Article

Not peer-reviewed version

Hybrid Data Mining Technique for Credit Card Fraud Detection

Dhakirah Salahudin-Mukeem and Olufisayo Ekundayo *

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1511.v1

Keywords: Credit Card; Fraud Detection; Machine learning; Hybrid model

Article

Hybrid Data Mining Technique for Credit Card Fraud Detection

Dhakirah Salahudin-Mukeem [1] and Olufisayo S. Ekundayo [2,*]

[1] College of Business,Technology and Engineering, Sheffield Hallam University, Sheffield, England
[2] School of Computing, University of South Africa, Florida, South Africa
* Correspondence: ekundo@unisa.ac.za

Abstract: The rising incidence of credit card fraud underscores the need for innovative strategies to enhance credit card fraud detection and prevention. Numerous approaches have been employed for credit card fraud detection; however, the field continues to seek methods that can adapt to the constantly evolving nature of fraud patterns. In this study, we develop a hybrid model by integrating machine learning algorithms for effective credit card fraud detection. Using a simulated credit card transaction dataset, the model is developed in two stages, the first stage finds a base algorithm for the proposed hybrid model. The second stage focus on developing the hybrid model by combining the base model (Light Gradient Boosting Machine) with each of the selected algorithms. The hybrid models, demonstrated superior performance compare to standalone algorithms. Also, the hybrid of LGBM and XGBoost model outperforms others combinations, having 98.3% accuracy, 98.88% Precision, 98.05% Recall, 98.46% F1-Score, 99.80% AUROC. This proposed hybrid model can enhance security and foster trust in financial institutions and businesses, and in turn contribute to a more stable and efficient financial ecosystem.

Keywords: Credit Card; Fraud Detection; Machine learning; Hybrid model

1. Introduction

Credit card fraud continues to cast a formidable shadow over individuals and financial institutions worldwide, exacting significant financial tolls and causing immeasurable harm to unsuspecting victims [1]. The devious nature of credit card fraud entails the unauthorised and fraudulent use of credit card data by malicious third parties, which has serious negative effects on the economy and causes psychological pain [2]. The significant global extent of this issue is emphasized by the striking statistics disclosed in the Nilson Report. It disclosed that in 2018, losses attributed to card fraud surged beyond an astounding $27.85 billion, highlighting the urgent necessity for formidable and efficient fraud detection systems [3]. The global financial losses due to credit card fraud reached 32.34 billion in 2021, with the US accounting for 11.9 billion or 37% of worldwide losses. Credit card fraud is projected to cost the global card industry 397 billion over the next 10 years, with 165 billion in losses coming from the US [4].

Credit card fraud can have severe financial, emotional, and psychological impacts on victims. 12% of US identity theft victims experienced out-of-pocket costs averaging $690, and 32% reported moderate to severe emotional distress [5,6]. To mitigate the impact of credit card fraud, fraud detection systems play a pivotal role by identifying and thwarting fraudulent transactions in real-time [7]. These systems harness advanced data mining techniques to analyse vast volumes of transactional data, facilitating early detection of anomalous patterns and enabling timely interventions to minimize financial losses [8].

In the realm of credit card fraud detection, conventional data mining techniques such as decision trees, neural networks, logistic regression, and rule-based algorithms have been widely employed [9]. Despite some successes, these methods grapple with inherent limitations that hinder their efficacy in this context. One primary constraint is the challenge posed by highly imbalanced datasets, where

 doi:10.20944/preprints202501.1511.v1

instances of fraud are markedly rare. This imbalance can lead to suboptimal accuracy as models tend to be biased toward the majority class [10]. The skewed nature of the data can result in the misclassification of fraudulent transactions, contributing to elevated false positive rates, thereby diminishing the overall reliability of the fraud detection system [11].

Moreover, the inadequacy of conventional techniques in capturing intricate fraud patterns is evident. Credit card fraud often involves sophisticated and dynamic patterns that conventional models struggle to discern due to their limitations in handling non-linearity and complex interdependencies [10]. Fraudsters constantly adapt their tactics, necessitating models with a higher degree of flexibility. This limitation underscores the need for more sophisticated approaches in credit card fraud detection.

Introducing convolutional techniques into this domain introduces additional challenges. One significant hurdle is the sequential nature of credit card transactions, where temporal dependencies play a crucial role [12]. Convolutional Neural Networks (CNNs), designed originally for grid-like data such as images, may not inherently capture the sequential dependencies present in transaction sequences [10]. The sequential nature of transactions, with varying time intervals and orders, poses a unique challenge for CNNs, which are not explicitly tailored for processing such temporal data.

Furthermore, the high dimensionality and variable sequence lengths inherent in credit card transactions present obstacles for traditional CNN architectures [12]. Transactions vary widely in terms of amounts, locations, and frequencies, resulting in datasets with diverse characteristics. Conventional CNNs may face difficulties efficiently handling such variable-length sequential data [10].

A critical concern is the limited interpretability of CNNs, often regarded as black-box models [13]. While CNNs excel in predictive capabilities, their opacity in revealing the decision-making process poses challenges, particularly in scenarios where interpretability is crucial, such as in fraud detection [10]. Understanding the rationale behind a model's prediction is essential for building trust among stakeholders and complying with regulatory requirements [14].

In response to these challenges, researchers have turned to hybrid data mining techniques, which amalgamate diverse algorithms and approaches to leverage their collective strengths while mitigating individual weaknesses [15]. One notable example is the integration of decision trees and ensemble methods, as demonstrated in the work by [16]. In their study, the authors proposed a hybrid approach that combined the interpretability of decision trees with the robustness of ensemble methods, specifically a random forest. The decision tree component provided a clear and interpretable model, allowing for a straightforward understanding of the features contributing to fraud detection. Simultaneously, the ensemble method addressed issues related to overfitting and enhanced predictive accuracy by aggregating the outputs of multiple decision trees. This amalgamation of techniques leverages the interpretability of decision trees and the predictive power of ensemble methods, showcasing the potential of hybrid approaches in credit card fraud detection [17]. The emergence of hybrid methods holds the promise of enhancing credit card fraud detection by reducing false positives, improving accuracy, and providing more resilient protection to individuals and financial institutions [18].

