S. Asghar, S. Arslan, and H. Kim, "A low-power spiking neural network chip based on a compact LIF neuron and binary exponential charge injector synapse circuits," vol. 21, no. 13, p. 4462.

[16] S. Kim, S. Park, B. Na, and S. Yoon, "Spiking-YOLO: Spiking neural network for energy-efficient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11270–11277. ISSN: 2374-3468, 2159-5399 Issue: 07 Journal Abbreviation: AAAI.

[17] S. Song, K. D. Miller, L. F. Abbott, and L. F. Abbott, "Competitive hebbian learning through spike-timing-dependent synaptic plasticity," vol. 3, no. 9, pp. 919–926. MAG ID: 1486852018.

[18] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," vol. 21, no. 6, pp. 335–346.

[19] T. Wang, G. Zhang, H. Rong, and M. J. Pérez-Jiménez, "Application of fuzzy reasoning spiking neural p systems to fault diagnosis," vol. 9, no. 6, p. 786.

[20] Z. Alom and T. M. Taha, "Network intrusion detection for cyber security using unsupervised deep learning approaches," pp. 63–69. MAG ID: 2790100928.

[21] S. Zhou, Shibo Zhou, Shibo Zhou, Xiaohua Li, Xiaohua Li, and X. Li, "Spiking neural networks with single-spike temporal-coded neurons for network intrusion detection.," ARXIV_ID: 2010.07803 MAG ID: 3093177876 S2ID: 70586ad226b9671b8c461f465850eff194eb726f.

[22] A. Zarzoor, N. Adnan, N. Al-Jamali, and D. Aldaloo, "Intrusion detection method for internet of things based on the spiking neural network and decision tree method," vol. 13, pp. 2278–2288.

[23] K. Hassini, S. Khalis, O. Habibi, M. Chemmakha, and M. Lazaar, "An end-to-end learning approach for enhancing intrusion detection in industrial-internet of things," vol. 294, p. 111785.

[24] N. Koroniotis, N. Moustafa, F. Schiliro, P. Gauravaram, and H. Janicke, "A holistic review of cybersecurity and reliability perspectives in smart airports," vol. 8, pp. 209802–209834. Conference Name: IEEE Access.

[25] N. Koroniotis and N. Moustafa, "Enhancing network forensics with particle swarm and deep learning: The particle deep framework."

[26] N. Koroniotis, N. Moustafa, and E. Sitnikova, "A new network forensic framework based on deep learning for internet of things networks: A particle deep framework," vol. 110, pp. 91–106.

[27] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, "Towards developing network forensic mechanism for botnet activities in the IoT based on machine learning techniques," in *Mobile Networks and Management* (J. Hu, I. Khalil, Z. Tari, and S. Wen, eds.), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 30–44, Springer International Publishing.

[28] N. Koroniotis, "Designing an effective network forensic framework for the investigation of botnets in the internet of things."

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization."

[30] D. R. Cox, "The regression analysis of binary sequences," vol. 20, no. 2, pp. 215–242. Publisher: [Royal Statistical Society, Wiley].

[31] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A neuromorphic manycore processor with on-chip learning," vol. 38, no. 1, pp. 82–99.

[32] G. , "IPU processors."

Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 79–88
DOI: 10.15439/2024F0005
ISSN 2300-5963 ACSIS, Vol. 40

QEDrants – Data Quality Quadrants for Business Users and Decision-Makers

Alina Powała[*][0009-0009-0268-3582], Dominik Ślęzak[*][†][0000-0003-2453-4974]

[*]QED Software, Mazowiecka 11/49, 00-052 Warsaw, Poland
Email: {alina.powala,dominik.slezak}@qedsoftware.com
[†]Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
Email: slezak@mimuw.edu.pl

***Abstract*—The adoption of artificial intelligence (AI) in business is often hindered by the complexity of data quality assessment. This paper introduces the quadrant-based data quality representation framework, which evaluates data assets based on two complementary dimensions: Data Integrity (accuracy and reliability, akin to Gartner's "Ability to Execute") and Data Coverage (breadth and comprehensiveness, similar to "Completeness of Vision"). The framework categorizes data into four groups: *Pure Gold* (AI-ready), *Sleeping Giants* (high integrity, low coverage), *Unpolished Diamonds* (high coverage, low integrity), and *Hitchhikers* (low integrity, low coverage). Each such quadrant provides actionable insights for business users, helping them prioritize data assets for AI readiness, identify data cleaning tasks, balancing costs and value realization by focusing on the right data. Given the roots of this idea in QED Software's technology experiences, we call the proposed quadrants as *QEDrants*.**

***Index Terms*—Data Quality, Data Integrity, Data Coverage, AI Readiness, Data Management, Decision Support, Cost Optimization**

I. Introduction

CAN businesses effectively manage data without constant reliance on data scientists? The purpose of this work is not to diminish the critical role of data scientists but to address the persistent gap between business stakeholders and technical teams. Even in organizations with strong artificial intelligence (AI) capabilities, this disconnect often leads to inefficiencies and misaligned goals. Bridging this gap requires establishing a common ground where business users can better understand technical concepts, and data scientists can align their work more closely with business objectives [1].

In particular, many organizations encounter the challenge of ensuring the quality of data, which is crucial for producing impactful AI outcomes. For non-technical stakeholders, assessing data quality is complex, making it hard to determine when data is ready to support business-critical AI applications. A part of the domain of AI refers to machine learning (ML)[1], wherein the challenges of poor data quality are especially well-understood. However, this problem is broader and does not refer only to the methods that we would call pure ML [2].

To address the above gap from the data quality perspective, we introduce the framework called *QEDrants*[2], which categorizes data assets into four groups based on two business-friendly dimensions: *Data Integrity* and *Data Coverage*.

Data Integrity corresponds to data accuracy and reliability. It assesses whether the data adheres to predefined rules and standards, such as valid formats (e.g., properly structured dates) and logical consistency (e.g., non-negative values in the age fields). High integrity ensures the data is free from errors and can be trusted for analysis and decision-making. Data Coverage, on the other hand, measures the completeness and comprehensiveness of the dataset. High coverage indicates that the dataset captures the full scope of the domain it describes, ensuring no critical information is missing. These two metrics are not opposing forces but rather complementary drivers of data quality. Together, they determine the data utility in delivering actionable insights and value. Just as both execution and vision are crucial for a business to thrive, integrity and coverage are essential for data to achieve its full potential.

