Journal of Big Data

RESEARCH Open Access

An integrated multistage ensemble machine learning model for fraudulent transaction detection

Md. Alamin Talukder[1*], Majdi Khalid[2] and Md Ashraf Uddin[3*]

*Correspondence:
alamin.cse@iubat.edu; ashraf.uddin@deakin.edu.au

[1] Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Dhaka, Bangladesh
[2] Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[3] School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, Australia

Abstract

Fraudulent transactions continue to pose a concern for financial institutions and organizations, necessitating the development of effective detection tools. Identification and prevention of fraudulent transactions depend heavily on the detection of credit card fraud. Even though instances of credit card fraud are uncommon, they can nonetheless cause significant financial losses because of the high cost of fraudulent transactions. When fraud is discovered early on, investigators can act quickly to stop additional losses. But because the investigation process takes a while, there are only so many warnings that can be looked through in detail in a given day. Thus, a fraud detection model's main goal is to minimize false alarms and missed fraud situations while producing accurate alerts. To improve fraud identification, we provide in this study an integrated multistage ensemble Machine Learning (IMEML) model that incorporates various multistage ensemble models intelligently, such as Ensemble Independent Classifier (EIC), Ensemble Bagging Classifier (EBC), and Ensemble ML Classifier (EMC). In order to overcome the problem of data imbalance, we use a number of methods-including Instant Hardness Threshold with EMC (IHT+EMC), Cluster Centroids (CC), and Random Under Sampler (RUS)-that go beyond traditional methods. We run our studies on a 284,807-transaction credit card dataset that is made available to the public. The accuracy rates of 99.94%, 99.91%, 99.14%, 99.52%, and perfect 100% for accuracy, precision, recall, f1-score, and AUC score, respectively, are achieved by the suggested model, demonstrating remarkable performance scores. For real-world fraud detection applications, the EIBMC model sets a new benchmark for identifying fraudulent transactions in high-frequency scenarios by outperforming cutting-edge techniques.

Keywords: Credit card, Fraudulent transactions, Integrated multistage model, Machine learning, Fraudulent detection

Introduction

Fraudulent transactions are defined as any unauthorized or misleading activity carried out with the purpose of unlawfully acquiring financial benefits [1]. These transactions can manifest in different ways, such as credit card fraud, identity theft, money laundering, and insurance scams. Fraudulent actions have emerged as a prominent issue for corporations and organizations on a global scale, resulting in substantial financial losses and

reputational harm [2]. The Association of Certified Fraud Examiners (ACFE) recently published a paper emphasizing that businesses face significant financial losses as a result of occupational fraud. On average, a single case of occupational fraud costs the victim organization more than $1.5 million [3]. Additionally, Certified Fraud Examiners (CFEs) estimate that organizations experience a loss of approximately 5% of their annual revenues due to fraud. In the ACFE's 2020 Report to the Nations, which examined 2504 cases of occupational fraud across 125 countries, it was found that the typical fraud lasted approximately 14 months before detection and resulted in a median monthly loss of $8300 [4].

To safeguard both themselves and their clients, companies and financial institutions must be able to identify fraudulent transactions. It preserves the integrity of the entire financial system in addition to assisting in the prevention of monetary losses [5]. Conventional approaches to fraud detection, which depend on rule-based algorithms and manual inspection, have shown to be insufficient for identifying complex and dynamic fraudulent activity. As a result, the need for increasingly sophisticated and automated methods of fraud detection and prevention is rising. The capacity of machine learning (ML) techniques to evaluate vast amounts of data and spot intricate patterns has made them effective instruments for fraud detection [6, 7]. From past transaction data, machine learning algorithms can identify trends or abnormalities that point to possible fraud. Scalability, efficiency, and adaptation to shifting fraud patterns are just a few of the benefits they provide [8]. Organizations can greatly enhance their fraud detection skills by utilizing machine learning techniques. Ensemble learning (EL) is a machine learning technique that generates a final prediction by aggregating the results of numerous independent models. In a number of fields, including fraud detection, it has demonstrated tremendous promise. Through the utilization of diverse model strengths, ensemble approaches may effectively mitigate overfitting, augment model generalization, and optimize overall performance [9]. Combining different models allows EL to identify a wide range of fraudulent transaction patterns and traits, which improves detection accuracy and dependability.

The detection of fraudulent transactions using ML and EL approaches has been the subject of several previous publications. To identify credit card fraud, for example, Esenogho et al. used Long Short-Term Memory (LSTM) in conjunction with Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors (SMOTE-ENN) [10]. Lakshmi and colleagues utilized machine learning models, such as Random Forest, Decision Tree, and Logistic Regression, to detect credit card fraud [11]. A Heterogeneous EL Model based on Data Distribution (HELMDD) was presented by Xie et al. [12] to address the problem of data imbalance in credit card fraud detection. Nevertheless, a few drawbacks of these current efforts include their inability to correct fraudulent transaction data that is significantly skewed and their limited accuracy improvement [12–14].

An integrated multistage ensemble machine learning (IMEML) model for detecting fraudulent transactions is presented in this paper. Our method seeks to enhance the accuracy of fraud detection while addressing the shortcomings of previous efforts. The IMEML model leverages the complementing strengths of many multistage ensemble ML (MEML) models, such as Ensemble Independent Classifier (EIC), Ensemble Bagging Classifier (EBC), and Ensemble ML Classifier (EMC), to improve fraud identification

and capture a variety of patterns in the data. We use a number of methods, including Randon Under Sampler (RUS), Cluster Centroids (CC), and Instant Hardness Threshold with EMC (IHT+EMC), to address the problem of data imbalance. We thoroughly test our suggested model on a real-world credit card fraud dataset to show its excellent accuracy and generalizability.

Our study has made the following significant contributions:

- Introduced an innovative IMEML model specifically suited for credit card fraudulent transaction detection. Our ensemble model includes different MEML algorithms, including EIC, EBC and EMC, exploiting their strengths and increasing fraud detection. By merging these models, we intend to capture varied patterns and characteristics of fraudulent transactions, leading to increased performance compared to individual models or traditional approaches.
- Implemented and extensively examined the performance of our proposed model, including trustworthy analysis on a real-world credit card fraud dataset. Furthermore, we compare the performance of our model with existing studies in the field of fraudulent transaction detection. Through this comparison, we illustrate the excellence and effectiveness of our IMEML strategy.

Our research greatly promotes fraudulent transaction detection by overcoming the limitations of existing systems and integrating state-of-the-art methodologies. The suggested IMEML model increases the accuracy and reliability of fraud detection, hence boosting the security and trustworthiness of financial systems. The performance study and comparison indicate our approach's practical adaptability and effectiveness in real-world circumstances. We hope that our research will considerably impact the security industry and contribute to establishing powerful fraud detection systems, helping enterprises, organizations, and individuals alike.

The remaining sections of this work are arranged as follows: In section Related works, we give a detailed assessment of relevant literature focusing on credit card fraudulent transaction detection. Section Methodology provides a full explanation of our study process and contains a description of the dataset used. The experimental setup and performance evaluation are given in section Results and discussion. In section Discussion and Cost analysis, we offered the results discussion and complexity analysis of our research. Furthermore, in section Dependable analysis, we perform a thorough investigation of the dependability of our suggested strategy. Lastly, section Conclusion presents the conclusion and future work.

Related works

Numerous studies have investigated the application of ensemble ML models for fraudulent transaction detection. These works have explored diverse ensemble techniques and data-balancing methods to enhance the accuracy and reliability of fraud detection systems.

Esenogho et al. [10] devised an innovative credit card fraud detection method utilizing a neural network ensemble (NNE) and a hybrid data resampling technique. By employing Long Short-Term Memory (LSTM) within an Adaptive Boosting (AdaBoost) framework

and employing Synthetic Minority Oversampling Technique-Edited Nearest Neighbors (SMOTE-ENN) for data resampling, their approach outperformed benchmark algorithms, achieving a sensitivity of 99.60% and specificity of 99.80%, showcasing its effectiveness in detecting credit card fraud.

