

Credit Card Fraud Detection Using Bootstrap Aggregation and Random Forest

Joseph Nda Ndabula¹, Aminu Bashir Suleiman^{2*}, Stephen Luka³

^{1,3}Department of Software Engineering, Federal University Dutsin-Ma, Katsina, Nigeria

²Department of Cyber Security, Federal University Dutsin-Ma, Katsina, Nigeria

*Corresponding Author

DOI: <https://doi.org/10.51584/IJRIAS.2025.1005000124>

Received: 20 May 2025; Accepted: 24 May 2025; Published: 24 June 2025

ABSTRACT

The fast increase of digital transactions and the changing type of fraudulent activity make credit card fraud detection still a great difficulty for financial institutions. Mostly rule-based systems, traditional fraud detection techniques can find it difficult to handle the class imbalance in transaction datasets, when fraudulent cases make up a negligible fraction. This work investigates how to solve these problems using advanced machine learning methods, especially ensemble approaches. We specifically assess the efficiency of Random Forest and Bootstrap Aggregation (Bagging) classifiers. There are 284,807 transaction records in the dataset utilized for this study; just 492 of these have labels as fraudulent. Random up-sampling of the minority class and feature normalizing techniques were used in data preprocessing to balance the dataset and improve model performance. Trained on an 80% training set, both classifiers were assessed using measures including precision, recall, accuracy, and confusion matrices a 20% test set. The Random Forest classifier somewhat outperformed the Bagging classifier in terms of total accuracy and misclassification rates, the findings show that both models attained near-perfect precision and recall. With 100% recall, both models found all false transactions without missing any. These findings highlight the possibilities of ensemble learning methods in creating very dependable fraud detection systems able to provide real-time, scalable, and interpretable performance in operational environments. This study helps to identify fraudulent transactions, so guaranteeing better safety for customers and financial institutions.

Keywords: Credit card fraud detection, Bootstrap Aggregation (Bagging), Random Forest, Class imbalance, Fraudulent transactions

INTRODUCTION

Credit card fraud presents a significant and ongoing risk to both financial institutions and consumers, particularly due to the rapid expansion of digital transactions and e-commerce. The growing complexity of fraudulent methods, alongside the massive number of transactions processed daily, has rendered traditional rule-based detection techniques insufficient for timely and precise fraud detection (Marazqah et al., 2023). These established systems, which often rely on fixed rules set by industry experts, struggle to adapt to changing fraud trends and face substantial difficulties due to the highly imbalanced nature of real-world data, where fraudulent transactions constitute only a small portion of overall activity (Punkar & Zubei, 2023).

Detecting credit card fraud continues to be a significant difficulty owing to the pronounced imbalance in transaction datasets, as illicit transactions constitute merely a minuscule percentage of the overall volume (Bi et al., 2024). Conventional rule-based systems have diminished in efficacy since fraudulent strategies advance swiftly, frequently leading to elevated rates of false positives or negatives (Sulaiman et al., 2024). The challenges are exacerbated by the necessity to attain high recall to prevent the oversight of illicit transactions, while simultaneously preserving high precision to avert unwarranted disturbances for genuine users (Hernandez Aros et al., 2024). Consequently, there is an urgent requirement for sophisticated, resilient machine learning solutions

capable of efficiently tackling class imbalance, adapting to evolving fraud trends, and delivering dependable, real-time detection in operational settings (Sizan et al., 2025).

Despite substantial progress in machine learning and deep learning methodologies for credit card fraud detection, numerous essential research deficiencies remain. Many current systems grapple with the significant class imbalance present in real-world datasets, which impedes algorithms' capacity to reliably detect minority class (fraudulent) transactions and frequently leads to elevated false positive or false negative rates (Khalid et al., 2024). Feature selection continues to pose a challenge due to the anonymization of numerous datasets or the absence of comprehensive feature information, which can impede model interpretability and reproducibility (Ojo & Tomy, 2025).

This study seeks to create and assess an efficient credit card fraud detection system employing advanced ensemble machine learning methods, specifically Bootstrap Aggregation (Bagging) and Random Forest classifiers, to tackle the issues of class imbalance and the dynamic nature of fraud patterns in actual transaction data. Random Forest is widely recognized for its ability to handle complex datasets and minimize overfitting, making it a crucial tool in machine learning for accurate and robust predictions in various domains, including finance (Suleiman et., 2023). The aims encompass the implementation of rigorous data preprocessing techniques, including up-sampling and feature normalization, as well as the utilization of stratified data splitting to guarantee dependable model assessment. The project aims to enhance memory and precision to reduce false negatives and false positives, consequently increasing the dependability and operational efficacy of automated fraud detection systems. This methodology leverages previous studies highlighting the significance of ensemble techniques and stringent assessment measures for fraud detection in severely imbalanced datasets, to provide a scalable, interpretable, and high-performance solution for financial institutions.