We define QEDrants as follows:

- *Pure Gold*: High integrity and high coverage data, ideal for direct application in AI models.
- *Unpolished Diamonds*: High coverage but lower integrity data, representing assets that are rich in content but may need refinement for reliable AI use.
- *Sleeping Giants*: High integrity but low coverage data, indicating well-curated yet incomplete data sources that could benefit AI if augmented.
- *Hitchhikers*: Low integrity and low coverage data, representing low-priority assets that are generally unsuitable for AI applications.

The goal of QEDrants is to deliver the quadrant-style visualization with a clear, actionable view of data quality, highlighting areas where data can best serve AI objectives and where it requires improvement. This way, QEDrants can provide business users with a practical tool to prioritize data curation efforts, focusing on areas where investment will yield the greatest gains in AI readiness. The goal of QEDrants is also to emphasize the *value realization potential* of data, demon-

[1] Although AI and ML can be considered as two separate domains, in this paper – for simplicity – we use the acronym "AI" to cover both of them.

[2] Name inspired by Gartner's *Magic Quadrants*™ https://www.gartner.com/.

Topical area: Information Technology for Business and Society

strating how data assets contribute to usability, actionability, and value extraction – concepts aligned with the “5 V’s of Big Data” (volume, velocity, value, variety, veracity) [3].

The remainder of this paper is organized as follows. Section II discusses related work on data quality in AI, barriers in AI adoption, and existing data quality frameworks. Section III recalls broader inspirations and connections, including insights from Gartner’s frameworks, the concept of Total Cost of Ownership (TCO) for data processing systems, and related disciplines such as data governance and data security. Section IV introduces the conceptual and architectural background for the methodology used to define QEDrants. Section V presents case studies illustrating the application of our framework in various business contexts. Finally, Section VI concludes the paper.

II. Related Work

Data quality has been a long-standing area of research within AI, as the success of AI systems is closely tied to the validity of learning. In recent years, a substantial body of work has addressed the critical aspects of data quality in AI, identifying key challenges and proposing frameworks for evaluating data quality across different domains. However, while numerous approaches to data quality exist, many are tailored primarily for data scientists, leaving business stakeholders with limited accessibility to those methodologies.

A. Data Quality in AI

Research on data quality in AI emphasizes its pivotal role in ensuring reliable and unbiased outputs. Early work focused on core data quality dimensions such as accuracy, completeness, consistency, timeliness, and relevance [4], [5]. These dimensions remain foundational for assessing data quality in modern AI applications, as they directly affect model performance, interpretability, and the capacity to generalize. Several studies link data quality issues to AI model training and deployment challenges. Poor data quality can lead to model overfitting and inaccuracies, ultimately diminishing the value of AI insights for business stakeholders [6].

B. Barriers to AI Adoption

Despite the advancements in data quality research, barriers to AI adoption in business persist. They are attributed to the lack of accessible and interpretable data preparation and assessment frameworks. Non-technical users, including business managers and decision-makers, often lack the tools that are needed to assess data quality or improve data readiness for AI applications. Studies indicate that without straightforward methods for evaluating data, organizations face increased costs, prolonged implementation timelines, and potential failures in AI deployment due to suboptimal data preparation [7]. Moreover, traditional data quality frameworks typically emphasize technical dimensions without considering usability in business. This technical focus can lead to an “AI unreadiness,” where data quality needed for effective AI outcomes is not in place, resulting in limited confidence in AI systems.

C. Existing Frameworks

Several frameworks have been developed to provide a systematic approach to data quality evaluation, including Total Data Quality Management (TDQM) [8], Data Quality Assessment (DQA) [9], and others based on international standards like ISO/IEC 25012. These frameworks define comprehensive methodologies for assessing and improving data quality across dimensions (see [5] for a survey of approaches). However, they are geared towards data engineers and scientists, involving complex metrics and extensive data profiling procedures that may be cumbersome for business users. Furthermore, some recent frameworks include domain-specific quality models for healthcare, finance, and retail [10]. While these models add valuable insights into data quality needs for AI, they still tend to require high technical proficiency and they do not address the accessibility requirements of non-technical users.

D. Gaps in Business-Friendly Evaluation

The existing literature on data quality frameworks reveals a clear gap in models that are accessible to non-technical users and aligned with their business goals. As we have already discussed, this is part of a broader issue of the gap between business and AI specialists, which still exists even in the case of relatively large and mature companies. Traditional frameworks focus on rigorous technical assessment, which is essential in data science but lacks usability for stakeholders who lack deep technical knowledge. As a result, many organizations face challenges in bridging the gap between data engineering teams and business decision-makers.

To support AI adoption in business, there is a need for simplified, business-oriented frameworks that can help non-technical users understand and prioritize data quality issues [11]. Such a framework would empower business leaders to make informed decisions about data readiness for AI, minimizing technical barriers and accelerating AI adoption. QEDrants will address this gap as they are designed to be accessible to business stakeholders, facilitating the identification of data quality issues with minimal technical complexity.

III. Connections to Other Domains

The previous section focused on related work concerning the importance and measurement of data quality. This part expands the scope to explore broader inspirations. Although the areas considered below are not directly tied to data quality, they intersect in meaningful ways, influencing and shaping one another within the data management ecosystem.

A. Gartner’s Magic Quadrants

Gartner is widely recognized for its proprietary methodologies, including the concept of Magic Quadrant$^{\text{TM}}$ which is famous primarily because it provides a clear, visual framework for comparing technology providers in various industries, simplifying the decision-making process for businesses (see e.g. Fig. 1). It breaks down complex market analyses into a simple, two-axis chart, categorizing vendors into four types –

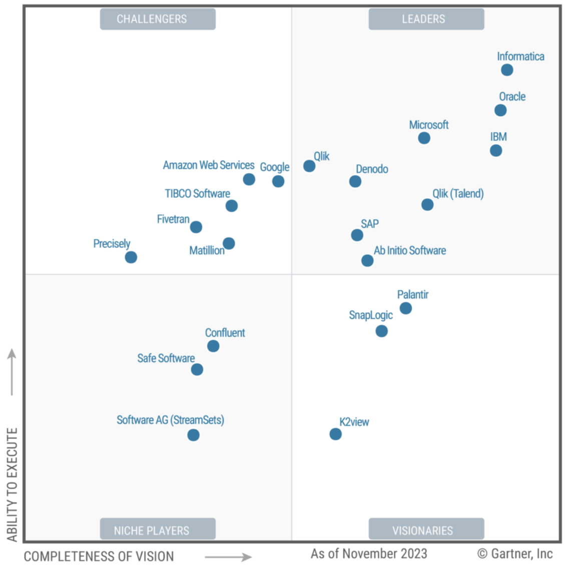


Fig. 1: Gartner's Magic Quadrant$^{\text{TM}}$ for data integration tools (https://www.informatica.com/content/dam/informatica-com/en/image/misc/data-integration-magic-quadrant-2023.png)

Leaders, Challengers, Visionaries, and Niche Players. The vertical axis, "Ability to Execute," evaluates a vendor's capacity to deliver on its promises, including product quality, customer support, and financial performance [12]. The horizontal axis, "Completeness of Vision," assesses a vendor's understanding of current and future market dynamics, innovation, and alignment with customer needs. This clarity makes it easier for companies to determine the competitive landscape at a glance.