Lakshmi et al. [11] evaluated the performance of logistic regression (LR), decision trees (DT), and random forests (RF) for credit card fraud detection using a highly imbalanced dataset. Through random oversampling (RO) to balance the data, RF outperformed LR and DT techniques, achieving an accuracy of 95.5%. Xie et al. [12] proposed the Heterogeneous EL Model based on Data Distribution (HELMDD) to address imbalanced data in credit card fraud detection. Through experiments on actual credit card datasets, HELMDD demonstrated superior performance compared to existing models, achieving high recall rates for minority and majority classes while significantly increasing savings rates for banks.

Soleymanzadeh et al. [13] proposed an ensemble stacking method for detecting cyberattacks in the Internet of Things (IoT). Conducting experiments on various datasets, including credit cards, NSL-KDD, and UNSW, their stacked ensemble classifier surpassed individual base model classifiers, achieving an impressive accuracy rate of 93.49%, suggesting its potential in addressing cyberattacks and credit card fraud.

Taha et al. [14] introduced an optimized light gradient boosting machine (OLightGBM) for fraud detection in credit card transactions, integrating a Bayesian-based hyperparameter optimization algorithm. Their method outperformed other approaches, achieving the highest accuracy of 98.40%, underscoring the effectiveness of intelligent parameter tuning in boosting model performance.

Faraji et al. [15] conducted a comprehensive evaluation of techniques for credit card fraud detection, proposing an ensemble model combining XGBoost with SMOTE to tackle data imbalance. Their approach yielded an impressive accuracy rate of 99%, emphasizing the significance of employing ensemble models for handling imbalanced data and achieving high accuracy in fraud detection.

Nandi et al. [16] introduced a novel multi-classifier architecture for credit card fraud detection, leveraging the Behavior-Knowledge Space (BKS) to combine predictions from multiple machine learning classification methods. Their ensemble model achieved an impressing accuracy rate of 99.81%, outperforming traditional techniques like majority voting, especially in scenarios requiring credit card fraud detection and noisy data classification.

Methodology

This section outlines the methodology we used to find CCFT in our investigation. We describe the key steps in our preferred methodology, which include preprocessing and collecting data, balancing and splitting data, creating multistage ensemble models, and doing performance analysis. The following sections emphasize our research's objectives:

Aims of the research

1. Data Gathering and Preparation: The goal of this study is to gather the CCFT dataset and use efficient data preparation methods, such as MinMax Scaler, to improve the quality of the data such that it is appropriate for the following stages.

2. Data Balance Methods: This study aims to address the imbalanced nature of fraudulent transaction data by applying efficient data balancing strategies, such as RUS, CC, and IHT+X, where X stands for ensemble machine learning models.
3. Building Multistage Ensemble Models: The main goal is to create a multistage ensemble machine learning model construction process. After that, all ensemble models will be integrated to create the final model that is suggested and especially made for CCFT detection. The suggested approach effectively combines multistage ensemble models, taking advantage of their advantages to greatly increase the accuracy of fraud detection.
4. Comparison and Analysis of Performance: Using a real-world credit card fraud dataset, the study attempts to provide a thorough performance analysis of the suggested hybrid multistage ensemble machine learning model. In addition, a thorough comparison with current methods for detecting fraudulent transactions will be carried out to show how much better and more efficient our suggested model is.
5. Adjustment to the Domain: The goal of the research is to solve the shortcomings of current methods and integrate cutting-edge strategies in order to significantly advance the field of fraudulent transaction detection. The suggested hybrid multistage ensemble machine learning approach will benefit individuals, companies, and organizations by improving the security and dependability of financial systems and improving accuracy and reliability in identifying fraud.

The methodology outlined for the development of the FTC model encompasses several key stages: CCFT data collection, data preprocessing, application of data balancing techniques, data splitting, construction of multistage ML models, integration of these models to form the proposed model, and thorough performance analysis to assess its efficacy in detecting fraudulent transactions. Figure 1 illustrates the proposed architecture of our hybrid ensemble multistage ML models designed specifically for CCFT detection.

Description of dataset

The Credit Card Fraud Detection dataset, which is publicly accessible on Kaggle [17], comprises anonymized credit card transactions conducted by European cardholders over a two-day period in September 2013. There are 284,807 transactions in the dataset overall, 492 of which are fake. There are 28 features in the dataset; the final feature is the class label, which indicates whether or not the transaction is fraudulent. Of these, 27 are numerical features produced by Principle component Analysis (PCA) transformation because of confidentiality concerns. The majority of the transactions in the dataset are not fraudulent, resulting in a severely skewed dataset. This makes it difficult to develop a classifier that minimizes false positives while effectively detecting fraudulent transactions. The dataset has been used to build and assess machine learning (ML) models for credit card fraud detection in a number of studies and competitions. The dataset is not meant for commercial usage; rather, it is meant to be used in the investigation and advancement of fraud detection algorithms. The public can access the dataset and utilize it for research or educational purposes. However, the owners of the dataset must give their prior written authorization before using this data for any commercial reason.

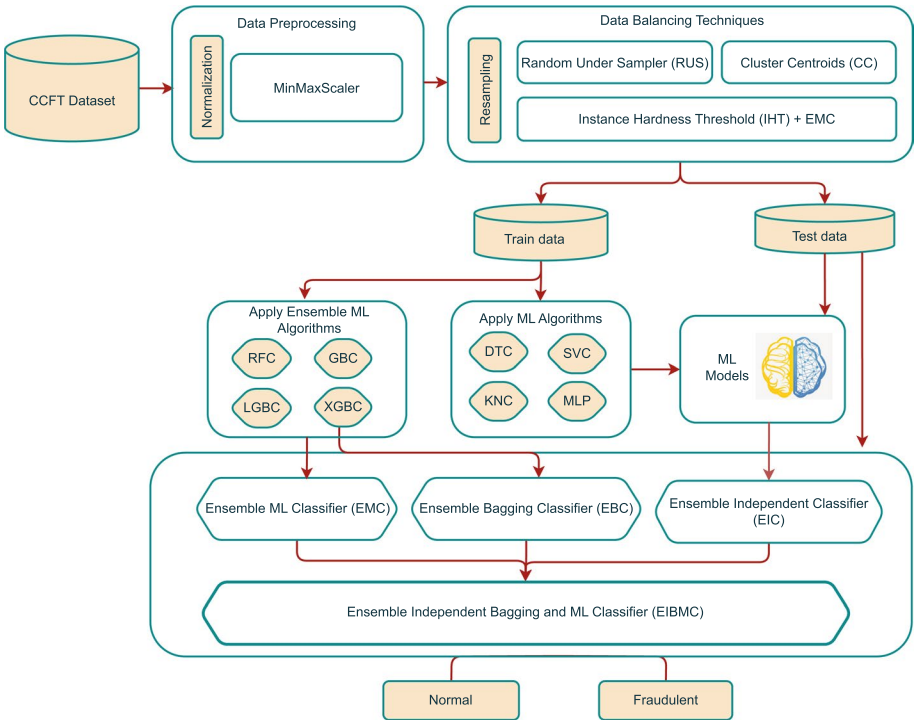


Fig. 1 The architecture developed for detecting fraudulent transactions

Data preprocessing

In the data preprocessing phase of our investigation, we diligently implemented MinMax Scaling techniques to normalize the data. MinMax Scaling involves transforming the data to a uniform scale, which facilitates optimal performance of machine learning models. By ensuring consistency and comparability across the dataset, these normalization techniques contribute to the efficient training and evaluation of ML models.

MinMax scaling

MinMax Scaling is a widely used normalization technique in data preprocessing, particularly in machine learning applications. Its purpose is to rescale the feature values of a dataset to a fixed range, typically between 0 and 1. This ensures that all features contribute equally to the analysis and prevents any single feature from dominating due to its scale.

The MinMax Scaling transformation for a feature x can be expressed using the following equation:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where:

- x_{scaled} is the scaled value of the feature.
- x is the original value of the feature.

- $\min(x)$ is the minimum value of the feature in the dataset.
- $\max(x)$ is the maximum value of the feature in the dataset.