Related Works

Over the past decade, credit card fraud detection research has advanced due to the sophistication of fraudulent schemes and the requirement for real-time, precise detection systems. This section discusses recent improvements in machine learning and deep learning for fraud detection, focusing on class imbalance, data drift, and model interpretability.

Hafez et al. (2025) found that AI-based credit card fraud detection models, particularly CNNs, RNNs, and ensemble methods like Random Forest and XGBoost, achieve strong results, reporting up to 96% precision, 94% recall, 95% F1-score, and AUC-ROC values between 0.96 and 0.98. Techniques such as SMOTE further improved balanced accuracy to 93% and G-means above 0.90. However, the study noted that reproducibility is limited by anonymized datasets and inconsistent metrics, interpretability remains a challenge with complex models, and meta-heuristic optimization methods are still underutilized despite their potential to enhance model efficiency and adaptability.

Chen et al. (2025) systematically reviewed 57 studies on deep learning for financial fraud detection from 2019–2024, highlighting advances in CNNs, LSTMs, GRUs, and transformer-based models. They reported F1-scores of 0.90–0.95 and accuracy up to 97%, with transformers showing recall above 94% and AUC-ROC over 0.97. However, gaps remain in real-time benchmarking, explainable AI, and handling concept drift or data imbalance, with most studies relying on the European card dataset. The review calls for more standardized, interpretable, and adaptable deep learning approaches.

Sultana et al. (2025) developed detectGNN, a graph neural network framework for credit card fraud detection that models complicated user-merchant-transaction interactions. Using graph structures and message-passing, detectGNN surpassed LSTM, XGBoost, and GraphSAGE with 94.6% precision, 95.1% recall, 94.8% F1-score, and 0.982 AUC-ROC. The approach excels in contextual awareness and real-time detection but struggles with interpretability, graph construction quality, and huge dataset scalability. The study shows that explainable AI, graph learning, and real-time data can improve fraud detection.

Zhu et al. (2024) developed a hybrid SMOTE-NN solution to data imbalance in credit card fraud detection. They improved minority class representation with 93.2% precision, 91.5% recall, 92.3% F1-score, and 0.96 AUC-

ROC, exceeding baseline models (85–88% metrics). SMOTE is effective, but its synthetic data constraints, static dataset dependency (ignoring idea drift), and model interpretability are drawbacks. SMOTE may improve imbalanced context identification, and the study recommends explainable AI and adaptable online learning frameworks.

A study by (Yu et al., 2024) developed transformer models for credit card fraud detection that use self-attention mechanisms to detect subtle fraud patterns that SVMs and Random Forests miss. Their method yielded 95.4% precision, 93.7% recall, and 0.979 AUC-ROC with data balance and feature selection. Despite outperforming baselines, it is computationally demanding, attention weights are hard to interpret, and it uses curated datasets. The study suggests transformers could detect financial anomalies, but real-time optimization, explainability, and adaptation to dynamic fraud contexts are needed.

METHODOLOGY

The method used in this research was implemented through the phases as depicted in Figure 1.