In this study we investigate and assess the usefulness of hybrid methods of data mining for identifying credit card fraud in light of the present constraints and the urgent need for more reliable fraud detection systems. By delving into the intricacies of these hybrid approaches, the study endeavours to unravel new possibilities in fraud detection technology, forging a path towards more secure financial transactions and bolstering confidence in electronic payment systems.

2. Related Works

Credit card fraud has emerged as a significant concern for individuals and financial institutions worldwide, leading to substantial financial losses and causing severe repercussions for victims. [19] highlights the pressing need for robust fraud detection systems to counter the ever-evolving tactics employed by fraudulent actors. Various studies have underscored the global impact of credit card fraud and the urgency to implement effective preventive measures [2]. The Nilson Report's alarming statistics revealed that bank fraud losses surpassed $27.85 billion in 2018, reinforcing the urgency to

 doi:10.20944/preprints202501.1511.v1

develop advanced fraud detection technologies [3]. This section comprehensively evaluates existing research in credit card fraud detection, emphasizing traditional data mining methods and the emerging trend toward hybrid techniques.

Traditional data mining techniques including decision trees, neural networks, logistic regression, and rule-based algorithms have been employed diversely to credit card fraud detection. [20] showcased the effectiveness of decision trees, underscoring their utility in capturing decision boundaries within the intricate landscape of credit card transactions. The transparency of decision trees has made them valuable not only for identifying potentially fraudulent activities but also for offering insights into the reasoning behind each classification.

[21] investigated the effect of imbalanced data on decision trees performance when they employed decision trees to credit card fraud detection, and highlighted a significant challenge posed by imbalanced datasets. In fraud detection scenarios, legitimate transactions often far outnumber fraudulent ones, creating an inherent imbalance. [21] emphasized that this imbalance can lead to biased models, with decision trees potentially favouring the majority class and resulting in suboptimal accuracy. The study pinpointed a critical limitation that decision tree models must contend with, urging a closer examination of their adaptability to imbalanced datasets in the context of credit card fraud detection. [22] provided further insights into the specific attributes of decision trees that contribute to their effectiveness in fraud detection. Their research delved into the optimization of decision tree parameters to enhance model performance, offering practical guidance on parameter tuning for credit card fraud detection scenarios. Furthermore, [23] conducted a comparative analysis of decision trees against other machine learning models, shedding light on the relative strengths and weaknesses of decision trees in the context of credit card fraud detection. Their findings contributed to the ongoing dialogue regarding the optimal selection of models in different fraud detection scenarios. Moreover, [24] explored the application of ensemble methods involving decision trees for improved fraud detection accuracy. Their research demonstrated the advantages of combining decision trees into ensembles, leveraging the diversity of individual trees to enhance overall model robustness.

Artificial neural network is another data mining model explored for credit card fraud detection, marked by its profound recognition of its prowess in modelling nonlinear relationships within intricate transactional data. [25] conducted an in-depth analysis of the challenges inherent in deploying neural networks for fraud detection, particularly in the context of the dynamic nature of fraudulent patterns. Their research emphasized the necessity for continuous refinement in the application of neural networks to ensure adaptability to evolving tactics employed by fraudsters. [26] delved into the importance of feature engineering in enhancing the performance of neural networks for fraud detection. Their research highlighted the significance of selecting and transforming input features to maximize the effectiveness of neural networks in discerning fraudulent activities. The study by [25] shed light on the delicate balance between the flexibility of neural networks and the evolving landscape of credit card fraud. Building upon this foundation, [27] provided insights into the optimization of neural network architectures for enhanced fraud detection accuracy. Their research focused on the fine-tuning of hyperparameters and model architectures to achieve better performance in identifying fraudulent transactions. However, [28] critically evaluated the limitations of neural networks in capturing subtle and evolving fraud patterns. Their research emphasized scenarios where the sheer complexity of neural network models might hinder interpretability, posing challenges in understanding and validating the decision-making processes. This critical viewpoint contributed to a more nuanced understanding of the trade-offs associated with the application of neural networks in the specific context of credit card fraud detection.

In exploring the landscape of logistic regression in fraud detection, [29] demonstrated its viability as a baseline method, leveraging its simplicity and interpretability for initial analyses. [30] investigated the impact of ensemble methods involving logistic regression in enhancing fraud detection accuracy. Their research showcased the advantages of combining logistic regression into ensembles, leveraging the diversity of individual models to bolster overall system robustness. Furthermore, [31] delved

 doi:10.20944/preprints202501.1511.v1

into the improvement of logistic regression algorithms for credit card fraud detection. Their research focused on enhancing the algorithm's capacity to handle intricate fraud patterns, providing insights into the nuanced tuning of logistic regression models for improved performance. Logistic regression is not effective in capturing subtle and evolving fraud patterns, this limitation is associated with its simplicity, which denied its application to complex fraudulent tactics detection [32].

Rule-based algorithms have been a cornerstone in the landscape of credit card fraud detection, offering transparent decision-making processes and interpretability [33]. In a practical application, [34] implemented rule-based algorithms in real-world credit card transaction data, providing empirical insights into their effectiveness. Their research highlighted the adaptability of rule-based models to different types of fraud scenarios, offering a tangible demonstration of their applicability in authentic settings. [35] combined optimized rule-based algorithms and machine learning techniques to the evolving landscape of fraud patterns, showcasing the potential for rule-based algorithms to align with machine learning advancements. [36] explored the integration of rule-based algorithms with explainability techniques, aiming to bridge the gap between transparency and complexity. Rule-based model is deterministic in nature hence, struggle to capture the nuanced relationships within transactional data, underscoring the need for adaptive and learning-based approaches [37].

According to [15], there is a growing interest in hybrid data mining techniques as a promising solution to address the limitations of conventional methods. Hybrid methods combine diverse algorithms or approaches, leveraging the strengths of individual techniques while compensating for their weaknesses [38]. These hybrid approaches offer the potential to improve credit card fraud detection accuracy and reduce false positives, thus providing more resilient protection to individuals and financial institutions [18]. [18] further emphasized the significance of integrating multiple data mining techniques to advance fraud detection systems.