We want QEDrants to leverage a two-axis model too. In our case, the focus is on Data Integrity and Data Coverage – two forces that work together to assess the data readiness for AI applications. Unlike Gartner's model, which primarily evaluates vendor performance, QEDrants apply these dimensions to data quality, offering a unique perspective on how organizations can use and improve their data to support AI initiatives. In this context, "Ability to Execute" from the Magic Quadrant$^{\text{TM}}$ framework aligns conceptually with Data Integrity. Just as "Ability to Execute" reflects a vendor's capacity to deliver on promises, Data Integrity measures the reliability and accuracy of the data, ensuring it is fit for purpose. Furthermore, "Completeness of Vision" corresponds to Data Coverage. In the Magic Quadrant$^{\text{TM}}$ model, "Completeness of Vision" means a vendor's forward-looking strategy and understanding of market trends. In QEDrants, Data Coverage assesses the comprehensiveness and representativeness of the data, ensuring it captures all necessary dimensions for effective AI deployment.

These two dimensions – Data Integrity and Data Coverage – are not opposites but rather complementary forces. Together, they provide a holistic view of data quality, ensuring organizations can trust their data and rely on its breadth. To our best knowledge, no Gartner-inspired quadrant visualization has been applied specifically to data quality assessment. While Gartner has utilized similar visual frameworks for evaluating technology platforms and AI solutions, the adoption of such tools for visualizing data quality metrics, like Data Integrity and Data Coverage, remains unexplored.

B. *Total Cost of Ownership*

Total Cost of Ownership (TCO) in IT encompasses several cost components like (1) system design and infrastructure costs (the initial setup of data processing systems, computational resources, and storage), (2) maintenance and human resource costs (regular system upkeep, troubleshooting, and personnel expenses), (3) user operation costs (e.g., for database engines and business intelligence tools, this includes query execution time, latency, and handling approximate results [13]), and (4) costs of re-engineering, including costly redesigns of poorly modeled data systems when the user demands evolve.

In AI, ensuring high-quality data is a critical factor in the TCO of deploying models in business environments (see [14] for a robust classification of data quality costs). Poor data quality can significantly increase operational and business costs throughout the AI lifecycle. These costs manifest in several ways: (1) Low-quality data can lead to poorly trained models, requiring additional iterations of training and validation. This increases computational costs and prolongs deployment timelines. (2) Post-deployment, models operating on low-quality inference data are more likely to trigger monitoring alerts. These alerts necessitate frequent investigations, potentially leading to model re-tuning. (3) Errors stemming from data quality issues – whether in training data, inference data, or both – can result in business losses. Incorrect model outputs may harm customer satisfaction, operational efficiency, or decision-making accuracy, directly affecting the bottom line.

With large language models (LLMs) becoming a "hot topic," understanding their TCO is increasingly important. Measuring data quality for LLMs, including the evaluation of training and inference data, is an emerging challenge. The costs of maintaining high-quality data for such models are substantial, given their reliance on vast and diverse datasets.

Our previous research highlights the importance of diagnostic tools for AI models, as discussed in [15]. These tools help identify model errors, some of which may be rooted in data quality issues. Such diagnostics are valuable for pinpointing problems in both training and inference data. However, even the most advanced diagnostic systems have limitations and cannot identify all potential errors. Thus, investing in robust data quality analysis from the outset remains essential.

While poor data quality means significant problems, efforts to improve it are not without their own financial and operational implications. Within AI-infused data processing pipelines, additional costs of data enhancements emerge:

- External data. High-precision external data can be expensive, particularly for use cases requiring customer data, detailed measurements, or enriched metadata.
- Advanced parsing and quality enhancement tools. These tools improve data accuracy but at the same time, increase computational costs and latency.

- Human-involved data labeling and curation. Active learning and interactive tagging approaches, such as those explored in [16], involve human experts in data improvement processes. While effective, these methods vary in cost depending on the level of investment, such as using multiple experts for higher accuracy.

Effectively managing these costs is essential to optimizing the TCO for AI deployments, as both underinvestment and overinvestment in data quality can compromise the overall value and efficiency of AI solutions in applications.

The success of AI projects can be multifaceted, encompassing technical, ethical, and societal dimensions. Unlike traditional IT projects, AI initiatives involve unique challenges due to the complexity of algorithmic decision-making and its far-reaching impacts (see a recent study [17] for a review of AI success factors within the project management literature). A crucial metric for assessing the success of AI deployments is the return on investment (ROI), directly tied to the balance of investment in data quality and the value derived from AI solutions. Achieving a positive ROI depends on ensuring that the costs associated with improving Data Integrity and Data Coverage are justified by the benefits these improvements bring to AI performance and business outcomes.

Measuring ROI for AI involves evaluating quantifiable gains, such as cost reductions and revenue increases, and intangible benefits, including improved customer experiences, enhanced decision-making speed, and competitive positioning. Companies should monitor and adjust their data quality investments to ensure that the total cost of ownership is optimized, and the expected ROI is achieved or exceeded.

By visualizing data quality through QEDrants, business users will make informed decisions about which data sources to improve, ignore, or prioritize. This targeted approach helps organizations allocate resources efficiently, ultimately optimizing TCO. Once these decisions are made, systems (like e.g. BlueQuail developed by QED Software[3]) can operationalize them, offering guidance on feasible data improvement strategies. Additionally, integrating active learning techniques ensures a balance between data quality and human resource costs, optimizing the overall investment in data curation.

C. Data Governance and Security

Data governance is a critical yet expansive topic, often considered a cornerstone of effective data management. It encompasses the frameworks, policies, and procedures that ensure the data is managed as a valuable asset, aligning with organizational goals and regulatory requirements. Data governance is closely tied to data quality, as poor governance can lead to inconsistencies, inaccuracies, and compliance risks.