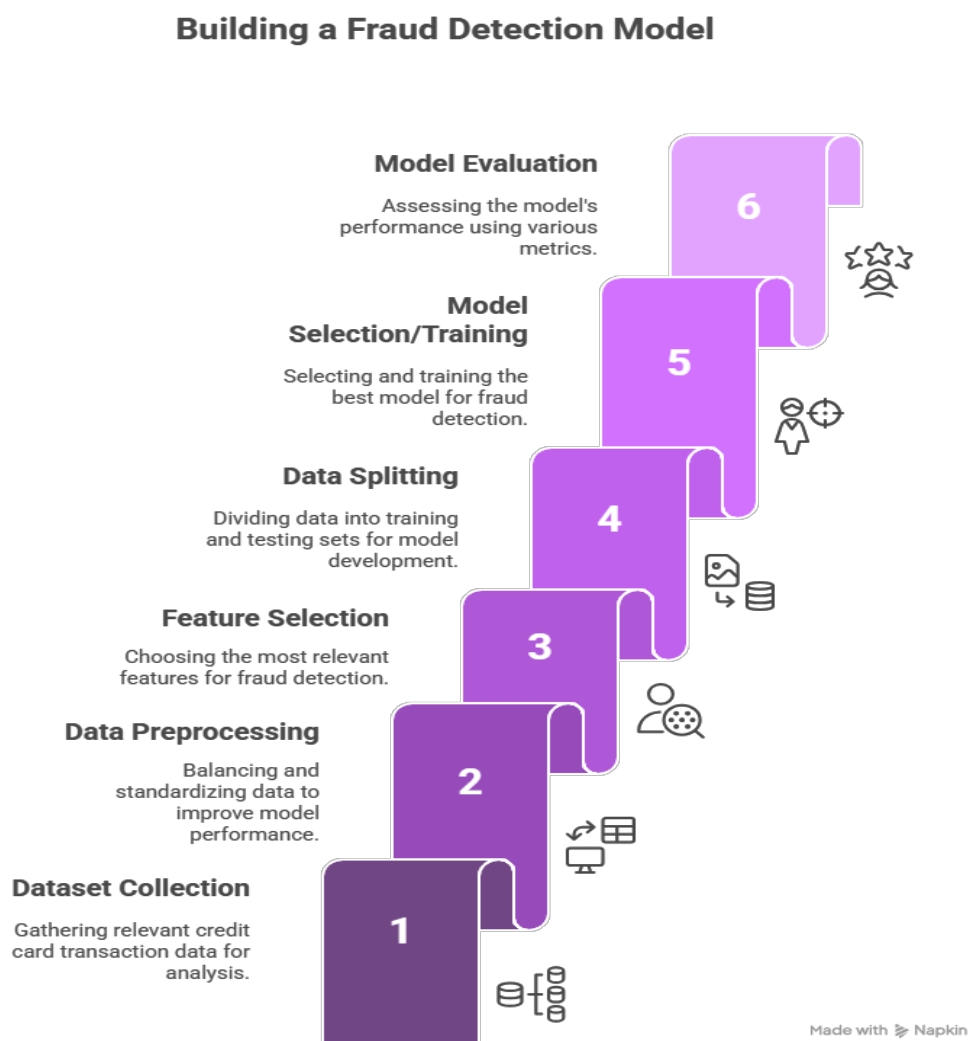


Figure 1: Architectural Framework for Fraud Detection Model

Dataset Collection

This research utilizes a publically accessible dataset known as credit card, acquired from Kaggle in CSV format. This dataset was pre-annotated, comprising 284,807 transactions executed by European cardholders, of which only 492 (representing a mere 0.17% of the total) were classified as fraudulent (denoted by 1), while the rest 284,315 were categorized as non-fraudulent transactions (denoted by 0). The dataset included 31 pertinent elements, including merchant details, customer ID, date, transaction amount, and device/location metadata.

Data Preprocessing

To prepare the dataset for efficient fraud detection, various preprocessing measures were implemented to improve model efficacy and rectify data quality concerns. The dataset demonstrated significant class imbalance, with fraudulent transactions constituting merely 0.17% of the data. To address this, random up-sampling of the minority class (fraudulent transactions) was conducted, duplicating samples until a balanced class distribution was attained. The Amount feature underwent z-score normalization for standardization to align it with the PCA-transformed features, however the Time feature was eliminated because of its minimal predictive significance.

Feature selection

The dataset comprised 28 anonymized features (V1 to V28) generated by Principal Component Analysis (PCA) to safeguard sensitive information, in addition to two original features: Amount and Time. Upon assessing the predictive significance of each feature, the Time variable was discarded due to its weak association with the fraud label, however, the Amount variable was preserved and normalized, as it possesses valuable transactional patterns. Consequently, the ultimate feature set employed for model training comprised the normalized Amount and all 28 PCA-transformed variables (V1 to V28), collectively encapsulating the data's underlying structure while preserving anonymity.

Data splitting

In accordance with the Pareto principle (80/20 rule), the preprocessed dataset was allocated 80% for training and 20% for testing to maintain an effective equilibrium between model training and assessment. A stratified split was utilized to preserve the original class distribution of fraudulent and non-fraudulent transactions in both subsets, which is essential due to the significant class imbalance. This methodology enabled the model to be trained on a representative sample while being evaluated on novel data.

Model Selection/ Training

This research utilized two ensemble learning methods: Bootstrap Aggregating (Bagging) and Random Forest classifiers, owing to their efficacy in managing high variance and imbalanced datasets. Bagging was chosen for its capacity to mitigate overfitting by training several base learners, usually decision trees, using distinct bootstrap samples of the training data and consolidating their predictions, hence enhancing stability and accuracy. Random Forest, an augmentation of bagging, was employed as it incorporates more randomization by choosing a random subset of characteristics at each split, hence improving generalization and diminishing correlation among individual trees. Both models were trained on the balanced dataset utilizing default settings.