[39] proposed a hybrid model that combined linear regression (LR) and support vector machine (SVM). Leveraging on the interpretability prowess of LR, and SVM strength of capturing nonlinear patterns. The hybrid model achieved an accuracy of 94.5%, outperforming individual models. The model of [40] is a hybrid of Random Forest and neural network, leveraging on the efficiency of Random Forest to detect normal instances and neural network strength to detect frudulent transaction patterns. addressed the critical challenge of timely detecting fraudulent credit card transactions in the financial industry, emphasizing the highly skewed nature of the dataset with a small ratio of fraud to normal transactions. This hybrid approach, as per their findings, demonstrated high accuracy and confidence in predicting the label of new samples. [41] introduced hybrid approaches involving AdaBoost and majority voting methods. Their evaluation, initially using a publicly available credit card dataset and subsequently with real-world financial institution data, showed the robustness of hybrid models. The hybrid model with majority voting demonstrated superior accuracy, achieving an Matthew Correlation Coefficient score of 0.823. Furthermore, when applied to real financial institution data, the hybrid models attained a perfect MCC score of 1. The introduction of noise in data samples, ranging from 10% to 30%, validated the robustness of the majority voting method, with an impressive MCC score of 0.942 under 30% noise. This study underscores the efficacy of hybrid methods, particularly majority voting, in enhancing credit card fraud detection accuracy, even in the presence of noise. [42] addressed the escalating issue of credit card fraud in the digital era by proposing a framework that amalgamates three distinct techniques: Decision Trees, Neural Network, and K-Nearest Neighbour. In the quest for effective fraud detection, the researchers assigned labels to new transactions by considering the majority output from these methodologies. The study emphasized the adaptability of the model, asserting its competence across varied dataset sizes and types due to the synergistic benefits derived from the individual techniques.

Despite the successes recorded in addressing credit card fraud detection, there are still unresolved challenges that researchers in the field are working relentlessly to resolve. the challenges include:

- Ever evolving nature of fraud pattern.

 doi:10.20944/preprints202501.1511.v1

- Interpretability of models, which is of great concerns especially in industries where regulatory compliance and transparency are paramount.
- Imbalanced datasets.
- Developing scalable models that ensure timely detection without compromising accuracy.
- Hybrid models show promise in enhancing accuracy, there is a lack of consensus on the optimal combination of algorithms for specific contexts.

To address these gaps we propose a hybrid data mining approach that leverages the strengths of diverse algorithms to enhance adaptability, interpretability, and scalability. By systematically evaluating the performance of different algorithmic combinations, This research aims to contribute insights into the design and implementation of more effective credit card fraud detection systems. Additionally, this study will explore innovative techniques for handling imbalanced datasets and enhancing the interpretability of the models, ensuring that the approach aligns with industry needs. By focusing on these gaps, this research seeks to advance the field of credit card fraud detection, providing practical solutions that can be implemented in real-world scenarios. The outcomes of this research have the potential to significantly improve the accuracy and efficiency of fraud detection systems, ultimately reducing financial losses for individuals and institutions.

3. Methodology

The design and implementation of hybrid data mining techniques involve the integration of selected algorithms to enhance fraud detection performance. This approach is supported by the literature, which emphasizes the limitations of individual techniques and the potential benefits of combining them [21,25]. Serial hybrid techniques, where algorithms are applied sequentially, parallel hybrid techniques that run algorithms simultaneously, and integrated hybrid techniques embedding one algorithm into another will be explored [43–45].

The model architecture for credit card fraud detection involves two phases: the careful selection and integration of seven distinct machine learning classification algorithms, each chosen based on promising performances observed in the literature. The selected algorithms are eXtreme Gradient Boosting (XGBOOST), Light Gradient Boosting Machine (LGBM), decision trees, neural networks, random forests, support vector machines (SVM), and AdaBoost. In the first phase of the model development, a baseline machine learning classification model is established. Each of the seven algorithms is individually applied to evaluate its performance in detecting fraudulent transactions. Figure 1 shows the process involves in hybrid model baseline selection.

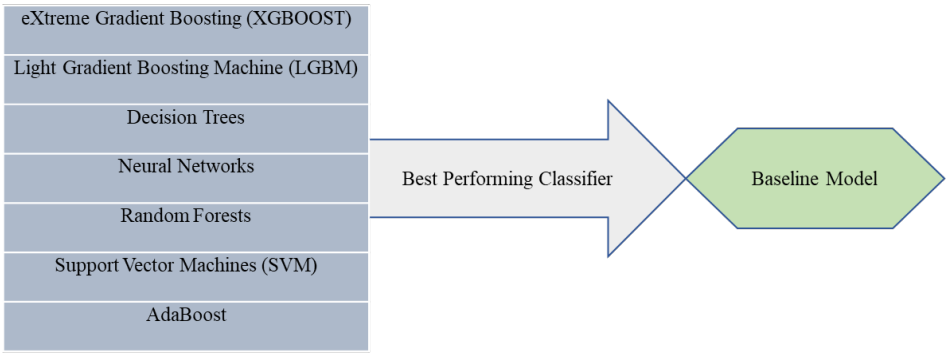

Figure 1. The first phase of the hybrid model development.

In the second phase, the algorithm exhibiting the highest performance in the first phase is identified as the baseline model. This algorithm is then employed to create the hybrid model as presented in Figure 2. In the process of creating hybrid models, a set of ensemble models were developed by combining the base Light Gradient Boosting Machine (LGBM) model with various other classifiers, namely eXtreme Gradient Boosting (XGBoost), Decision Trees, Neural Networks, Random Forests, Support Vector Machines (SVM), and AdaBoost. The ensemble models were formed using

 doi:10.20944/preprints202501.1511.v1

the Voting Classifier approach with soft voting, allowing for the aggregation of individual model predictions based on their probabilities. For instance, the LGBM + XGBoost hybrid model combines the predictive strengths of the LGBM base model with XGBoost. This amalgamation aims to leverage the diverse strengths of individual algorithms, creating a more robust and powerful predictive model. Each ensemble model underwent training on the resampled data and subsequent evaluation.