A key aspect of data governance is enabling business users to play an active role in data management. Traditionally, governance has been the domain of IT and data management professionals, but involving business stakeholders can bridge the gap between technical data policies and business needs. By equipping business users with tools like QEDrants, organizations can democratize data governance, allowing non-technical stakeholders to assess and influence data quality proactively.

Data security, though sometimes overlooked, is an equally important consideration in the context of AI and data quality. In business applications, security concerns frequently arise when sensitive data must be sent to external AI modules or third-party services. To mitigate the risks, organizations often anonymize or obfuscate the data before sharing it. However, this process can degrade data quality, introducing a trade-off between maintaining privacy and ensuring data reliability.

This trade-off was explored in [18], where the data was deliberately “corrupted” for business reasons, demonstrating the impact of security measures on data utility. Similarly, anonymization becomes particularly relevant when AI model development is outsourced to external firms, a common practice observed e.g. at QED Software[4]. Outsourcing can shift to crowdsourcing in competitive scenarios like those hosted on platforms such as knowledgepit.ai. A notable example is presented in [19], where sensitive communication data was stripped of its content to ensure privacy, rendering sentiment analysis infeasible. This highlights the broader challenge faced by all crowdsourcing platforms, including Kaggle, where data anonymization can limit the scope of achievable insights.

In both outsourced and crowdsourced AI projects, maintaining a balance between data security and data quality (which implies AI readiness) is crucial. Future iterations of the QEDrant framework may explore this trade-off, providing business users with visual tools to assess the impact of security measures on Data Integrity and Data Coverage.

IV. QEDrant Framework

This section lays the groundwork for understanding the QEDrant framework, focusing on its structure, core components, and core functionalities. We begin by addressing the foundational mechanics and metrics that drive the framework. Next, we shift to the user perspective, exploring how to interact with the framework and interpret its outputs. Finally, we revisit the foundations to consolidate key insights.

The QEDrant framework is a structured approach to assessing data quality, designed to help business users quickly understand the readiness of their data for AI applications. The framework organizes data assets into four quadrants based on two key metrics – Data Integrity and Data Coverage – allowing users to evaluate data reliability and completeness at a glance. The data quality analysis has a subsequent goal of recommending actions for data improvement or enrichment to better support AI. This is done without referring to any specific AI model. Instead, the framework provides foundational insights into data quality, helping users recognize the value and limitations of their data for future AI applications.

A. Data Integrity and Data Coverage

The QEDrant framework is grounded in established theories of data quality management, drawing on key metrics such as

[3] https://bluequail.ai/

[4] https://qedsoftware.com/

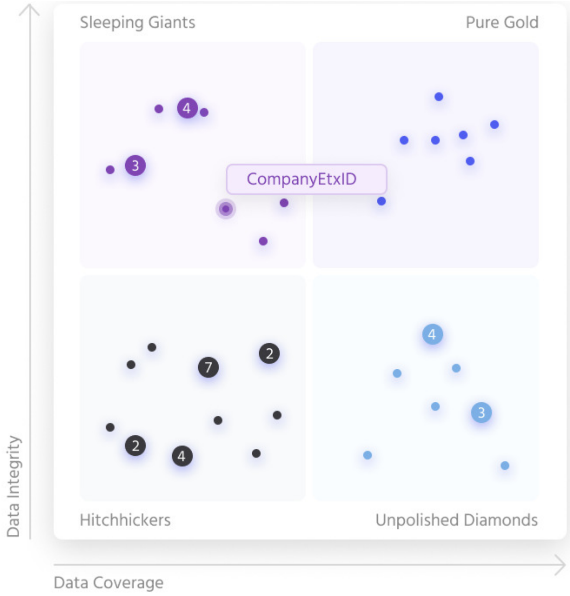


Fig. 2: QEDrant inspired by Gartner's Magic Quadrants^{TM}. It visualizes data assets based on Data Integrity and Data Coverage. Each quadrant (Sleeping Giants, Pure Gold, Unpolished Diamonds, Hitchhikers) categorizes the data due to its readiness and suitability for AI, providing an actionable overview for business users to prioritize data quality improvements.

accuracy, completeness, consistency, and relevance [4], [5]. As we already know, these dimensions are translated into two aggregate metrics within this framework: Data Integrity and Data Coverage. In the current version of the framework, these metrics are implemented for tabular data formats.

Data Integrity. Derived from the concepts of accuracy and consistency, it evaluates how closely the data aligns with expected semantic types and domain rules. It measures the reliability and error-free nature of the data, ignoring null values. More about the evaluation components:

- Data validity: Whether the data in a column adheres to defined business rules or domain constraints (e.g., no negative values in an age column).
- Data consistency: Consistent data format across the column (e.g., uniform date format). This metric is calculated per field and then aggregated across fields to provide a Data Integrity score for each table.

Data Coverage. It assesses whether the data is not too sparse. In the simplest version, one can think about it as the percentage of non-null values. However, a more advanced analysis of semantic types of missing values is required in future [20].

In Subsection IV-E, we will elaborate on how to estimate Data Integrity and Data Coverage in a more sophisticated way. However, we want to keep information as simple as possible for business users. Therefore, more advanced methods will need to come up together with their intuitive explanations.

B. *Two Levels of Granularity*

QEDrants operate across the following levels. Level 1 assesses the overall quality of data tables by aggregating metrics for Data Integrity and Data Coverage across columns. Level 2 goes deeper, analyzing individual columns in a table. These scores are visually represented in a QEDrant diagram (see Fig. 2), allowing users to quickly see where their data stands from the perspective of usefulness and reliability. This is a straightforward categorization of data quality.

Level 1: Table-Level Assessment. Each table is assigned a Data Integrity and Data Coverage score, derived by aggregating the basic metrics across all columns. This provides a high-level view of the table's suitability for AI. It enables users to prioritize tables for refinement or immediate application.

Level 2: Column-Level Assessment. For every column, the framework calculates its reliability and consistency based on validation criteria (e.g., adherence to semantic types or domain constraints). Intermediate outputs such as unique identifiers (UIDs), semantic types, and time-related columns are identified to provide a more granular understanding of data quality.

We refer to [21] for a vision of a richer hierarchy of granularities that can be useful for analyzing data quality.