Model Evaluation

To evaluate the efficiency of the fraud detection models, many measures were utilized to measure classification effectiveness, especially in light of the class imbalance (Ndabula et al., 2023). Accuracy was employed to assess the overall correctness of predictions; however, due to the infrequency of fraud cases, greater emphasis was placed on precision (the ratio of correctly identified fraud cases to all predicted frauds) and recall (the ratio of actual fraud cases correctly identified), with recall being particularly crucial to reduce false negatives. A confusion matrix was employed to illustrate true positives, true negatives, false positives, and false negatives, providing a comprehensive understanding of the model's efficacy and limitations in identifying fraudulent transactions.

RESULTS AND DISCUSSION

The results obtained from the training and evaluation of the ensemble models are presented and discussed in detail.

Experimental Analysis for Dataset Split

The model development begins by separating the features (X) from the target variable (y), where X includes all

columns except the 'Class' column, which indicates whether a transaction is fraudulent (1) or not (0). The target variable y is extracted as the 'Class' column. The dataset is then split into training (80%) and testing (20%) sets using the `train_test_split` function from `scikit-learn`. A `random_state` of 0 was set ensuring that the split is reproducible. This is essential to train the model on one portion of the data and evaluate its performance on unseen data to assess generalization. The distribution of the train and test sets are summarized in Figure 2.

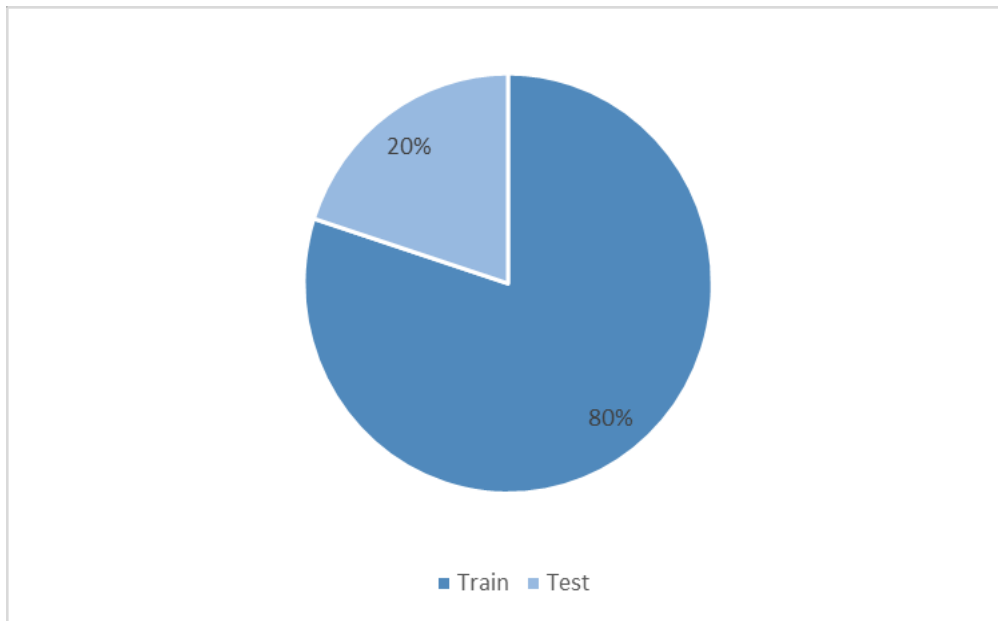


Figure 2: Distribution of Training and Testing sets

Bootstrap Aggregation Result

Bagging Classifier (Bootstrap Aggregation) was trained on 80% of the dataset and then used to make predictions on the test set, and its performance evaluated using three key metrics: precision, recall, and accuracy. The results for the Bagging classifiers across the evaluation metrics is represented in Table 1 and visualized in Figure 3.