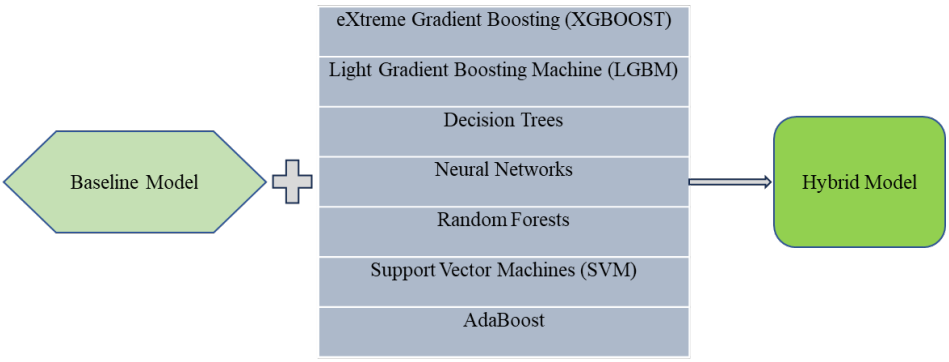

Figure 2. The first phase of the hybrid model development.

4. Experimentation

This section contains discussion of data exploration, data preparation and preprocessing, system setup and experiments, and results presentation and discussion.

4.1. Data Exploration

The dataset utilized in this research is a simulated credit card transaction dataset spanning the duration from January 1, 2019, to December 31, 2020. The simulation process was carried out using the Sparkov Data Generation tool created by [46]. This dataset is intentionally designed to encompass both legitimate and fraudulent transactions, providing a diverse and comprehensive foundation for training and testing the credit card fraud detection model. The dataset comprises transactions involving credit cards issued to 1000 customers engaging in transactions with a pool of 800 merchants.

The dataset exhibits a significant class imbalance, with approximately 1.84 million instances of legitimate transactions and about 10 thousand instances of fraudulent transactions. The initial dataset contained a vast number of transactions. However, to circumvent computational challenges and potential delays in model training, a smaller random subset sample, termed the Proof of Concept (POC) dataset, was created from the original dataset having 109,651 records and 23 features. It is important to note that the POC dataset maintains the same imbalance ratio as the original dataset.

Figure 3 provides valuable insights into the patterns and preferences targeted by potentially fraudulent activities. Among the categories, "shopping_net" and "grocery_pos" stand out with notably higher counts of fraudulent transactions, indicating that these areas are more susceptible to fraudulent activities. On the other hand, categories such as "travel" and "grocery_net" exhibit lower counts, suggesting a comparatively lower incidence of fraudulent transactions in these domains. This granular analysis of category-wise fraudulent transactions contributes to a nuanced understanding of the specific sectors within which credit card fraud is more prevalent, offering valuable guidance for the development of a targeted and effective credit card fraud detection model.

Figure 4 give the distribution of credit card fraud counts across age groups which challenges the conventional notion that older customers are inherently more susceptible to such fraudulent activities. Contrary to expectations, the data presents a nuanced picture, as reflected in the histogram. Notably, individuals between the ages of 30 and 60 emerge as more likely victims of credit card fraud, debunking the stereotype that older age groups are disproportionately affected. This finding underscores the necessity of leveraging data-driven insights rather than relying on preconceived notions, emphasizing the importance of tailoring fraud detection models to consider specific age demographics. By understanding the age-related patterns in credit card fraud, the model can better

 doi:10.20944/preprints202501.1511.v1

adapt its predictive capabilities to the actual distribution of fraudulent activities within different age groups, enhancing its accuracy and relevance in real-world scenarios.

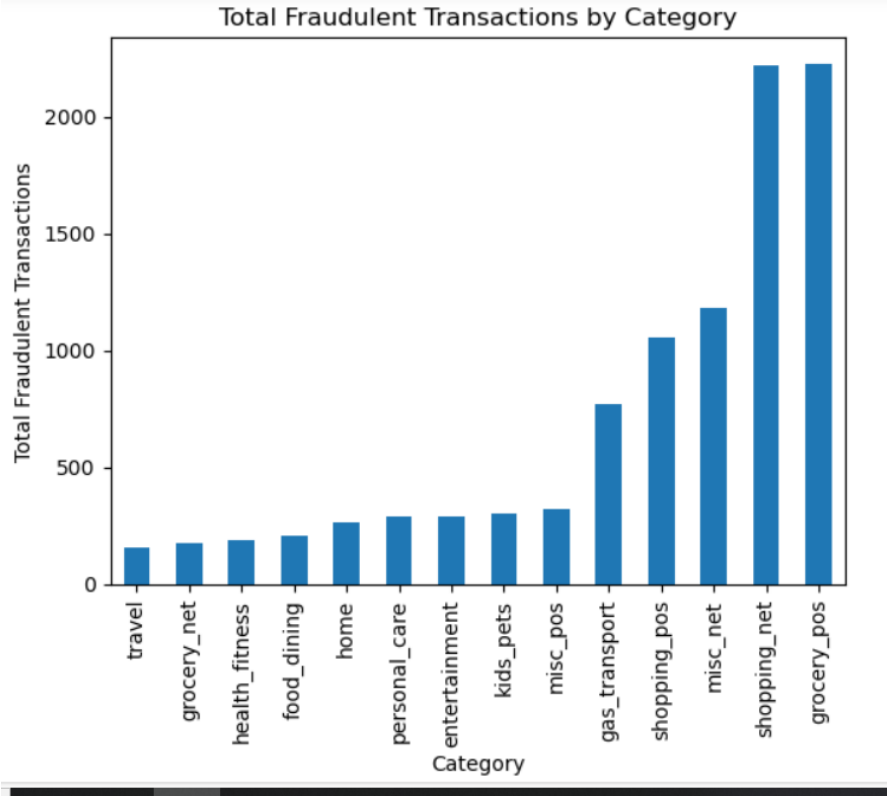

Figure 3. Distribution of Fraudulent Transactions by Merchant Category.