C. *QEDrant Representations*

This subsection serves as a glossary for business users, providing a comprehensive overview of the QEDrant framework's key elements and functionalities. We explain the user interface (UI) features that make QEDrants accessible and actionable for non-technical stakeholders. Fig. 2 provides a visual reference. (For further study on visual navigation through QEDrants, we refer to [21] again.) We identify the following quadrants:

Pure Gold quadrant represents data with high integrity (accurate, reliable, consistent) and high coverage (comprehensive, minimum number of missing values). **Importance.** Pure Gold data is well-suited for high-stakes AI applications where accuracy and coverage are essential, such as predictive modeling and decision support systems. **Examples of Use.** Pure Gold data can be deployed immediately in AI, providing reliable insights with minimal risk of errors or biases.

Sleeping Giants represent high integrity but low coverage data, indicating that the data is accurate and consistent but has gaps or missing entries. **Importance.** This data may lack sufficient coverage for comprehensive analyses but is valuable in applications requiring precision and reliability within a limited scope. **Examples of Use.** Sleeping Giants are ideal for pilot AI projects or initial proof-of-concept models where accuracy is paramount, but complete coverage is not a requirement.

Unpolished Diamonds represent high coverage but low integrity data, suggesting the data is complete but may contain errors or inconsistencies. **Importance.** While suitable for exploratory analysis or feature discovery, Unpolished Diamonds require refinement before being applied to critical AI tasks. **Examples of Use.** This data can support early-stage analysis where the breadth of the data is valued over precision.

Hitchhikers represent low integrity and coverage, indicating the data that is both incomplete and potentially inaccurate or inconsistent. **Importance.** Hitchhikers are generally unsuitable for direct AI applications but could provide some value in non-critical exploratory tasks or after significant data enrichment and cleaning efforts. **Examples of Use.** This data may be useful for supplementary analyses, or in cases where additional cleaning can bring it up to a higher standard of usability.

D. Business Relevance

The QEDrant framework helps business leaders visualize data quality at a glance, empowering them to make informed decisions about AI readiness and data improvement tasks. By categorizing data assets into actionable quadrants, the framework enables organizations to:

- Leverage AI-ready assets (Pure Gold) immediately, maximizing ROI as discussed in Subsection III-B.
- Identify targeted analyses and data collection needs (Sleeping Giants) for precise insights.
- Prioritize data cleaning tasks (Unpolished Diamonds) to unlock valuable data potential.
- Avoid unnecessary costs on low-value data (Hitchhikers), optimizing TCO and resource allocation.

This approach not only clarifies data priorities but also aligns data quality efforts with business goals, reducing the barriers to effective AI adoption. For non-technical users, QEDrants provide a structured framework to guide data management strategies, helping them realize measurable business outcomes without deep technical expertise. This way, companies can confidently advance their data assets from potential to performance, setting a solid foundation for AI success.

E. More about Metric Derivations

Finally, let us go back to the problem of computing the Data Coverage and Data Integrity measures at particular levels of granularity. While the current version of QEDrants employs relatively simple methods, we acknowledge the potential for refinement and expansion. This subsection outlines a roadmap for improving these calculations, ensuring they better serve the practical goals described in subsequent sections.

Expanding to multimodal data. Future QEDrant releases should account for images, text, logs, etc. For such datasets traditional notion of a column does not apply. Instead, each modality (e.g., a camera feed, text document, sensor reading) may be treated as an independent "field." At Level 1, Data Coverage and Data Integrity may be computed separately for each modality, while Level 2 would aggregate them across modalities to provide a comprehensive view.

One possible approach is to transform raw multimodal data into intermediate vector or tensor representations. These representations, derived using appropriate tools for data transformation [22], may be evaluated using classical metrics. Multiple versions of these transformations might be sampled, with metrics averaged across them. These intermediate steps would remain hidden from users, unless they specifically request an explanation through an Explainable AI module.

Leveraging advanced learning techniques. To enhance metric accuracy, we propose more sophisticated, learning-based methods for estimating Data Coverage and Data Integrity:

- Feature selection approaches. Inspired by feature selection, we may dynamically generate hypothetical target variables based on the dataset's semantic context. For each target variable, quality measures for fields (or modalities) could be evaluated using filter-based or model-based methods. By averaging these results across various target variables, a more nuanced estimate of field quality may be obtained.
- Data and entity matching. Another promising direction involves leveraging a repository of historical datasets with validated Data Coverage and Data Integrity scores. By matching new datasets to similar historical datasets (using entity matching techniques), we can infer quality levels for new data. This approach would allow the system to "learn" from past data and apply those insights to new, incoming datasets.
- Interactive learning with expert feedback. Interactive learning can refine the framework. Starting with basic calculations, the framework can present edge cases to domain experts. Their feedback can be used to fine-tune the quality assessment models, gradually incorporating richer, more accurate metrics. Over time, these interactions can enable our framework to adapt and improve its recommendations. For a deeper discussion on active learning methodologies, see [16].

Towards continuous improvement. Even in production, user feedback plays a crucial role in improving the system. If business users disagree with QEDrant classification, their corrections can be fed back into the model for retraining, enabling continuous improvement. This feedback-driven approach ensures that Data Coverage and Data Integrity metrics evolve alongside the changing needs and contexts of the organization. By refining these metrics and incorporating advanced methods, QEDrants can provide more precise and actionable insights across a wide range of data types and use cases.

The above roadmap not only enhances the technical robustness of the framework but also ensures it remains adaptable to emerging challenges, such as multimodal datasets and evolving business requirements. While this framework is designed to analyze data independently of any specific AI application, future work can also extend its capabilities to recommend tailored data enrichment or cleaning actions based on AI project needs. Moreover, more advanced evaluation metrics and aggregation methods are envisioned in subsequent phases, aligning the framework more closely with specific AI objectives.

V. QEDRANT APPLICATIONS

By categorizing the data into four quadrants based on Data Integrity and Data Coverage, the QEDrant framework provides actionable guidance for each data asset. This section details practical use cases for each quadrant, illustrating how they can inform data strategies and improve decision-making. To give a comprehensive view, the section is structured as follows:

- Subsections V-A-V-D: Real-world examples of how each quadrant guides immediate business actions.
- Subsection V-E: Advanced scenarios – exploring how companies may push the boundaries of AI adoption.
- Subsection V-F: Integrations – embedding QEDrants into larger AI ecosystems to maximize their impact.

Each subsection is designed to help organizations leverage the QEDrant framework not only as a diagnostic tool but also as a driver for strategic improvements in data readiness.

A. Pure Gold

Practical Application: Data assets categorized as Pure Gold are immediately suitable for AI applications. This data can be confidently utilized in high-stakes AI initiatives such as predictive analytics, fraud detection, and customer segmentation.