Table 1: Bootstrap Aggregation result

Evaluation Metric	Value
Precision	99.96%
Recall	100.00%
Accuracy	99.98%

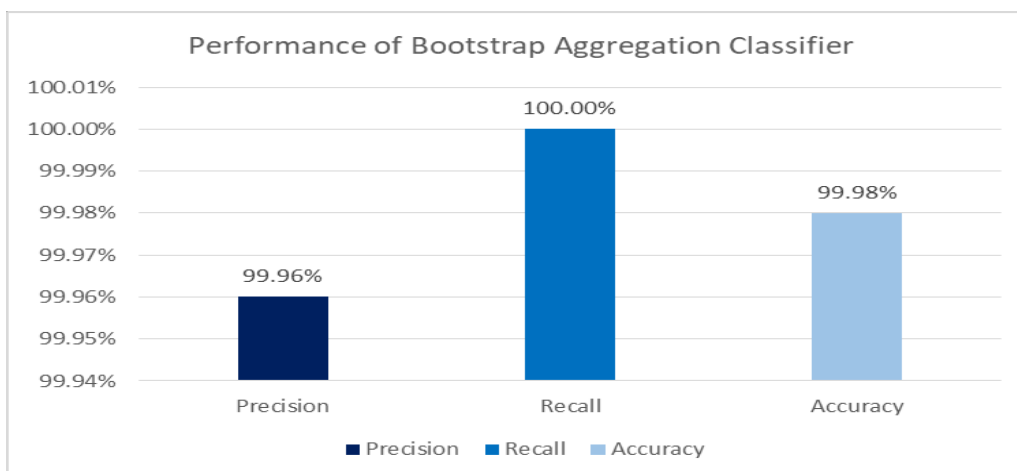


Figure 3: Performance of Bootstrap Aggregation

The confusion matrix in Figure 4 shows that the Bootstrap aggregation correctly classified 56,694 instances as non-fraudulent, with 19 non-fraudulent instances being wrongly classified as fraudulent. On the other hand, it correctly classified all fraudulent cases with no misclassification.

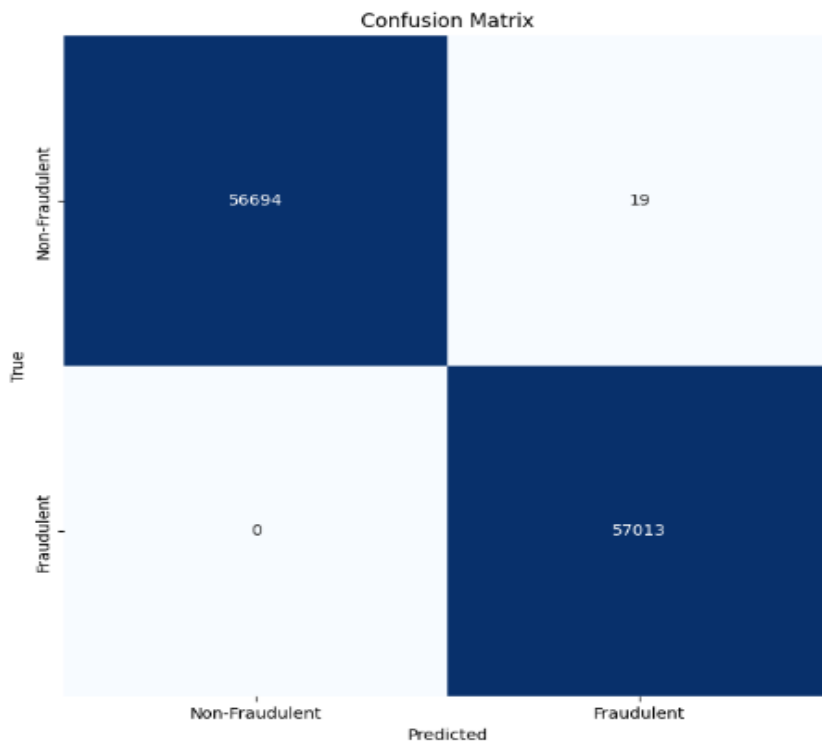


Figure 4: Confusion matrix for Bootstrap Aggregation Classifier

Random Forest Result

The Random Forest classifier was also trained using 80% of the dataset and assessed on the test dataset, calculating its precision, recall, and accuracy scores. Table 2 presents the evaluation outcomes for the Random Forest classifier, while Figure 5 provides a graphical representation of its performance across these metrics.