This transformation ensures that the minimum value of the original feature maps to 0, and the maximum value maps to 1. All other values are linearly scaled between these two extremes. MinMax Scaling offers several benefits: It preserves the relationships between the original data points while ensuring consistency in scale. It is robust to outliers since it scales the entire range of the data. In ML, it often leads to faster convergence during training, especially in optimization algorithms that are sensitive to feature scales

Strategies for data balancing

In the landscape of fraudulent transaction detection, the presence of imbalanced datasets presents a formidable hurdle to the effectiveness of machine learning models. Conventional classification algorithms often demonstrate biased tendencies towards the majority class, resulting in less than optimal performance in identifying instances of fraud. To confront this inherent imbalance and enhance the resilience of our model, we have implemented a range of resampling techniques customized to the specific characteristics of our dataset.

- Random Under Sampler (RUS) The Random Under Sampler (RUS) technique involves randomly eliminating instances from the majority class until a more balanced distribution between the majority and minority classes is achieved. By reducing the dominance of the majority class, RUS aims to mitigate the bias towards it, thereby fostering improved performance in detecting fraudulent transactions.
- Cluster Centroids (CC) Cluster Centroids (CC) is a resampling method that operates by clustering the majority class instances and replacing each cluster centroid with the mean of the cluster. This technique effectively synthesizes new instances for the minority class while preserving the underlying structure of the data. By generating representative samples from the minority class within the clusters of the majority class, CC helps to rectify class imbalance and enhances the model's ability to detect fraudulent activities.
- Instance Hardness Threshold (IHT) Instance Hardness Threshold (IHT) is a data-driven approach that assigns a hardness score to each instance based on its difficulty in being correctly classified. Instances with higher hardness scores, indicating greater ambiguity or complexity, are given more weight during the training process. By focusing on the most challenging instances, IHT aims to improve the model's capacity to accurately classify minority class instances, such as fraudulent transactions, thereby bolstering overall performance and resilience against imbalanced data distributions. Furthermore, in our experiment, we integrated IHT with three distinct ensemble classifiers such as RFC, GBC and LGBC, utilizing them within the framework of IHT as a meta-classifier. This integration yielded three unique combinations of IHT with the classifiers: IHT + RFC, IHT + GBC, IHT + LGBC. Each combination leverages the strengths of both IHT, which prioritizes challenging instances, and the respective base classifiers, which provide distinct methodologies for classification. By integrating these classifiers with IHT, we aimed to enhance the model's

discriminatory power, particularly in the detection of fraudulent transactions within imbalanced datasets.

Table 1 provided demonstrates the distribution of instances before and after applying various undersampling techniques. The original dataset contains 284,807 instances, with the majority class (class 0) comprising 284,315 instances and the minority class (class 1) comprising only 492 instances. This severe class imbalance can lead to biased model training, where the algorithm may prioritize the majority class and struggle to effectively learn patterns from the minority class.

To address this issue, undersampling techniques are applied to balance the dataset. It involves reducing the number of instances in the majority class to match the number of instances in the minority class. By doing so, the algorithm can allocate sufficient attention to both classes during training, improving its ability to detect fraudulent transactions without being overwhelmed by the abundance of genuine transactions.

However, it's important to note that the total number of instances in the balanced dataset decreases significantly from the original dataset. Regarding the observation that even after undersampling, the number of instances in class 1 is not equal to class 0, this can be attributed to the fact that class 0 initially has a much larger number of instances than class 1. Undersampling aims to address the severe class imbalance by reducing the number of instances in the majority class to match the minority class. However, since class 0 has 284,315 instances compared to only 492 instances in class 1, it's impossible to achieve perfect equality in the number of instances between the two classes without losing a significant amount of information. Therefore, the undersampling process aims to strike a balance between addressing class imbalance and preserving meaningful data for model training and evaluation.

Table 1 Distribution of CCFT data employing data balancing techniques

Data balancing techniques	Dataset	Total instances	Class 0	Class 1
CCFT Dataset	Before	284807	284315	492
RUS	After	984	492	492
	Train	886	436	450
	Test	98	56	42
CC	After	984	492	492
	Train	886	436	450
	Test	98	56	42
IHT+RFC	After	277958	277466	492
	Train	250163	249722	441
	Test	27795	27744	51
IHT+GBC	After	57378	56886	492
	Train	51641	51196	445
	Test	5737	5690	47
IHT+LGBC	After	207135	206643	492
	Train	186422	185981	441
	Test	20713	20662	51
IHT+XGB	After	6372	5880	492
	Train	5735	5301	434
	Test	637	579	58

Utilized independent and ensemble ML algorithms

In our credit card fraud detection framework, we have employed a total of 8 popular ML algorithms, comprising 4 independent classifiers and 4 ensemble classifiers. Here's a description of each algorithm:

- Decision Tree (DTC): Decision Tree is a versatile classifier known for its flexibility in handling both classification and regression tasks [18]. By recursively partitioning the feature space based on informative attributes, it constructs a tree-like structure that aids in decision-making. Its interpretability and robustness make it a favored choice across various domains.
- Support Vector Classifier (SVC): Support Vector Classifier is a powerful supervised learning algorithm utilized for classification tasks [19]. It works by finding the optimal hyperplane that separates different classes with the maximum margin in the feature space. SVC is particularly effective in high-dimensional spaces and is widely used in various applications, including text classification, image recognition, and bioinformatics.
- K-Nearest Classifier (KNC): K-Nearest Classifier is a simple yet effective supervised learning algorithm used for classification and regression tasks [20]. It assigns labels to data points based on the majority vote of their nearest neighbors in the feature space. KNN's simplicity and interpretability make it suitable for various applications, including pattern recognition, anomaly detection, and recommendation systems.
- Multilayer Perceptron (MLP): Multilayer Perceptron is a type of artificial neural network characterized by multiple layers of interconnected neurons [6]. It is capable of learning complex patterns and relationships in data through nonlinear transformations. MLP is widely applied in tasks such as image recognition, natural language processing, and financial forecasting due to its ability to handle large and high-dimensional datasets.
- Random Forest Classifier (RFC): Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and robustness [8]. It mitigates overfitting by aggregating the predictions of individual trees, resulting in a more stable and accurate model. RFC is widely used in various fields, including finance, healthcare, and marketing, for tasks such as classification, regression, and feature selection.
- Gradient Boosting Classifier (GBC): Gradient Boosting Classifier is another ensemble learning technique that builds a sequence of decision trees iteratively, with each tree correcting the errors of its predecessor [3]. It combines the strengths of individual trees to create a strong predictive model. GBC is known for its high accuracy and is commonly used in applications such as web search ranking, click-through rate prediction, and customer churn prediction.
- Light Gradient Boosting Classifier (LGBC): Light Gradient Boosting Classifier is a scalable and efficient implementation of gradient boosting [6]. It optimizes memory usage and computational speed while maintaining high predictive performance. LGBC is widely used in large-scale machine learning tasks, such as web-scale recommendation systems, fraud detection, and financial modeling, due to its speed and accuracy.
- XGBoost Classifier (XGBC): XGBoost is an optimized implementation of gradient boosting that excels in both speed and performance [21]. It employs a regularization term in the objective function to control overfitting and enhance generalization.

XGBoost is highly versatile and is used in various applications, including credit risk modeling, time series forecasting, and personalized recommendation systems.

In our experiment, we employed various ML classifier algorithms for credit card fraud detection and identified the optimal parameters to enhance model performance. Table 2 displays the best parameters utilized for each classifier algorithm in our study.

Consider a potent CCFD system that makes use of several integrated ensemble state-of-the-art algorithms cooperating in a hybrid multistage ensemble model-rather than just one or four! Capable of capturing even the most skilled scammers, this system has 4 independent, 4 ensembles, and 1 bagging at its disposal. We can make sure that this dynamic combination of algorithms is genuinely unbeatable in spotting fraudulent transactions and protecting your financial security by assessing their performance using measures like precision, recall, accuracy, etc.

Multistage ensemble ML approach

To increase the precision and resilience of our CCFD model, we used an integrated multistage ensemble ML approach in our study. More specifically, we aggregated the predictions from several ensemble ML models, including EMC, EBC, and EIC, using a weighted vote classifier technique. We used the Grid Search technique to determine the ideal weights for each algorithm in order to give it the greatest performance possible for our ensemble model. For each algorithm, the weights had to be carefully adjusted in order to produce an accurate and balanced ensemble prediction. We sought to create a more potent credit card fraud detection model by applying this ensemble method and refining the weights via grid search, which might potentially protect the country's economy from the financial damage brought on by FD.