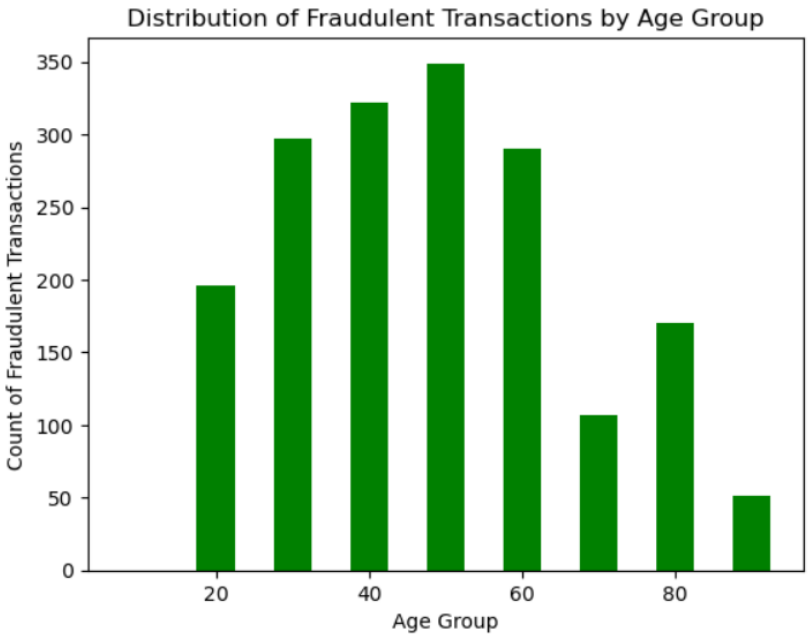

Figure 4. Distribution of Fraudulent Transactions by Age Group.

4.2. Data Preparation and Preprocessing

The data preparation phase is integral to ensuring the dataset is suitable for model training and evaluation. The procedure consider in this study include the transformation of categorical feature to numerical equivalent to ensure model compatibility. Next to data transformation is data scaling and normalization, which ensure uniformity in the range and distribution of data values. Min-Max scaling

 doi:10.20944/preprints202501.1511.v1

is employed in this research as the feature scaling technique. Min-Max scaling transforms features to a specific range, usually [0, 1], preserving the relationships between data points while preventing outliers from unduly affecting the model.

Data imbalance is then handled using the SMOTE-ENN (Synthetic Minority Over-Sampling Technique with Edited Nearest Neighbours) techniques proposed in [47]. This approach involves oversampling the minority class using SMOTE and then applying ENN to eliminate overlapping between classes, thereby producing well-defined class clusters. This preprocessing step is crucial for creating a more balanced and representative dataset for training the machine learning model, ultimately enhancing its ability to discern patterns associated with credit card fraud. Figures 5 and 6 show the effect of SMOTE-ENN on the dataset before and after its application, respectively.

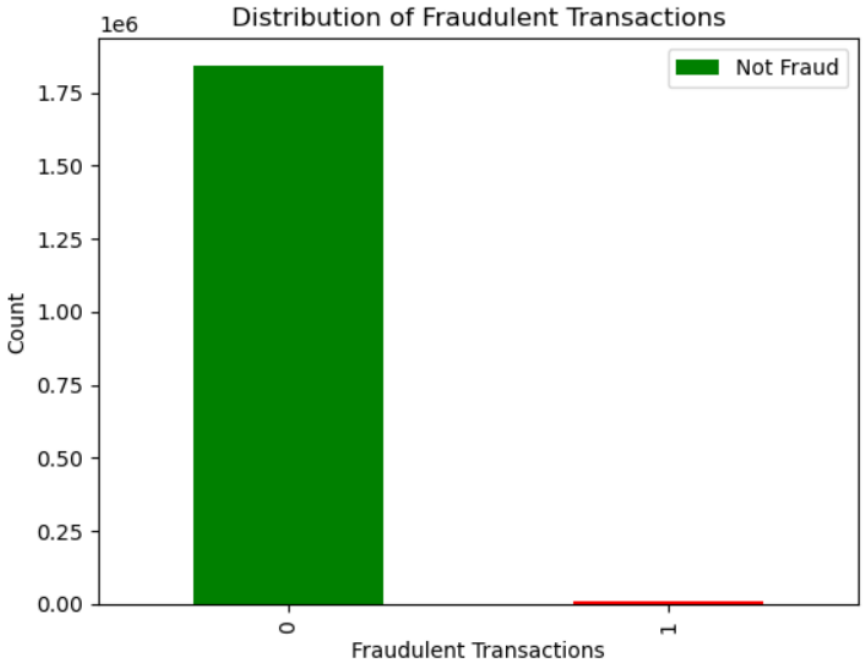

Figure 5. Class distribution before the application of SMOTE-ENN.

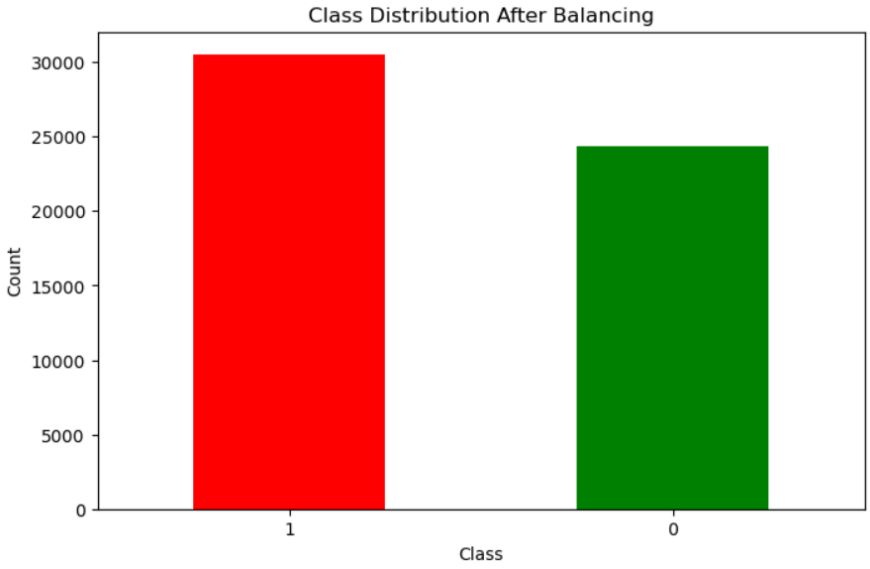

Figure 6. Class distribution after the application of SMOTE-ENN.