Guidance for Use: The primary strategy with Pure Gold is to harness its rich and high-value data for strategic decision-making, focusing on areas where immediate, actionable insights can drive significant impact.

AI Readiness: Pure Gold data assets require minimal preparation. Their high quality supports reliable AI training, reducing the risk of errors and allowing for fast deployment.

Measurable Business Outcomes: Using Pure Gold data improves decision-making and operational efficiency. Examples:

- Customer segmentation: Accurate targeting enhances marketing ROI and customer engagement.
- Fraud detection: High data quality reduces false positives, minimizing financial losses.
- Predictive maintenance: Reliable performance data enables more accurate predictions, reducing downtime and optimizing resources.

For non-technical users, Pure Gold means “AI-ready” assets, allowing them to proceed with confidence and minimal effort.

B. Sleeping Giants

Practical Application: Sleeping Giants are accurate and reliable but incompleteness may restrict their usage in comprehensive analyses. They may be well-suited for limited or targeted analyses, where precision is more important than breadth.

Guidance for Use: The primary strategy is to leverage high integrity for targeted insights, but also identify areas where additional data collection could expand usefulness. For example, a retail company might use Sleeping Giants data to analyze customer behavior in a specific region, and then supplement it with new data collection efforts for broader insights.

AI Readiness: Sleeping Giants data is suitable for proof-of-concept models or analyses focused on precise questions within a limited scope. Organizations can proceed with smaller-scale AI initiatives, assessing the initial utility of the data while planning for future data enrichment.

Measurable Business Outcomes:

- Market analysis in targeted segments: Precise insights for specific demographics or regions, reducing the cost of large-scale data collection.
- Feature testing for product development: Using accurate but limited data to aid efficient R&D.

Non-technical users can leverage Sleeping Giants data for “targeted insights now, broader potential later.” This approach offers immediate value and provides a clear path for further data collection if more comprehensive analyses are desired.

C. Unpolished Diamonds

Practical Application: Unpolished Diamonds datasets can be regarded as comprehensive but in the same time they may contain errors or inconsistencies. This quadrant is ideal for identifying data cleaning tasks that can elevate its quality, making it suitable for more robust AI applications in future.

Guidance for Use: For Unpolished Diamonds, the focus should be on data cleaning and validation to improve data integrity. This includes tasks such as correcting inconsistencies, filling in missing values where possible, and standardizing data formats. Once cleaned, this data can transition to the Pure Gold quadrant, making it highly valuable for AI applications.

AI Readiness: Unpolished Diamonds are not immediately AI-ready but offer potential once data cleaning tasks are performed. These data assets can support exploratory analyses and feature discovery during the initial stages, but they should undergo refinement before being used in critical AI models.

Measurable Business Outcomes:

- Improved exploratory analysis: Cleaning the data enhances the reliability of trend analysis and feature discovery, making initial insights more trustworthy.
- Preparedness for advanced AI applications: Data cleaning converts Unpolished Diamonds into AI-ready assets, increasing the ROI of the data asset over time.

Non-technical users see Unpolished Diamonds as “the data with potential.” Data cleaning can unlock this potential, transforming these assets into reliable resources for AI.

D. Hitchhikers

Practical Application: Hitchhikers are unsuitable for immediate AI use. This data typically requires significant effort to clean and augment, which may not be worth the investment relative to its potential value.

Guidance for Use: Given their low quality, Hitchhikers should generally be deprioritized to avoid unnecessary costs. Limiting efforts on these data assets helps reduce the TCO associated with data preparation and maintenance. In cases where Hitchhikers data holds specific or supplementary value, it can be revisited for enhancement later, but for most business needs, focusing on other quadrants yields better returns.

AI Readiness: Hitchhikers data is generally not AI-ready and should not be prioritized for immediate usage. These assets can be kept as optional but are unlikely to directly support critical AI applications without substantial improvement.

Measurable Business Outcomes:

- Cost savings: By limiting efforts on Hitchhikers data, one can focus resources on higher-value data assets, reducing the TCO associated with data preparation.
- Focused data strategy: Deprioritizing low-quality data allows organizations to concentrate on assets that are

more likely to yield actionable insights, increasing the efficiency of data-related investments.

Hitchhikers are “not worth the investment right now.” Limiting resources spent on Hitchhikers helps streamline data strategy and focuses attention on more promising assets.

E. Making QEDrant-based Decisions

QEDrants provide users with data quality visualization, enabling them to make informed decisions impacting their data strategy and AI projects. While earlier subsections laid the groundwork for interpreting QEDrant diagrams, this part delves into more advanced scenarios where business stakeholders leverage QEDrant insights to drive key decisions. Below, we explore various decision-making contexts.

Investing in extra data sources. One common scenario is to identify gaps in Data Coverage or Data Integrity that could hinder the success of an AI project. For example, a dataset in the Sleeping Giants quadrant may signal the need for additional sources to improve coverage. Users may decide to:

- Purchase external datasets (e.g., market trends or demographic data).
- Enhance data collection (e.g., gathering more detailed customer feedback or expanding survey outreach).

Such investments aim at improving data completeness, ensuring that models trained on these datasets achieve broader applicability and higher performance.

Improving data generation and processing. Data assets classified as Unpolished Diamonds (high coverage, low integrity) suggest that while sufficient data exists, its quality is compromised by errors or inconsistencies. In this case, QEDrant analysis might prompt business users to:

- Enhance data parsing or transformation pipelines to improve accuracy.
- Implement stricter validation rules or automate error detection mechanisms.

For instance, a company relying on web-scraped data might identify parsing errors causing misclassification or duplication. Addressing these issues would improve data reliability, which in turn supports more robust AI models.

Reevaluating AI project goals. QEDrant insights can lead to strategic shifts in AI objectives. For example, if a dataset supporting an AI classification task is predominantly in the Hitchhikers quadrant, then business users may decide to:

- Reduce the task’s complexity, e.g. moving from 500 decision classes to 50 more generalized ones.
- Adjust accuracy expectations based on current data limitations, moving from a target of 95% accuracy to a more achievable 90%.

These adjustments allow for more realistic project goals, aligning expectations with the available data capabilities.

Deciding on project continuation or pivot. When a significant portion of critical datasets fall into problematic quadrants, such as Hitchhikers or Unpolished Diamonds, business users might face a more fundamental question: Is the project viable? Based on QEDrant insights, they may:

- Decide to halt the project until data quality improves.
- Pivot the project focus to areas where higher-quality data is available.

For example, an AI project initially designed to predict customer churn may be shifted toward identifying high-value customers if the churn-related data proves insufficient in quality.