The weighted average voting ensemble model utilizes a weighted average of each base ensemble model's predictions, with the weights representing the models' respective relative relevance. By using each base model's strengths and mitigating its limitations, this method improves prediction accuracy overall. The integrated multistage ensemble model is more reliable at detecting CCF and less prone to overfitting because it incorporates the predictions of several base ensemble models. Utilizing the knowledge of several base models, the weighted average voting ensemble model enhances the performance of the credit card fraud detection model, which makes its application in our study noteworthy. To make sure that the ensemble model is well-calibrated and performs optimally in classifying fraudulent transactions, we conduct multiple experiments at multistage ensemble voting models. The optimized weights for these models vary depending on the ML model and range from 0.25 to 0.85 for different ML models. These weights are obtained through the grid search process. Financial institutions may be able to prevent possible financial losses due to fraud by doing this. The weighted average voting ensemble model is a useful method in credit card fraud detection and other machine learning applications due to its effectiveness and resilience.

Algorithm 1 describes the Grid Search Approach to Find the Optimal Weights in our proposed CCFT detection method. The algorithm presented is a grid search approach designed to find the best combination of weights for an ensemble classifier. It takes a list

Table 2 The best parameters used in our experiment by each classifier algorithm

Algorithm	Parameter	Used Value
Decision Tree Classifier (DTC)	criterion	'gini'
	splitter	'best'
	min_samples_split	2
	min_samples_leaf	1
K-Nearest Classifier (KNC)	n_neighbors	5
	weights	'uniform'
	leaf_size	30
	p	2
	metric	'minkowski'
Multilayer Perceptron (MLP)	hidden_layer_sizes	(100,)
	activation	'relu'
	solver	'adam'
	alpha	0.0001
	learning_rate_init	0.001
	max_iter	200
	shuffle	TRUE
Support Vector Classification (SVC)	C	1
	kernel	'rbf'
	degree	3
	gamma	'scale'
	probability	TRUE
	decision_function_shape	'ovr'
Random Forest Classifier (RFC)	n_estimators (number of trees)	100
	min_samples_split	2
	min_samples_leaf	1
	max_features	'sqrt'
	bootstrap	TRUE
Gradient Boosting Classification (GBC)	loss	'log_loss'
	learning_rate	0.1
	n_estimators	100
	min_samples_split	2
	min_samples_leaf	1
	max_depth	3
Extreme Gradient Boosting Classification (XGBC)	n_estimators	100
	learning_rate	0.3
	max_depth	6
	reg_alpha	0
	reg_lambda	1
Light Gradient Boosting Classification (LGBC)	n_estimators	100
	learning_rate	0.1
	max_depth	− 1
	reg_alpha	0
	reg_lambda	0

of pre-trained models and the corresponding training and test data as input. The goal is to determine the optimal weights for combining these models to achieve the highest accuracy on the test dataset. The algorithm systematically explores various weight combinations through nested loops, adjusting the weights for each model in the ensemble.

For each combination, it trains the ensemble classifier, evaluates its performance on the test data, and records the accuracy score. After evaluating all combinations, it selects the one with the highest accuracy and returns the corresponding weights. This method helps to automate the process of finding the most effective ensemble configuration, facilitating the development of robust detection models.

Algorithm 1 Grid search approach to find the optimal weights

Input:
trained_models_list: List of trained models
X_train: Training data
y_train: Training labels
X_test: Test data
y_test: Test labels
Output:
Optimal weights $w_1, w_2, w_3, \ldots, w_N$

Start:
$gf \leftarrow$ pd.DataFrame([])
$estimators \leftarrow$ trained_models_list
$count \leftarrow 0$
$N \leftarrow$ *number of models*
$upto \leftarrow N$
for $w1$ **in** range(1, upto) **do**
  for $w2$ **in** range(1, upto) **do**
    for $w3$ **in** range(1, upto) **do**
      ... ▷ Continue up to w_N
      for w_N **in** range(1, upto) **do**
        $wts \leftarrow [\frac{w1}{upto}, \frac{w2}{upto}, \ldots, \frac{w_N}{upto}]$
        $ens \leftarrow$ VotingClassifier(estimators, voting='soft', weights=wts).fit(X_train, y_train)
        $w\_pred \leftarrow$ ens.predict(X_test)
        $w\_acc \leftarrow$ accuracy_score(y_test, w_pred)
        $gf \leftarrow$ pd.concat([gf, pd.DataFrame('wt1': wts[0], 'wt2': wts[1], ..., 'wtN': wts[N-1], 'acc': w_acc*100, index=[0])], ignore_index=True)
        $count \leftarrow count + 1$
        if $count > N$ **then**
          break
        end if
      end for
    end for
  end for
end for

$max\_acc\_row \leftarrow gf.iloc[gf['acc'].idxmax()]$
$w_1, w_2, w_3, \ldots, w_N \leftarrow$ max_acc_row[0], max_acc_row[1], max_acc_row[2], ..., max_acc_row[N-1]
End

Results and discussion

Our study presents a novel integrated multistage ensemble machine learning model designed to identify fraudulent transactions. This part presents a concise summary of the experimental setup, the metrics used to evaluate performance, and an initial analysis of the findings and subsequent discussion.

Experimental configuration

The experimental configuration was implemented on a high-performance machine operating Windows 11 Pro, featuring an Intel Core i7 vPro 8th Generation processor, 16 GB of RAM, and a 500 GB solid-state drive (SSD). To implement our proposed model, we employed Anaconda Navigator and its Jupyter Notebook interface. The primary programming language was Python, with assistance from significant libraries like Scikit-learn, Pandas, NumPy, and Matplotlib. These libraries enabled a variety of tasks including data manipulation, numerical calculations, visualization, and ML processes.

Metrics for evaluating performance

To assess the efficacy of our proposed model, we employ a range of performance metrics, ensuring a comprehensive evaluation. These metrics include accuracy, precision, recall, F1-score, area under the curve (AUC) score, confusion matrix, as well as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Here's a brief definition of each metric:

- The confusion matrix provides a detailed breakdown of the model's predictions, showcasing true positives, true negatives, false positives, and false negatives. It offers valuable insights into the model's performance across different classes and aids in identifying areas for improvement. Table 3 presents the confusion matrix, visually summarizing the model's predictive accuracy and errors. where FP stands for False Positive, FN for False Negative, TP for True Positive, and TN for True Negative.
- $$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$
- $$Precision = \frac{TP}{TP + FP} \tag{2}$$
- $$Recall = \frac{TP}{TP + FN} \tag{3}$$
- $$F1score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4}$$
-

$$MAE = \frac{\sum_{i=1}^{n} predicted(i) - actual(i)}{n} \tag{5}$$

- $$MSE = \frac{\sum_{i=1}^{n} (predicted(i) - actual(i))^2}{n} \tag{6}$$

- $$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (predicted(i) - actual(i))^2}{n}} \tag{7}$$

where n stands for the total number of values.

- The AUC (Area Under the Curve) score is a crucial performance metric in evaluating classification models, including our fraudulent transaction detection model. It measures the model's ability to distinguish between positive and negative instances across various threshold values. A higher AUC score indicates better discrimination between classes, with a perfect score of 1 representing ideal classification.
- Stratified k-fold cross-validation with k = 5 was employed to robustly evaluate our model's performance. This method partitions the dataset into 5 equally sized folds while preserving the class distribution in each fold. It ensures that each fold contains a representative sample of both fraudulent and genuine transactions, facilitating an unbiased assessment of the model's effectiveness across different data subsets.

Results analysis

Our study delves into the realm of credit card fraud detection, a critical domain in financial security where accurate and efficient detection mechanisms are paramount. In this pursuit, we conducted five distinct experiments, each aimed at evaluating the performance of various ML models utilizing different data balancing techniques. The experiments were conducted on the CCFT dataset, a publicly available dataset widely used for benchmarking fraudulent transaction detection algorithms. Our objective was to comprehensively assess the effectiveness of these models in detecting fraudulent transactions under different experimental conditions.