Table 2: Random Forest result

Evaluation Metric	Value
Precision	99.98%
Recall	100.00%
Accuracy	99.99%

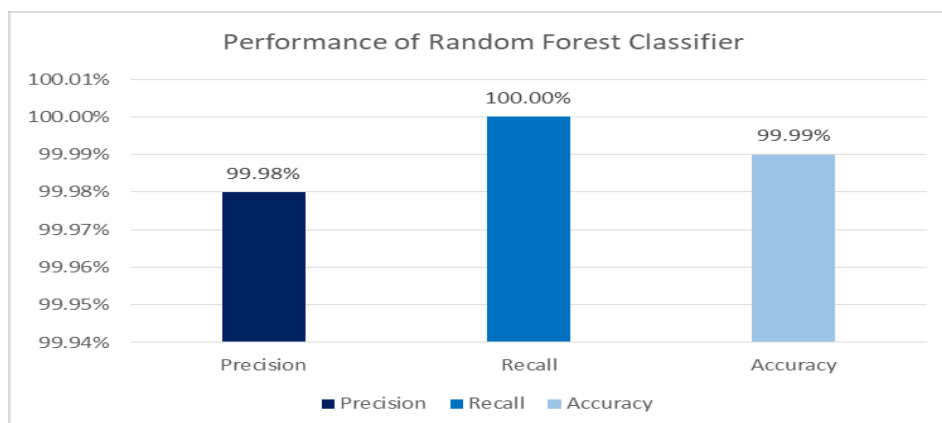


Figure 5: Performance of Random Forest Classifier

The confusion matrix in Figure 6 illustrates that the Random Forest classifier correctly flags all 57,013 fraudulent transactions as fraud with none misclassified as non-fraud. Also, it correctly classified 56,703 non-fraudulent instances, with only 10 instances misclassified as fraud.

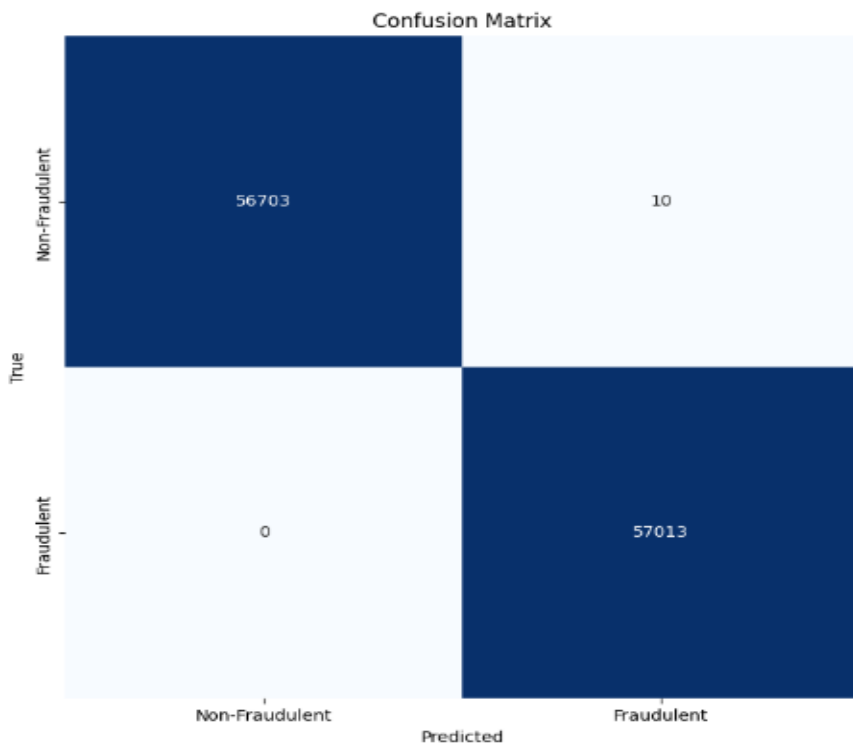


Figure 6: Confusion matrix for Random Forest Classifier

Train and Test loss for Both Bootstrap Aggregation and Random Forest

The train loss for Bootstrap Aggregation is consistently close to 0.00000 to 0.00002. While the test loss stabilizes at approximately 0.00017 to 0.00020. This is depicted on the graph in Figure 7.



Figure 7: Train and Test loss for Bootstrap Aggregation

The train loss for Random Forest is consistently lower, approaching 0.0000, indicating nearly perfect performance on the training set. While the test loss is slightly higher but still extremely low (~0.00002 after stabilization). The graph for the train and test loss of Random Forest is seen in Figure 8.