Lastly, the dataset is split into training and testing sets in the ratio of 80:20 respectively. The goal of this section is to create a robust and representative dataset that facilitates the development of an accurate and generalizable credit card fraud detection model.

 doi:10.20944/preprints202501.1511.v1

4.3. System setup and Experiments

The experiments are carried out in two stages: The initial stage of model training involved a comprehensive process using popular machine learning libraries and classifiers. The process was implemented using the scikit-learn library in Python, which provides a consistent interface for various machine learning algorithms. The dataset was split into training (80%) and testing (20%) sets using the train_test_split function from scikit-learn, a crucial step in ensuring the model's generalizability to unseen data. The models selected for training in the first stage were the eXtreme Gradient Boosting (XGBOOST), Light Gradient Boosting Machine (LGBM), Decision Trees, Neural Networks, Random Forests, Support Vector Machines (SVM), and AdaBoost. Distinct hyperparameters were carefully chosen for each classifier to optimize their individual performances. The XGBoost model was configured with a maximum tree depth of 3, a learning rate of 0.1, and 100 estimators. The LGBM model utilized 31 leaves, a learning rate of 0.05, and 100 estimators. Decision Trees were constructed without a specified maximum depth and with minimum samples split of 2. Neural Networks consisted of a single hidden layer with 100 neurons, using the rectified linear unit (ReLU) activation function and the Adam solver. Random Forests incorporated 100 estimators without a maximum depth restriction and minimum samples split of 2. The Support Vector Machines (SVM) model was configured with a regularization parameter (C) of 1.0, using a radial basis function (RBF) kernel, and enabling probability estimation. Lastly, the AdaBoost model included 50 weak learners with a learning rate of 1.0. These tailored hyperparameters were designed to strike a balance between computational efficiency and model accuracy across various algorithms. Each classifier was trained on the training set, and performance metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (ROC AUC) were calculated using scikit-learn metrics. The models were then evaluated on the test set to assess their generalization capabilities.

In the second stage of the experiments for hybrid model development, the Light Gradient Boosting Machine (LGBM) with the highest Area Under the Receiver Operating Characteristic (AUROC) in the first stage served as the base model. Ensemble models were created by combining the LGBM with each of the following classifiers: eXtreme Gradient Boosting (XGBOOST), Decision Trees, Neural Networks, Random Forests, Support Vector Machines (SVM), and AdaBoost. This ensemble approach aligns with the concept of model stacking or blending, where multiple models are combined to improve overall predictive performance. Each hybrid model underwent training using the same dataset split used in the initial stage. The performance of each ensemble model is then evaluated on the test set using key metrics such as accuracy, precision, recall, F1 score, and ROC AUC. These metrics provide a comprehensive understanding of the model's ability to correctly classify instances of fraud and non-fraud transactions. The final evaluation of these ensemble models will provide insights into the potential synergies achieved by combining LGBM with different algorithms and guide the selection of the most effective hybrid model for credit card fraud detection.

4.4. Result Presentation and Discussion

The result and discussion section provides a detailed evaluation of the model performance, comparing the proposed hybrid models with established machine learning algorithms. Initially, the single algorithms (XGBoost, LGBM, Decision Trees, Neural Networks, Random Forests, SVM, and AdaBoost) were individually assessed based on Accuracy scores. Notably, all single models exhibited relatively similar performance values, ranging from 0.62 to 0.84 reflecting their competence in credit card fraud detection.

The first stage of model development yields diverse results across various classifiers, highlighting the nuanced performance of each algorithm in credit card fraud detection. Among the classifiers, the XGBoost performed reasonably well with a 78% accuracy and 89% AUROC but the Light Gradient Boosting Machine (LGBM) stood out as a top performer, demonstrating notable accuracy (0.84), precision (0.76), recall (0.81), and an impressive Area Under the Receiver Operating Characteristic (AUROC) score of 0.82. LGBM's robust performance, particularly in correctly identifying fraudulent

 doi:10.20944/preprints202501.1511.v1

transactions, positions it as a strong candidate for the second stage of hybrid model development. On the other hand, some classifiers, such as Decision Trees and Neural Networks, exhibit comparatively lower performance metrics. Decision Trees, with an accuracy of 0.62 and a modest AUROC of 0.63, show limitations in capturing the complexity of the credit card fraud detection task. Neural Networks, while achieving a balanced accuracy (0.71) and recall (0.72), indicate challenges in precision (0.68), suggesting a propensity for false positives. The robust performance of LGBM positions it as a strong base model and combining it with other algorithms in the ensemble models aims to create a more comprehensive and effective fraud detection system. Table 1 contains summary of the classifiers.

Table 1. Performance evaluation of individual data mining model.

	Part				
Models	Accuracy	Precision	Recall	F1 Score	AUROC
XGBoost	0.78	0.73	0.78	0.75	0.89
LGBM	0.84	0.76	0.81	0.78	0.82
Decision Trees	0.62	0.61	0.55	0.58	0.63
Neural Networks	0.71	0.68	0.72	0.70	0.60
Random Forest	0.79	0.80	0.75	0.77	0.81
SVM	0.76	0.74	0.63	0.68	0.74
Adaboost	0.78	0.65	0.71	0.67	0.72

Details of hybrid models using LGBM as base model is presented in Table 2 and Figure 7. , the first hybrid model, LGBM + XGBoost, demonstrates superior accuracy (98.30%), precision (98.88%), and AUROC (99.80%) than XGBoost. The recall (98.05%) and F1 score (98.46%) indicate a balanced performance. The confusion matrix shows fewer false positives (68) and false negatives (119) compared to XGBoost, suggesting LGBM's ability to minimize both Type I and Type II errors. The hybridization of LGBM and Decision Trees shows commending results, the confusion matrix shows that 5992 fraud cases were correctly identified with 127 false positives and 122 false negatives. This implies that the hybrid model, while effective, may slightly misclassify certain transactions as fraudulent. Hybrid of LGBM and Neural Networks exhibit comparatively low recall performance implying a challenge in capturing all fraud instances. In the confusion matrix, the hybrid model shows a higher number of false negatives (1002), indicating a potential limitation in identifying all fraudulent transactions. The hybrid of LGBM and Random Forests shows good performance, it effectively minimizes false alarms but has the tendency to overlook some fraudulent transactions. The performance of the hybrid of LGBM + SVM is moderate, its confusion matrix reveals a higher number of false negatives (1067), indicating potential challenges in identifying all fraud cases. LGBM and AdaBoost hybrid presents a balanced performance and its confusion matrix reveals the model's effectiveness in fraud prediction.