Experiment 1 focused on employing the Random Under Sampling (RUS) technique, a commonly used method to address class imbalance, where instances of the majority

Table 3 Confusion matrix

	Actual positive	**Actual negative**
Predicted positive	TP	FP
Predicted negative	FN	TN

class are randomly removed to balance the dataset. Experiment 2 utilized the Cluster Centroids (CC) technique, which involves generating synthetic samples to balance the dataset by clustering the majority of class instances. Experiment 3 employed the Instance Hardness Threshold + Random Forest Classifier (IHT+RFC) approach, which combines instance hardness thresholding with the Random Forest algorithm for enhanced fraud detection. Experiment 4 delved into the Instance Hardness Threshold + Gradient Boosting Classifier (IHT+GBC) technique, where instance hardness thresholding is combined with the Gradient Boosting Classifier to improve model performance. Experiment 5 explored the Instance Hardness Threshold + Light Gradient Boosting Classifier (IHT+LGBC) approach, integrating instance hardness thresholding with the Light Gradient Boosting Classifier for enhanced fraud detection capabilities. Through these experiments, we aim to provide insights into the efficacy of different ML models and data-balancing techniques in detecting fraudulent transactions. By evaluating various performance metrics such as accuracy, precision, recall, F1-score, and AUC, we seek to identify the most effective approach for fraud detection, ultimately contributing to the advancement of security measures in financial transactions. Experiment 6 further expands our investigation into fraud detection methodologies. This experiment focuses on the Instance Hardness Threshold + Extreme Gradient Boosting Classifier (IHT+XGBC) technique. Here, instance hardness thresholding is coupled with the Extreme Gradient Boosting Classifier to enhance the model's ability to identify fraudulent transactions. By incorporating the powerful gradient boosting capabilities of XGBoost with instance hardness thresholding, we aim to improve the model's performance even further. Through this experiment, we seek to evaluate the effectiveness of this advanced approach in comparison to other techniques explored in previous experiments. Table 4 displays our Multistage Ensemble ML Model's Performance Analysis.

In each experiment, the performance of the EIBMC stood out prominently, showcasing its robustness and effectiveness in detecting fraudulent transactions. In Experiment 1, conducted with the RUS technique, the EIBMC model achieved an accuracy of 94.90%, precision of 94.93%, recall of 94.64%, F1-score of 94.78%, and an AUC of 99.29%. This demonstrated its capability to maintain high accuracy and precision while effectively identifying instances of fraud amidst class imbalances. Similarly, in Experiment 2 employing the CC technique, the EIBMC model continued to excel with an accuracy of 97.96%, precision of 97.92%, recall of 97.92%, F1-score of 97.92%, and an AUC of 99.84%. These results underscored the model's consistency in performance across different data balancing methods, affirming its reliability in fraud detection tasks. In Experiment 3, which utilized the IHT+RFC approach, the EIBMC model showcased its adaptability by achieving an accuracy of 99.95%, precision of 98.73%, recall of 88.23%, F1-score of 92.85%, and an AUC of 98.54%. Despite the complexity introduced by combining instance hardness thresholding with the Random Forest algorithm, the EIBMC model maintained its high-performance standards. Experiment 4 explored the IHT+GBC technique, where the EIBMC model again demonstrated its prowess with an accuracy of 99.91%, precision of 97.80%, recall of 96.79%, F1-score of 97.29%, and an AUC of 98.72%. In Experiment 5, employing the IHT+LGBC, the EIBMC model maintained its exceptional performance with an accuracy of 99.94%, precision of 99.97%, recall of 87.25%, F1-score of 92.68%, and an AUC of 95.99%. Through these experiments, the consistently

high performance of the EIBMC model across diverse data balancing techniques underscores its status as a promising choice for fraud detection in financial transactions.

Finally, In Experiment 6, the EIBMC stands out with exceptional performance scores. With an accuracy of 99.84%, precision of 99.91%, and recall of 99.14%, the EIBMC model demonstrates its ability to accurately identify fraudulent transactions. Moreover, achieving an impressive F1-score of 99.52% and a perfect AUC score of 100%, the model showcases its robustness and effectiveness in fraud detection tasks. Additionally, the EIBMC model exhibits minimal error metrics, with a MAE and MSE of 0.16, and an RMSE of 3.96. These outstanding results underscore the reliability and suitability of the EIBMC model for fraud detection applications, highlighting its potential as a powerful tool for ensuring financial security. While it leverages the diversity of the fundamental EML models and combines the finest aspects of multiple multistage ensemble models into a more robust and trustworthy system, our suggested integrated ensemble model performs exceptionally well. The grid search results improve the ensemble model and make it feasible for a successful fusion of the fundamental machine learning models' predictions by providing the best weights for those models. The ensemble model shows promising potential as a reliable and efficient technique for credit card fraud detection since it beats individual models and offers improved accuracy and precision in detecting fraudulent transactions. Through the use of thorough performance matrices, we have gained insight into the efficacy and potential uses of these models as well as more knowledge about their real-world performance.

In our pursuit of effective fraud detection methodologies, we rigorously evaluated the performance of our EIBMC model across various experiments illustrated in Fig. 2 and Table 4. Examining the AUC scores, a critical metric indicating the model's ability to discern between fraudulent and legitimate transactions, we uncovered compelling insights. Notably, in Experiment 6, where we employed the IHT+XGBC technique, our EIBMC model achieved a remarkable AUC score of 100. This exceptional result signifies unparalleled accuracy in identifying fraudulent activities, demonstrating the robustness of our model. Comparing this achievement with other experiments, we observed consistently high AUC scores across different methodologies. In Experiment 1, utilizing RUS, our EIBMC model attained an AUC score of 99.15. Experiment 2, employing CC, yielded an AUC score of 99.84. Experiment 3, which utilized the IHT+RFC approach, resulted in an AUC score of 98.54. Experiment 4, implementing IHT+GBC, produced an AUC score of 99.72. Experiment 5, integrating IHT+LGBC, delivered an AUC score of 95.99. However, it's Experiment 6 that stands out prominently, with the IHT+XGBC configuration showcasing the EIBMC model's exceptional capabilities. Achieving a perfect AUC score of 100, our model surpassed all other experiments in detecting fraudulent transactions. This remarkable performance underscores the effectiveness of the EIBMC model in real-world applications, promising heightened security and integrity in financial transactions.

Furthermore, the Confusion Matrix and ROC Curve of the EIMBC Model for CCFTD is illustrated in Fig. 3. The confusion matrix for the EIBMC model reveals exceptional performance, with 579 fraudulent transactions correctly identified (True Positives) and 57 legitimate transactions accurately classified (True Negatives), while only one