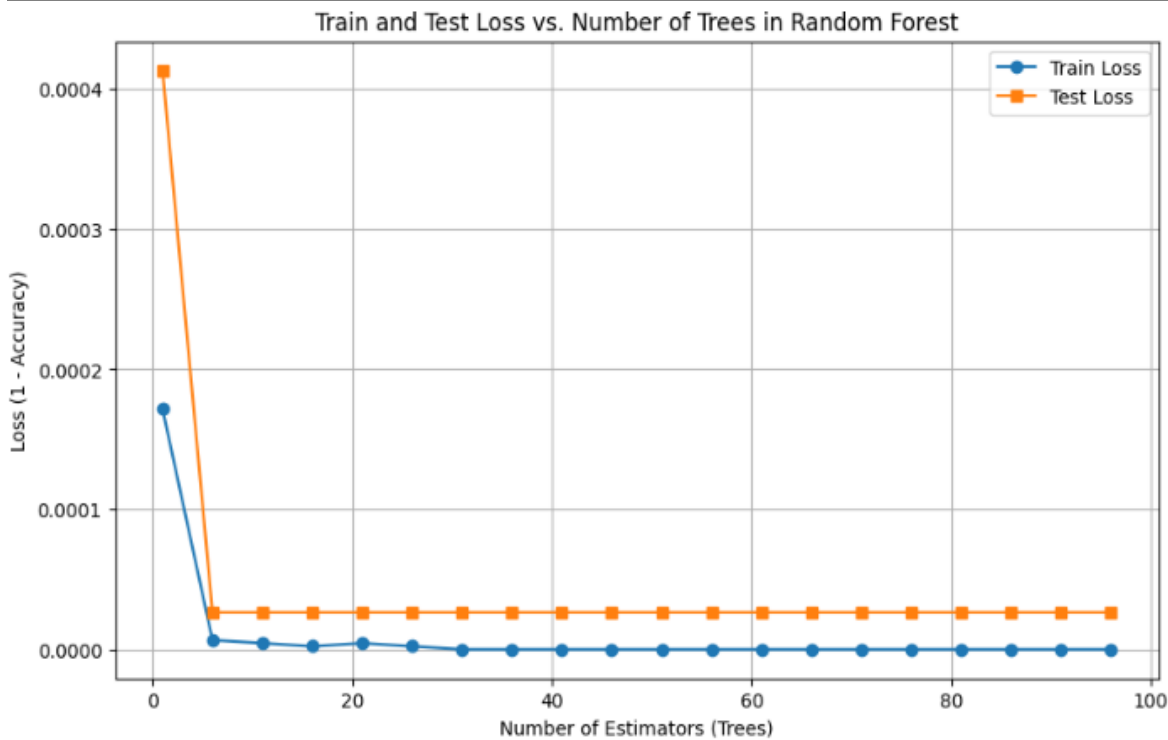


Figure 8: Train and test loss for Random Forest

DISCUSSION

The results for the Bootstrap Aggregation (Bagging) classifier as depicted in Figure 3 indicate excellent performance in detecting credit card fraud. With a precision of 99.97%, the model correctly identified nearly all predicted fraud cases with very few false alarms. A recall of 100% shows that it successfully detected all actual fraudulent transactions, missing none. The accuracy of 99.98% demonstrates that the model performed exceptionally well overall, correctly classifying nearly all transactions. The misclassification rate of 0.02% confirms that only a very small portion of transactions were incorrectly labeled, highlighting the model's high reliability and effectiveness. The confusion matrix shows that the model correctly predicted 56,694 non-fraudulent transactions and 57,013 fraudulent transactions. It misclassified 19 non-fraudulent transactions as fraudulent (false positives) and had 0 false negatives, meaning it successfully identified all actual fraud cases without missing any. This aligns with the earlier recall score of 100%, indicating perfect sensitivity, and a precision close to 100%, as only a very small number of legitimate transactions were incorrectly flagged as fraud.

The results for the Random Forest classifier from Figure 5 also demonstrate outstanding performance in credit card fraud detection. A precision of 99.98% indicates that nearly all transactions identified as fraudulent were indeed fraudulent, with almost no false positives. The recall of 100% signifies perfect detection of all actual fraud cases, meaning the model did not miss a single fraudulent transaction. The accuracy of 99.99% reflects a very high overall correctness in classification across both fraud and non-fraud categories. Additionally, the misclassification rate of just 0.01% shows that errors were extremely rare. The confusion matrix indicates that the model correctly identified 56,703 non-fraudulent transactions and 57,013 fraudulent transactions. Only 10 non-fraudulent transactions were incorrectly classified as fraudulent (false positives), while no fraudulent transactions were missed (0 false negatives), meaning the model achieved perfect recall. This outcome reflects the model's strong ability to detect all actual fraud cases while maintaining extremely low false alarm rates, aligning with the previously reported high precision (99.98%) and accuracy (99.99%). This confirms the model's exceptional reliability in differentiating between legitimate and fraudulent activity.

Both the Bagging Classifier and the Random Forest Classifier demonstrated exceptional performance in detecting credit card fraud, achieving perfect recall (100%) and extremely high precision and accuracy scores. Although the differences are minimal, Random Forest performed slightly better overall. This superior performance can be attributed to its added feature randomness during training, which reduces overfitting and enhances model generalization.