Table 2. Performance evaluation of individual data mining model.

	Evaluation Metrics						
Models	Accuracy	Precision	Recall	F1 Score	AUROC	Type-1 Error	Type-2 Error
LGBM + XGBoost	0.9830	0.9888	0.9805	0.9846	0.9980	0.0140	0.0195
LGBM + Decision Trees	0.9773	0.9792	0.9800	0.9796	0.9770	0.0261	0.020
LGBM + Neural Networks	0.8847	0.9511	0.8361	0.8899	0.9472	0.0541	0.1639
LGBM + Random Forest	0.9750	0.9831	0.9719	0.9775	0.9967	0.0210	0.0281
LGBM + SVM	0.8693	0.9322	0.8255	0.8756	0.9363	0.0755	0.1745
LGBM + Adaboost	0.9214	0.9615	0.8948	0.9270	0.9689	0.0451	0.1052

 doi:10.20944/preprints202501.1511.v1

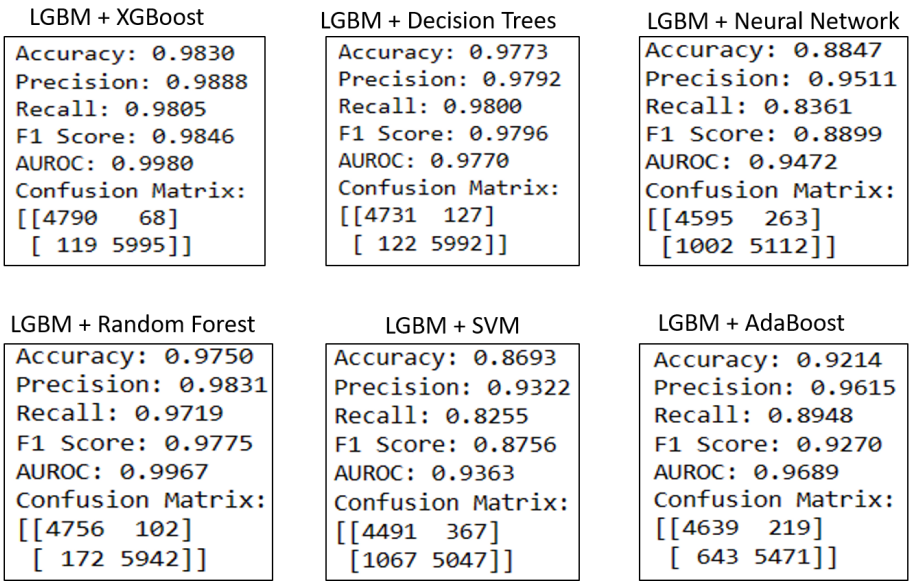

Figure 7. Hybrid models performance and their confusion matrix.

5. Conclusions

This research has successfully addressed its primary objective of exploring and evaluating the efficiency of hybrid data mining methods in credit card fraud detection. By systematically achieving each of the defined research objectives, the study has contributed valuable insights to the field. The collection and exploration of an appropriate dataset enabled the development and training of a hybrid data mining models. The models, incorporating multiple algorithms, aimed to enhance accuracy and decrease false positive rates. Through the model development and training process, the research demonstrated a meticulous approach in comparing various classifiers, employing both traditional and hybrid models. The analysis of the first stage model training results emphasized the strengths and limitations of individual classifiers, providing a basis for selecting the top performer (LGBM) as the base model for the second stage. The hybrid model was developed by permutating other selected algorithms with the base model (LGBM) considering their strengths. Observation from the results shows that hybrid of LGBM and XGBoost outperformed other hybrid models when evaluated using key metrics, including accuracy, precision, recall, and AUROC, Type- Error, Type-2 Error and confusion matrix. While this study considers synthetic data, future work will consider generalization of the most effective model to real-world credit card fraud detection characterized with consistence fraud evolution. Furthermore, this study does not account for the possibility of external factors like political influence, technological advancement and economical conditions that could have significant effect on credit card fraud pattern. Understanding the impact of these external factors on fraud dynamics could contribute to a more holistic fraud detection framework.

References

1. Pandey, K.; Sachan, P.; Ganpatrao, N.G.; et al. A review of credit card fraud detection techniques. In Proceedings of the 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, 2021, pp. 1645–1653.
2. Lokanan, M.E. Financial fraud detection: the use of visualization techniques in credit card fraud and money laundering domains. *Journal of Money Laundering Control* **2022**, *26*, 436–444.
3. The Nilson Report. Global Card Fraud Losses Continue to Rise, 2018.
4. The Nilson Report. Global Card Fraud Losses Continue to Rise, 2022.
5. AARP. Identity Fraud Report 2023. https://www.aarp.org/money/scams-fraud/info-2024/identity-fraud-report.html, 2023. Accessed: YYYY-MM-DD.
6. Experian. Steps to Take if You Are the Victim of Credit Card Fraud. https://www.experian.com/blogs/ask-experian/steps-to-take-if-you-are-the-victim-of-credit-card-fraud/, n.d.