Table 4 Performance analysis of multistage ensemble ML model

Data balancing techniques	ML Model	Accuracy	Precision	Recall	F1-score	AUC	MAE	MSE	RMSE
RUS	DTC	91.84	91.54	91.96	91.71	91.96	8.16	8.16	28.57
	SVC	93.88	95.16	92.86	93.61	99.64	6.12	6.12	24.74
	KNC	90.82	90.74	90.48	90.6	96.7	9.18	9.18	30.3
	MLPC	93.88	94.52	93.15	93.66	99.57	6.12	6.12	24.74
	RFC	92.86	92.83	92.56	92.69	98.83	7.14	7.14	26.73
	GBC	93.88	93.6	94.05	93.78	98.96	6.12	6.12	24.74
	LBC	94.9	94.93	94.64	94.78	99.29	5.1	5.1	22.59
	XGBC	92.86	92.62	92.86	92.73	99.02	7.14	7.14	26.73
	EBC	94.9	95.33	94.35	94.74	99.05	5.1	5.1	22.59
	EIC	94.9	95.33	94.35	94.74	98.67	5.1	5.1	22.59
	EIBMC	94.9	94.93	94.64	94.78	99.15	5.1	5.1	22.59
CC	DTC	92.86	92.83	92.56	92.69	92.56	7.14	7.14	26.73
	SVC	93.88	95.16	92.86	93.61	98.9	6.12	6.12	24.74
	KNC	87.76	87.72	87.2	87.42	93.59	12.24	12.24	34.99
	MLPC	92.86	93.2	92.26	92.64	98.11	7.14	7.14	26.73
	RFC	97.96	97.92	97.92	97.92	99.63	2.04	2.04	14.29
	GBC	97.96	97.92	97.92	97.92	99.95	2.04	2.04	14.29
	LBC	98.98	98.84	99.11	98.96	99.97	1.02	1.02	10.1
	XGBC	97.96	97.92	97.92	97.92	99.84	2.04	2.04	14.29
	EBC	96.94	96.77	97.02	96.88	99.77	3.06	3.06	17.5
	EIC	93.88	94.52	93.15	93.66	99.16	6.12	6.12	24.74
	EIBMC	97.96	97.92	97.92	97.92	99.84	2.04	2.04	14.29
IHT+RFC	DTC	99.93	90.18	90.18	90.18	90.18	0.07	0.07	2.68
	SVC	99.95	98.73	88.23	92.85	96.42	0.05	0.05	2.16
	KNC	99.95	99.97	86.27	92.03	90.19	0.05	0.05	2.24
	MLPC	99.95	97.54	88.23	92.38	98.68	0.05	0.05	2.24
	RFC	99.96	99.98	89.22	93.95	95.93	0.04	0.04	1.99
	GBC	99.95	96.41	88.23	91.92	95.91	0.05	0.05	2.32
	LBC	99.41	58.63	79.15	63.24	74.37	0.59	0.59	7.68
	XGBC	99.96	99.98	88.24	93.32	98.16	0.04	0.04	2.08
	EBC	99.95	99.98	87.25	92.68	97.81	0.05	0.05	2.16
	EIC	99.95	98.73	88.23	92.85	98.81	0.05	0.05	2.16
	EIBMC	99.95	98.73	88.23	92.85	98.54	0.05	0.05	2.16
IHT+GBC	DTC	99.67	87.03	95.61	90.87	95.61	0.33	0.33	5.75
	SVC	99.86	95.71	95.71	95.71	98.82	0.14	0.14	3.73
	KNC	99.84	97.56	92.54	94.9	97.83	0.16	0.16	3.96
	MLPC	99.84	96.54	93.59	95.02	99.9	0.16	0.16	3.96
	RFC	99.91	97.8	96.79	97.29	99.82	0.09	0.09	2.95
	GBC	99.72	89.58	94.58	91.93	99.59	0.28	0.28	5.28
	LBC	99.88	97.68	94.66	96.12	98.85	0.12	0.12	3.49
	XGBC	99.88	97.68	94.66	96.12	99.52	0.12	0.12	3.49
	EBC	99.88	97.68	94.66	96.12	99.76	0.12	0.12	3.49
	EIC	99.9	97.74	95.73	96.71	99.82	0.1	0.1	3.23
	EIBMC	99.91	97.8	96.79	97.29	99.72	0.09	0.09	2.95

Table 4 (continued)

Data balancing techniques	ML Model	Accuracy	Precision	Recall	F1-score	AUC	MAE	MSE	RMSE
IHT+LGBC	DTC	99.9	92.82	85.28	88.68	85.28	0.1	0.1	3.18
	SVC	99.9	92.82	85.28	88.68	92.43	0.1	0.1	3.18
	KNC	99.93	99.97	86.27	92.03	89.2	0.07	0.07	2.6
	MLPC	99.89	94.55	82.34	87.47	94.65	0.11	0.11	3.26
	RFC	99.94	99.97	88.24	93.32	94.8	0.06	0.06	2.41
	GBC	99.85	92.79	73.52	80.34	67.71	0.15	0.15	3.87
	LBC	99.64	68.5	82.22	73.48	82.42	0.36	0.36	5.98
	XGBC	99.94	99.97	88.24	93.32	93.48	0.06	0.06	2.41
	EBC	99.94	99.97	87.25	92.68	94.49	0.06	0.06	2.51
	EIC	99.92	97.4	86.27	91.09	95.48	0.08	0.08	2.78
	EIBMC	99.94	99.97	87.25	92.68	95.99	0.06	0.06	2.51
IHT+XGBC	DTC	99.06	96.49	97.93	97.2	97.93	0.94	0.94	9.71
	SVC	99.22	99.57	95.69	97.53	100	0.78	0.78	8.86
	KNC	99.22	99.57	95.69	97.53	97.4	0.78	0.78	8.86
	MLPC	99.69	99.83	98.28	99.04	99.79	0.31	0.31	5.6
	RFC	99.84	99.91	99.14	99.52	100	0.16	0.16	3.96
	GBC	99.69	99.05	99.05	99.05	99.99	0.31	0.31	5.6
	LBC	99.69	99.83	98.28	99.04	100	0.31	0.31	5.6
	XGBC	99.84	99.91	99.14	99.52	100	0.16	0.16	3.96
	EBC	99.84	99.91	99.14	99.52	100	0.16	0.16	3.96
	EIC	99.53	99.74	97.41	98.54	100	0.47	0.47	6.86
	EIBMC	99.84	99.91	99.14	99.52	100	0.16	0.16	3.96

legitimate transaction is incorrectly labeled as fraudulent (False Positive) and one fraudulent transaction is missed (False Negative). This demonstrates high precision and recall, indicating minimal misclassifications. Moreover, the Receiver Operating Characteristic (ROC) curve underscores the model's robustness, with an AUC of 100, reflecting perfect discrimination between fraudulent and legitimate transactions. This signifies the model's ability to achieve a true positive rate of 100% while maintaining a false positive rate of 0%, ensuring optimal fraud detection without erroneous classifications.

In conclusion, our study highlights the significance of thorough performance assessment when deploying an integrated ensemble multistage ML model for detecting fraudulent transactions. By systematically evaluating the efficacy of various models, we have provided valuable insights into their strengths and weaknesses, empowering businesses and researchers to make informed choices regarding model selection for real-world applications. The findings from our analysis offer valuable contributions to the fields of ML, fraud detection, and financial security.

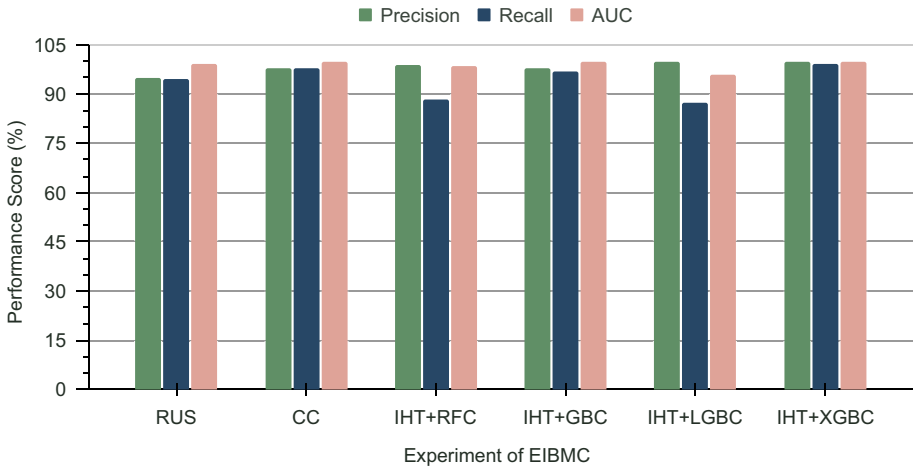


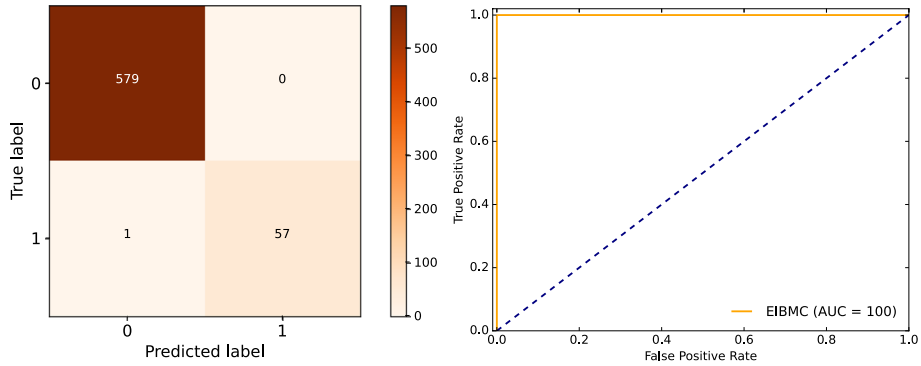
Fig. 2 Performance analysis of EIBMC model for CCFTD