Trade-offs and cost-benefit analysis. QEDrant diagrams also help users navigate trade-offs between data quality dimensions and project requirements. Consider the following:

- Time versus Quality: Should the project proceed with current data quality to meet deadlines, or is it worth delaying for data improvement efforts?
- Cost versus Accuracy: Would investing in high-quality data sources justify the incremental improvement in model performance?

These trade-offs are particularly relevant for projects where small quality gains come at a high cost, enabling business users to evaluate the ROI of data enhancement efforts.

Adjusting model complexity or evaluation metrics. Another scenario is to adjust the complexity of the AI model or the metrics by which its performance is evaluated. Examples:

- For datasets with lower integrity, shifting from precision-oriented metrics (e.g., precision/recall) to more robust metrics like F1-score or Matthews correlation coefficient might be advisable.
- Simplifying model architectures to reduce sensitivity to noisy or incomplete data, which can still provide actionable insights with reduced computational costs.

Collaboration and resource allocation. QEDrant insights also aid in optimizing cross-team collaboration. For instance, datasets requiring significant improvement might warrant additional resources from IT, data engineering, or third-party vendors. By identifying and prioritizing data assets based on their quadrant classification, organizations can allocate resources more effectively, focusing on the most critical datasets first.

In summary, QEDrants can provide a foundation for strategic decision-making across a variety of business and technical contexts. From data acquisition and pipeline optimization to project goal revision and resource allocation, QEDrant analysis empowers users to make data-driven choices that balance data quality, project feasibility, and business impact. Going further, we will explore specific examples of these advanced applications, demonstrating how QEDrant insights translate into actionable strategies for optimizing AI projects.

F. Deployments and Integrations

To maximize their utility, QEDrants must operate as part of a broader data ecosystem, seamlessly connecting with other modules and systems to drive actionable insights. This subsection explores how QEDrants can integrate with existing data infrastructure, support decision implementation, and potentially evolve into a recommendation engine capable of suggesting data quality improvements.

Interfacing with other modules. QEDrants should not function in isolation. Instead, they must interface with various components of the data pipeline, including:

- Data ingestion and transformation pipelines. Once a QEDrant analysis identifies data quality issues, the system can trigger automated data cleansing or transformation processes in connected ETL pipelines.
- Monitoring and diagnostic tools. QEDrants can feed data quality insights into monitoring systems to flag potential issues affecting model performance.
- AI model training modules. Data quality metrics provided by QEDrants can inform model training, helping select the most reliable datasets or identifying areas where synthetic data augmentation may be beneficial.

By integrating with these modules, QEDrants enable a feedback loop where data quality improvements translate directly into enhanced model performance and business outcomes.

Supporting decision implementation. A key aspect of QEDrants is to translate analysis into action. When users decide to address specific data quality issues – e.g. enhancing Data Coverage or correcting parsing errors – the framework should support seamless implementation. This can involve:

- Task automation. Automatically initiating data quality improvement tasks, such as filling missing values using imputation techniques or applying stricter validation rules during data ingestion.
- Workflow integration. Creating tickets in project management tools (e.g., Jira, Asana) to involve relevant teams (e.g., data engineering) in quality improving.
- Collaboration with external vendors. If external data sources are required, QEDrants can generate detailed procurement requirements based on identified gaps in Data Coverage or Data Integrity.

Recommendation engine potential. To further enhance its utility, the QEDrant framework may evolve into a recommendation engine, autonomously suggesting data quality improvement actions. This requires several key capabilities:

- Data-driven recommendations. QEDrants can analyze historical data quality improvement efforts and their outcomes, learning which interventions are most effective for specific types of data quality issues.
- External data. By integrating with external repositories and APIs, QEDrants can identify new data sources that may fill coverage gaps or improve data reliability.
- Predictive analytics. We can predict the potential impact of suggested improvements on AI performance and business gains, helping users prioritize actions.

Thus, by interfacing with other modules, supporting decision implementation, and evolving into a recommendation engine, QEDrants can not only identify data quality issues but also drive actionable, automated improvements.

VI. Conclusions and Future Directions

The QEDrant framework presents a structured approach for businesses to assess and prioritize their data assets in preparation for AI adoption. By categorizing data into four actionable quadrants, this model enables organizations to understand their data readiness for AI without requiring deep technical expertise. The framework guides non-technical users in identifying AI-ready data, determining areas for targeted analysis, prioritizing data cleaning tasks, and managing costs by deemphasizing low-value data assets.

The framework's strength lies in its ability to simplify complex data quality assessments, helping business leaders make informed decisions about data curation and improvement. This quadrant-based approach ultimately empowers companies to approach AI adoption with clarity and confidence. It lowers the barrier to AI by translating data quality dimensions into actionable insights for non-technical stakeholders, aligning data efforts with business goals. As a foundational step, QEDrants establish a roadmap for data quality improvement that can evolve with a company's AI maturity, supporting more advanced data strategies and AI applications over time.

As we continue to refine the QEDrant framework, several avenues for future work emerge, ranging from enhancing core functionality to exploring entirely new concepts. Below, we outline key directions for future R&D, some of which have been briefly mentioned in earlier sections.

Expanding data modalities. One significant area of future work is to extend QEDrants to support multimodal data. Currently, the framework focuses on tabular data, but many real-world AI applications rely on a mix of data types, including images, text, and sensor data. In future, we intend to:

- Develop methods for assessing Data Integrity and Data Coverage across different modalities.
- Introduce modality-specific metrics (e.g., resolution for images, semantic coherence for text).
- Enable users to view data quality metrics for individual modalities or entire multimodal datasets.

This expansion will introduce QEDrants to a wider range of industries, e.g., healthcare, autonomous systems, media.

Advanced methods for calculating data quality metrics. While the current approach to computing Data Integrity and Data Coverage relies on straightforward aggregation methods, more sophisticated techniques can improve accuracy and applicability. Potential directions of improvement include:

- Predicting data quality metrics, especially when direct calculations are infeasible or incomplete.
- Using dynamic target variables and feature selection to better evaluate the relevance of specific fields.
- Incorporating historical data repositories to infer quality metrics for new datasets based on their similarity to previously validated data.

These advancements would enable more precise assessments, especially for complex or evolving datasets.

Recommendations for data improvement. As discussed earlier, transforming QEDrants into a recommendation engine can significantly enhance their utility. Examples of future work:

- Developing algorithms that suggest targeted actions, such as acquiring new data sources, automatic data cleaning, or modifying AI project objectives.