7. Thennakoon, A.; Bhagyani, C.; Premadasa, S.; Mihiranga, S.; Kuruwitaarachchi, N. Real-time credit card fraud detection using machine learning. In Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019, pp. 488–493.
8. Kamusweke, K.; Nyirenda, M.; Kabemba, M. Data mining for fraud detection in large scale financial transactions. *EasyChair* **2019**.
9. Lim, K.S.; Lee, L.H.; Sim, Y.W. A review of machine learning algorithms for fraud detection in credit card transaction. *International Journal of Computer Science & Network Security* **2021**, *21*, 31–40.
10. Barman, S.; Pal, U.; Sarfaraj, M.A.; Biswas, B.; Mahata, A.; Mandal, P. A complete literature review on financial fraud detection applying data mining techniques. *International Journal of Trust Management in Computing and Communications* **2016**, *3*, 336–359.
11. Padvekar, S.A.; Kangane, P.M.; Jadhav, K.V. Credit card fraud detection system. *International Journal Of Engineering And Computer Science* **2016**.
12. Mathur, S.; Daniel, S. It's Fraud! Application of Machine Learning Techniques for Detection of Fraudulent Digital Advertising. *Webology* **2022**, *19*, 2475–2490.
13. Cortez, P.; Embrechts, M.J. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences* **2013**, *225*, 1–17.
14. Maleki, F.; Muthukrishnan, N.; Ovens, K.; Reinhold, C.; Forghani, R. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics* **2020**, *30*, 433–445.
15. Orzechowski, P.; Boryczko, K. Hybrid biclustering algorithms for data mining. In Proceedings of the Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30–April 1, 2016, Proceedings, Part I 19. Springer, 2016, pp. 156–168.
16. Xie, Y.; Li, A.; Gao, L.; Liu, Z. A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wireless Communications and Mobile Computing* **2021**, *2021*, 2531210.
17. Kim, E.; Lee, J.; Shin, H.; Yang, H.; Cho, S.; Nam, S.k.; Song, Y.; Yoon, J.a.; Kim, J.i. Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications* **2019**, *128*, 214–224.
18. Phua, C.; Lee, V.; Smith, K.; Gayler, R. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119* **2010**.
19. Bhowmik, R. Data mining techniques in fraud detection. *Journal of Digital Forensics, Security and Law* **2008**, *3*, 3.
20. Bagga, S.; Goyal, A.; Gupta, N.; Goyal, A. Credit card fraud detection using pipeling and ensemble learning. *Procedia Computer Science* **2020**, *173*, 104–112.
21. Jain, R.; Gour, B.; Dubey, S. A hybrid approach for credit card fraud detection using rough set and decision tree technique. *International Journal of Computer Applications* **2016**, *139*, 1–6.
22. Zareapoor, M.; Seeja, K.; Alam, M.A. Analysis on credit card fraud detection techniques: based on certain design criteria. *International journal of computer applications* **2012**, *52*.
23. Sharma, P.; Banerjee, S.; Tiwari, D.; Patni, J.C. Machine learning model for credit card fraud detection-a comparative analysis. *Int. Arab J. Inf. Technol.* **2021**, *18*, 789–796.
24. Leevy, J.L.; Hancock, J.; Khoshgoftaar, T.M. Comparative analysis of binary and one-class classification techniques for credit card fraud data. *Journal of Big Data* **2023**, *10*, 118.
25. Dornadula, V.N.; Geetha, S. Credit card fraud detection using machine learning algorithms. *Procedia computer science* **2019**, *165*, 631–641.
26. Bahnsen, A.C.; Aouada, D.; Stojanovic, A.; Ottersten, B. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications* **2016**, *51*, 134–142.
27. Li, Z.; Li, J.; Wang, Y.; Wang, K. A deep learning approach for anomaly detection based on SAE and LSTM in mechanical equipment. *The International Journal of Advanced Manufacturing Technology* **2019**, *103*, 499–510.
28. Dighe, D.; Patil, S.; Kokate, S. Detection of credit card fraud transactions using machine learning algorithms and neural networks: A comparative study. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018, pp. 1–6.
29. Alenzi, H.Z.; Aljehane, N.O. Fraud detection in credit cards using logistic regression. *International Journal of Advanced Computer Science and Applications* **2020**, *11*.
30. Zhu, H.; Liu, G.; Zhou, M.; Xie, Y.; Abusorrah, A.; Kang, Q. Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection. *Neurocomputing* **2020**, *407*, 50–62.

 doi:10.20944/preprints202501.1511.v1

31. Chen, M. Credit Card Fraud Detection Based on Multiple Machine Learning Models. In Proceedings of the Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering, 2022, pp. 1801–1805.
32. Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; Bontempi, G. Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems* **2017**, *29*, 3784–3797.
33. Cherif, A.; Badhib, A.; Ammar, H.; Alshehri, S.; Kalkatawi, M.; Imine, A. Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University-Computer and Information Sciences* **2023**, *35*, 145–174.
34. Kültür, Y.; Çağlayan, M.U. Hybrid approaches for detecting credit card fraud. *Expert Systems* **2017**, *34*, e12191.
35. Kumar, J.; Saxena, V. Rule-based credit card fraud detection using user's keystroke behavior. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2021*; Springer, 2022; pp. 469–480.
36. Li, W.; Paraschiv, F.; Sermpinis, G. A data-driven explainable case-based reasoning approach for financial risk detection. *Quantitative Finance* **2022**, *22*, 2257–2274.
37. Adewumi, A.O.; Akinyelu, A.A. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management* **2017**, *8*, 937–953.
38. Delimata, P.; Suraj, Z. Hybrid methods in data classification and reduction. *Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam: Volume 2* **2013**, pp. 263–291.
39. Tripathi, K.K.; Ragha, L. Hybrid approach for credit card fraud detection. *Int. J. Soft Comput. Eng.(IJSCE)* **2013**, *3*.
40. Sohony, I.; Pratap, R.; Nambiar, U. Ensemble learning for credit card fraud detection. In Proceedings of the Proceedings of the ACM India joint international conference on data science and management of data, 2018, pp. 289–294.
41. Randhawa, K.; Loo, C.K.; Seera, M.; Lim, C.P.; Nandi, A.K. Credit card fraud detection using AdaBoost and majority voting. *IEEE access* **2018**, *6*, 14277–14284.
42. Tiwari, P.; Mehta, S.; Sakhuja, N.; Gupta, I.; Singh, A.K. Hybrid method in identifying the fraud detection in the credit card. In Proceedings of the Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2020. Springer, 2021, pp. 27–35.
43. Dai, Y.; Yan, J.; Tang, X.; Zhao, H.; Guo, M. Online credit card fraud detection: a hybrid framework with big data technologies. In Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2016, pp. 1644–1651.
44. Ojugo, A.A.; Nwankwo, O. Spectral-cluster solution for credit-card fraud detection using a genetic algorithm trained modular deep learning neural network. *JINAV: Journal of Information and Visualization* **2021**, *2*, 15–24.
45. Mienye, I.D.; Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **2022**, *10*, 99129–99149.
46. Harris, B. Sparkov Data Generation Tool. https://github.com/namebrandon/Sparkov_Data_Generation, 2022.
47. Makki, S.; Assaghir, Z.; Taher, Y.; Haque, R.; Hacid, M.S.; Zeineddine, H. An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access* **2019**, *7*, 93010–93022.