Discussion

In comparing our proposed model with existing works in fraudulent transaction detection, Table 5 presents a comprehensive analysis based on various metrics. Esenogho et al. [10] employed the SMOTE-ENN technique with an LSTM Ensemble, achieving a commendable AUC score of 99.00%. Xie et al. [12] introduced the HELMDD model without employing balancing techniques, achieving an AUC score of 98.53%. Lakshmi et al. [11] utilized random oversampling (RO) with Random Forests (RF), achieving an accuracy of 95.50%. Soleymanzadeh et al. [13] employed Ensemble Stacking without specifying balancing techniques, achieving an accuracy of 93.49%. Faraji et al. [15] employed SMOTE with an Ensemble model, achieving an accuracy score of 99.00%. Taha et al. [14] introduced OLightGBM without employing specific balancing techniques, achieving an accuracy of 98.40%. Nandi et al. [16] presented the BKS model without specifying balancing techniques, achieving an impressive accuracy score of 99.80%. In comparison, our proposed model, utilizing the IHT+XGBC technique with the EIBMC model, outperforms existing methods with an accuracy score of 99.94% and an AUC score of 100%. This comparison underscores the superior performance of our proposed model in fraudulent transaction detection, demonstrating its potential to significantly enhance security measures in financial transactions.

However, it's important to emphasize that our primary focus was on evaluating the performance of our proposed multistage ensemble models relative to existing approaches. While it's true that different researchers may employ diverse techniques for data balancing, we ensured consistency in the initial dataset size (284,807), which remains the same across our experiments and those of existing works. This approach allows for a fair comparison of model efficacy and facilitates meaningful insights into the relative strengths and weaknesses of various fraud detection methodologies.

Our proposed model, leveraging the IHT+XGBC technique with the EIBMC model, outperforms existing methods through its integration of multiple stages of ensemble models. By harnessing the collective intelligence of diverse ensemble techniques, our model achieves superior accuracy and AUC scores in fraud detection tasks, enhancing



(a) Confusion Matrix (b) ROC Curve

Fig. 3 Confusion natrix and ROC curve of EIMBC model for CCFTD

Table 5 Comparison analysis of our model with existing works

Sl. No.	Authors	Balancing Techniques	Models	Accuracy (%)	AUC (%)
1	Esenogho et al. [10]	SMOTE-ENN	LSTM Ensemble	–	99.00
2	Xie et al. [12]	–	HELMDD	–	98.53
3	Lakshmi et al. [11]	RO	RF	95.50	–
4	Soleymanzadeh et al. [13]	–	Ensemble Stacking	93.49	–
5	Faraji et al. [15]	SMOTE	Ensemble	99.00	–
6	Taha et al. [14]	–	OLightGBM	98.40	–
7	Nandi et al. [16]	–	BKS	99.80	–
8	Our Proposal	IHT+XGBC	EIBMC	99.94	100

the reliability of financial transaction monitoring systems. The adoption of our model in CCFD systems can strengthen fraud prevention measures, offering financial institutions timely and accurate insights into suspicious activities, thus fortifying the security of consumers' financial transactions.

Cost analysis

Cost analysis refers to the evaluation of resources consumed during these processes, both in terms of time and computational requirements. In Table 6, the costs are represented as the time taken for building and predicting using different ML models under various data balancing techniques. The build time signifies the duration required to train the model on the dataset, while the prediction time indicates the time taken to generate predictions for new or unseen data points. This cost analysis allows researchers and practitioners to assess the efficiency of different models. Lower build and prediction times imply reduced computational expenses and faster response times, which are desirable qualities, especially in real-time applications like credit card fraud detection. The associated Fig. 4 provides a visual representation of these costs, aiding in the comparative analysis of different models and techniques.

The IHT+XGBC EIBMC model exhibits notably reduced build and prediction times compared to other IHT+X models in the provided cost analysis. While it may require

slightly more time than independent ML models for both building and predicting, it offers superior performance. This multistage ensemble model surpasses the independent ones by leveraging the collective strength of multiple classifiers. Despite its slightly increased computational overhead, its enhanced accuracy and reliability make it a preferred choice for credit card fraud detection. Additionally, the IHT+XGBC EIBMC model strikes a balance between efficiency and effectiveness, offering a compelling solution for real-world applications where accuracy and speed are crucial factors.

Our IHT+XGBC EIBMC model stands out for its remarkable efficiency in both build and prediction times compared to other models utilizing the same data balancing techniques. With a mere 19.7641 ms of build time and 0.0749 ms of prediction time, it significantly outperforms its counterparts, such as IHT+RFC EIBMC, IHT+LGBC EIBMC, and IHT+GBC EIBMC, in terms of speed.

The exceptionally low prediction time of 0.0749 ms underscores the model's ability to provide rapid responses, making it well-suited for real-time applications where timely decisions are paramount. Despite its swift execution, this model maintains high accuracy, ensuring reliable predictions for credit card fraud detection tasks.

The combination of fast and accurate predictions positions our IHT+XGBC EIBMC model as a highly efficient and effective solution for addressing the challenges of credit card fraud detection. Its ability to deliver rapid results without compromising accuracy makes it an invaluable tool for financial institutions and researchers seeking optimal performance in fraud detection systems.

Dependable analysis

In the realm of credit card fraud detection, ensuring dependability is paramount, encompassing factors like reliability, availability and efficiency as emphasized by Talukder et al. [21].

Our model, leveraging an IMEML approach, aligns well with the concept of dependability. By amalgamating diverse Multistage ML models and data balancing strategies, it enhances reliability by mitigating the risk of false positives and negatives in Fig. 3. Additionally, the model's ability to achieve high precision, recall and AUC rates, as demonstrated by the results in Fig. 2, underscores its dependability in accurately identifying fraudulent transactions while minimizing false alarms.

Moreover, the model's efficiency, as indicated by its comparatively lower build and prediction times, ensures its availability for real-time deployment in financial institutions' fraud detection systems in Fig. 4. In essence, our proposed model exhibits commendable dependability, aligning closely with the principles. Its reliability, availability and efficiency collectively contribute to its efficacy in combating credit card fraud, making it a dependable solution for financial institutions and researchers in the field.

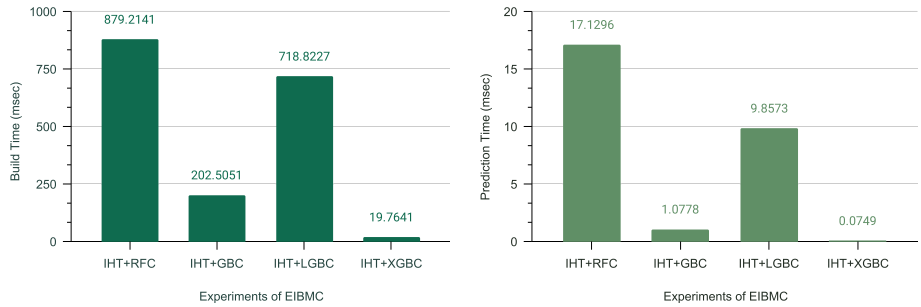
Conclusion

In this paper, we introduced an innovative approach to fraudulent transaction detection using an integrated multistage ensemble ML model. Our methodology encompassed various stages, including data collection, preprocessing, normalization, balancing, splitting, multistage ensemble model (MEM) construction, integration, and performance evaluation. Leveraging a credit card fraudulent transaction dataset, we tackled data