- Integrating external data sources and metadata repositories to provide context-aware recommendations.
- Evaluating the impact of recommended actions through predictive analytics, helping users prioritize improvements based on expected business outcomes.

Interactive and adaptive learning. QEDrants may incorporate interactive learning mechanisms to continuously refine their assessments and recommendations. This may include:

- Collecting feedback from users on the accuracy and usefulness of QEDrant outputs.
- Employing active learning techniques to engage domain experts in reviewing edge cases, gradually improving the system's understanding of data quality.
- Implementing a closed-loop feedback system, where users' inputs directly influence future iterations.

Such capabilities would ensure that QEDrants remain responsive to user needs and evolving data environments.

Addressing trade-offs in data quality. Future work may also explore more nuanced trade-offs between data security and data quality, as already highlighted. For example:

- Investigating the impact of data anonymization and obfuscation on Data Integrity and Data Coverage.
- Developing visual tools to help users balance privacy and quality, potentially introducing new QEDrant variants focused on these trade-offs.
- Exploring real scenarios where such trade-offs are critical, e.g., outsourced / crowdsourced AI projects.

Real-time and dynamic data quality assessment. Another promising direction is to enable real-time data quality assessment, similarly to [23]. This would involve:

- Integrating QEDrants directly into live data pipelines to provide continuous monitoring and feedback.
- Developing dynamic visualization capabilities to reflect changes in data quality metrics over time.
- Supporting adaptive decision-making by alerting users to emerging issues, with immediate recommendations.

Real-time assessments may be particularly valuable in industries like finance, where timely insights are critical.

In summary, the QEDrant framework provides a strong foundation for data quality assessment, but its full potential lies in integration with broader ecosystems. By pursuing the outlined future work, we aim to make QEDrants even more versatile, precise, and user-friendly. Some further investigations, particularly related to more types of granularity levels and associated visualizations, can be found in [21].

VII. Acknowledgements

This research was co-funded by the Polish National Centre for Research and Development under the project MAZOWSZE/0198/19.

References

[1] U. Jagare, *Operating AI: Bridging the Gap Between Technology and Business*, Wiley, 2022.

[2] M. Świechowski, *The History of Artificial Intelligence: From Leonardo da Vinci to Chat-GPT*, Amazon KDP, 2024.

[3] G.L. Geerts and D.E. O'Leary, "V-Matrix: A Wave Theory of Value Creation for Big Data," *International Journal of Accounting Information Systems*, vol. 47, pp. 100575, 2022.

[4] R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–34, 1996.

[5] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys*, vol. 41, pp. 16:1–16:52, 2009.

[6] S. Sadiq and M. Indulska, "Open Data: Quality Over Quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, 2017.

[7] Y. Gil and B. Selman, "A 20-year Community Roadmap for Artificial Intelligence Research in the US," *AI Magazine*, vol. 40, no. 1, pp. 8–24, 2019.

[8] R.Y. Wang and S.E. Madnick, "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective," in *Proceedings of VLDB 1990*, 1990, pp. 519–538.

[9] L. Pipino, Y.W. Lee, and R.Y. Wang, "Data Quality Assessment," *Communications of the ACM*, vol. 45, pp. 211–218, 2002.

[10] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, and L. Schilling, "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *Journal of Electronic Health Data and Methods*, vol. 4, no. 1, pp. 18, 2016.

[11] Y. Lee, D. Strong, B. Kahn, and R. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Information & Management*, vol. 40, pp. 133–146, 12 2002.

[12] S. Bresciani and M.J. Eppler, "Case Nr.2, 2008 – Updated in 2010 Gartner's Magic Quadrant and Hype Cycle," 2010.

[13] M. Kowalski, D. Ślęzak, and P. Synak, "Approximate Assistance for Correlated Subqueries," in *Proceedings of FedCSIS 2013*, 2013, pp. 1455–1462.

[14] M. Eppler and M. Helfert, "A Classification and Analysis of Data Quality Costs," in *Proceedings of ICIQ 2004*, 2004, pp. 311–325.

[15] A. Janusz, A. Zalewska, Ł. Wawrowski, P. Biczyk, J. Ludziejewski, M. Sikora, and D. Ślęzak, "BrightBox – A Rough Set Based Technology for Diagnosing Mistakes of Machine Learning Models," *Applied Soft Computing*, vol. 141, pp. 110285, 2023.

[16] D. Kałuża, A. Janusz, and D. Ślęzak, "Robust Assignment of Labels for Active Learning with Sparse and Noisy Annotations," in *Proceedings of ECAI 2023*. 2023, vol. 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 1207–1214, IOS Press.

[17] G.J. Miller, "Artificial Intelligence Project Success Factors – Beyond the Ethical Principles," in *Post-Proceedings of FedCSIS-AIST 2021*. 2021, vol. 442 of *Lecture Notes in Business Information Processing*, pp. 65–96, Springer.

[18] M.S. Szczuka, A. Janusz, B. Cyganek, J. Grabek, Ł. Przebinda, A. Zalewska, A. Bukała, and D. Ślęzak, "IEEE BigData Cup 2022 Report Privacy-preserving Matching of Encrypted Images," in *Proceedings of IEEE BigData 2022*. 2022, pp. 6471–6480, IEEE.

[19] A. Janusz, G. Hao, D. Kałuża, T. Li, R. Wojciechowski, and D. Ślęzak, "Predicting Escalations in Customer Support: Analysis of Data Mining Challenge Results," in *Proceedings of IEEE BigData 2020*. 2020, pp. 5519–5526, IEEE.

[20] T. Mroczek, D. Gil, and B. Pękała, "Fuzzy and Rough Approach to the Problem of Missing Data in Fall Detection System," *Fuzzy Sets and Systems*, vol. 480, pp. 108868, 2024.

[21] A. Powała and D. Ślęzak, "Hierarchical Approach to Data Quality Understanding in QEDrant Framework," in *Proceedings of IEEE BigData 2024*. 2024, IEEE.

[22] M. Bartoszuk, J. Litwin, M. Wnuk, and D. Ślęzak, "Tensor-based Approach to Big Data Processing and Machine Learning," in *Proceedings of IEEE BigData 2022*. 2022, pp. 6188–6194, IEEE.

[23] J. Bicevskis, Z. Bicevska, A. Nikiforova, and I. Oditis, "Towards Data Quality Runtime Verification," in *Proceedings of FedCSIS 2019*, 2019, vol. 18 of *Annals of Computer Science and Information Systems*, pp. 639–643.

Author Index