Table 6 Cost analysis of multistage ensemble ML models

Data balancing techniques	ML model	Build time	Prediction time
IHT+RFC	DTC	19.2007	0.0033
	SVC	70.3317	0.5442
	KNC	0.0355	16.5128
	MLPC	87.7239	0.2316
	RFC	226.5927	0.27
	GBC	318.7969	0.0205
	LBC	3.5822	0.0656
	XGBC	2.6217	0.024
	EBC	24.809	0.2627
	EIC	309.0796	15.3606
	EIBMC	879.2141	17.1296
IHT+GBC	DTC	3.2144	0.0015
	SVC	10.5495	0.1502
	KNC	0.0069	0.989
	MLPC	87.3998	0.0379
	RFC	37.8179	0.0582
	GBC	60.4271	0.0049
	LBC	1.3055	0.0198
	XGBC	0.8377	0.0061
	EBC	6.7445	0.0565
	EIC	73.3799	0.8955
	EIBMC	202.5051	1.0778
IHT+LGBC	DTC	13.2404	0.0025
	SVC	58.8033	0.5503
	KNC	0.0257	9.0736
	MLPC	95.3515	0.1289
	RFC	160.5604	0.2123
	GBC	237.8369	0.0376
	LBC	3.046	0.0695
	XGBC	1.9757	0.0192
	EBC	19.8812	0.2298
	EIC	248.2804	8.7945
	EIBMC	718.8227	9.8573
IHT+XGBC	DTC	0.2126	0.0003
	SVC	0.4586	0.0082
	KNC	0.0016	0.0457
	MLPC	8.5088	0.0034
	RFC	2.927	0.0105
	GBC	5.5692	0.0012
	LBC	0.27	0.0051
	XGBC	0.1283	0.0011
	EBC	1.0852	0.0089
	EIC	8.5907	0.0294
	EIBMC	19.7641	0.0749

imbalance using RU, CC, and IHT+X techniques, where X denotes Ensemble ML algorithms. Subsequently, we initialized MEM with multiple ML algorithms and fused them to construct an IMEML for fraud detection.



(a) Build Time (b) Prediction Time

Fig. 4 Cost analysis of EIBMC models for CCFTD

Our proposed ensemble model presents a robust and efficient solution for credit card fraud detection, exhibiting superior accuracy, precision, and computational efficiency compared to existing techniques. By amalgamating the strengths of multiple base ensemble models and optimizing performance through grid search, our EIBMC model achieved a perfect 100% AUC score, showcasing its exceptional performance.

Despite its strengths, our model has limitations. Its performance is contingent upon the quality and representativeness of the input dataset and may struggle with previously unseen or evolving fraudulent transaction types. Additionally, the choice of ML or DL algorithms and hyperparameter settings can influence performance, necessitating further optimization.

Prospective avenues for investigation may comprise the integration of sophisticated deep feature engineering approaches, investigation of alternative methodologies for data balancing, and integration of real-time data to optimize performance. A deeper understanding of the efficacy and generalizability of our suggested model would be possible by experimentation with a variety of datasets and comparison with cutting-edge techniques.

Abbreviations

ML	Machine learning
EML	Ensemble ML
Deep Learning	DL
IM	Integrated multistage
MEML	Multistage ensemble ML
MEM	Multistage ensemble model
IMEML	Integrated multistage ensemble ML
FD	Fraud detection
CCF	Credit card fraud
CCFT	Credit card fraud transaction
CCFD	Credit card fraud detection
CCFTD	Credit card fraud transaction detection
RUS	Random under sampler
CC	Cluster centroids
IHT	Instance hardness threshold
DTC	Decision tree classifier
SVC	Support vector classifier
KNC	K-neighbors classifier
MLP	Multilayer perceptron
RFC	Random forest classifier
GBC	Gradient boosting classifier
LGBC	Light gradient boosting classifier
XGBC	XGBoost classifier

EMC	Ensemble ML classifier
EBC	Ensemble bagging classifier
EIC	Ensemble independent classifier
EIBMC	Ensemble independent bagging ML classifier

Acknowledgements
This work is supported by the Open Research Support, Springer Nature.

Author contributions
Md. Alamin Talukder: Data curation, methodology, software, investigation, formal analysis, visualization, writing—original draft and writing—reviewing and editing. Majdi Khalid: Validation, visualization, writing—reviewing and editing. Md Ashraf Uddin: Investigation, formal analysis, visualization, writing—reviewing and editing.

Funding
There is no funding available for this research project.

Availability of data and materials
The selected datasets are sourced from free and open-access sources such as Credit Card Fraudulent Transaction Dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no Conflict of interest to declare that they are relevant to the content of this article.

Received: 15 June 2023 Accepted: 2 September 2024
Published online: 22 November 2024

References

1. Putra R, Lubis SD. Law enforcement for fraud offenders on behalf of banks through online according to islamic criminal law. J Law Politic Humanit. 2024;4(3):295–305.
2. Nakitende MG, Rafay A, Waseem M. Frauds in business organizations: a comprehensive overview. Res Anthol Bus Law Policy and Soc Responsib. 2024;848–865
3. Talukder MA, Hossen R, Uddin MA, Uddin MN, Acharjee UK. Securing transactions: a hybrid dependable ensemble machine learning model using iht-lr and grid search. Cybersecurity 2024. https://doi.org/10.48550/arXiv.2402.14389
4. Mustapha R, Fauzi MA, Soon OT, Wei LH, Yee CM. Employee perception of whistleblowing in the workplace: a systematic bibliometric review. Pak J Life Soc Sci. 2024;22(1)
5. Chatterjee P, Das D, Rawat DB. Digital twin for credit card fraud detection: opportunities, challenges, and fraud detection advancements. Future Gener Comput Syst. 2024;
6. Talukder MA, Islam MM, Uddin MA, Akhter A, Hasan KF, Moni MA. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Syst Appl. 2022;205:117695.
7. Talukder MA, Islam MM, Uddin MA, Akhter A, Pramanik MAJ, Aryal S, Almoyad MAA, Hasan KF, Moni MA. An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. Expert Syst Appl. 2023;230:120534.
8. Talukder MA, Islam MM, Uddin MA, Hasan KF, Sharmin S, Alyami SA, Moni MA. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. J Big Data. 2024;11(1):33.
9. Chang V, Di Stefano A, Sun Z, Fortino G. Digital payment fraud detection methods in digital ages and industry 4.0. Comput Electr Eng. 2022;100:107734.
10. Esenogho E, Mienye ID, Swart TG, Aruleba K, Obaido G. A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access. 2022;10:16400–7.
11. Lakshmi S, Kavilla S. Machine learning for credit card fraud detection system. Int J Appl Eng Res. 2018;13(24):16819–24.
12. Xie Y, Li A, Gao L, Liu Z. A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. Wireless Commun Mobile Comput. 2021;2021:1–13.
13. Soleymanzadeh R, Aljasim M, Qadeer MW, Kashef R. Cyberattack and fraud detection using ensemble stacking. AI. 2022;3(1):22–36.
14. Taha AA, Malebary SJ. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access. 2020;8:25579–87.
15. Faraji Z. A review of machine learning applications for credit card fraud detection with a case study. SEISENSE J Manag. 2022;5(1):49–59.
16. Nandi AK, Randhawa KK, Chua HS, Seera M, Lim CP. Credit card fraud detection using a hierarchical behavior-knowledge space model. PloS ONE. 2022;17(1):0260579.

17. MLG - ULB: Credit Card Fraud Dataset. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Accessed 15 June 2023 2023;
18. Hammed M, Soyemi J. An implementation of decision tree algorithm augmented with regression analysis for fraud detection in credit card. Int J Comput Sci Inf Secur. 2020;18(2):79–88.
19. Talukder MA, Sharmin S, Uddin MA, Islam MM, Aryal S. Mlstl-wsn: machine learning-based intrusion detection using smotetomek in wsns. Int J Inf Secur. 2024;23:1–20.
20. Ganji VR, Mannem SNP. Credit card fraud detection using anti-k nearest neighbor algorithm. Int J Comput Sci Eng. 2012;4(6):1035–9.
21. Talukder MA, Hasan KF, Islam MM, Uddin MA, Akhter A, Yousuf MA, Alharbi F, Moni MA. A dependable hybrid machine learning model for network intrusion detection. J Inf Secur Appl. 2023;72: 103405.

Publisher's Note