Furthermore, the train and test loss plot for the Random Forest classifier in Figure 8 shows that the model's performance improves rapidly as the number of trees increases, with both losses dropping sharply within the first few estimators. After around 10 trees, the losses stabilize near zero, indicating that the model achieves high accuracy on both training and test data. The train loss is consistently lower, approaching 0.0000, indicating nearly perfect performance on the training set. While the test loss is slightly higher but still extremely low (~0.00002 after stabilization), suggesting excellent generalization to unseen data and minimal overfitting. The training loss remains slightly lower than the test loss, but both are extremely minimal, demonstrating excellent generalization and minimal overfitting. Also, the plot of train and test loss for the Bootstrap Aggregation (Bagging) classifier as depicted in Figure 7 shows a sharp decline in both losses as the number of trees increases initially, indicating rapid improvement in model performance. After around 10–20 trees, both the train and test loss values stabilize, with the train loss approaching zero and the test loss remaining slightly higher but still very low. After stabilization (from around 20 trees onward), the train loss is consistently close to 0.00000 to 0.00002, indicating nearly 100% training accuracy. While the test loss stabilizes at approximately 0.00017 to 0.00020, corresponding to a test accuracy between 99.98% and 99.983%. This behavior suggests that the bagging model learns quickly and generalizes well, maintaining high accuracy across both training and unseen test data without overfitting.

CONCLUSION

The study shows that, particularly in settings marked by considerable class imbalance, sophisticated ensemble machine learning techniques that is, Bootstrap Aggregation (Bagging) and Random Forest classifiers, are remarkably successful in detecting credit card fraud. Utilizing careful data preparation, up-sampling minority (fraudulent) cases, and adjusting features, the models obtained remarkable results. With almost perfect recall (100%), shockingly high accuracy (99.98–99.99%), and almost perfect precision (99.97–99.98%), Bagging and Random Forest classifiers excelled. Because of its feature randomization and enhanced generalization capacity, which reduced overfitting and efficient detection of fraud cases without omissions, the Random Forest classifier showed rather better performance. The results confirm that, when combined with appropriate data balancing and assessment methods, ensemble approaches may efficiently distinguish between real and fraudulent transactions, therefore offering a scalable and dependable solution for practical financial systems. Still, challenges in model interpretability, adaptability to shifting fraud tendencies, and explainable artificial intelligence point to possible future routes for research and development in this important area.

REFERENCES

1. Bi, W., Li, L., Zheng, S., Lu, T., & Zhu, Y. (2024). A Dual Ensemble Learning Framework for Real-time Credit Card Transaction Risk Scoring and Anomaly Detection. *Journal of Knowledge Learning and Science Technology* ISSN: 2959-6386 (online), 3(4), 330-339.
2. Chen, Y., Zhao, C., Xu, Y., & Nie, C. (2025). Year-over-Year Developments in Financial Fraud Detection via Deep Learning: A Systematic Literature Review. *arXiv preprint arXiv:2502.00201*. <https://arxiv.org/abs/2502.00201>
3. Hafez, I. Y., Hafez, A. Y., Saleh, A., et al. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, 12(6). <https://doi.org/10.1186/s40537-024-01048-8>
4. Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*, 11(1), 1-22.
5. Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing credit card fraud detection: an ensemble machine learning approach. *Big Data and Cognitive Computing*, 8(1), 6.
6. Marazqah Btoush, E. A. L., Zhou, X., Gururajan, R., Chan, K. C., Genrich, R., & Sankaran, P. (2023). A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ. Computer science*, 9, e1278. <https://doi.org/10.7717/peerj-cs.1278>
7. Ndabula, J. N., Olanrewaju, O. M., & Echobu, F. O. (2023). Detection of Hate Speech Code Mix Involving English and Other Nigerian Languages. *Journal of Information Systems and Informatics*, 5(4), 1416-1431. [doi:10.51519/journalisi.v5i4.595](https://doi.org/10.51519/journalisi.v5i4.595).
8. Ojo, I. P., & Tomy, A. (2025). Explainable AI for credit card fraud detection: Bridging the gap between accuracy and interpretability.

9. Pundkar, S. N., & Zubei, M. (2023). Credit Card Fraud Detection Methods: A Review. *E3S Web of Conferences*, 453, 01015. <https://doi.org/10.1051/e3sconf/202345301015>.
10. Sizan, M. M. H., Chouksey, A., Tannier, N. R., Al Jobaer, M. A., Akter, J., Roy, A., ... & Islam, D. A. (2025). Advanced Machine Learning Approaches for Credit Card Fraud Detection in the USA: A Comprehensive Analysis. *Journal of Ecohumanism*, 4(2), 883-905.
11. Sulaiman, S. S., Nadher, I., & Hameed, S. M. (2024). Credit Card Fraud Detection Challenges and Solutions: A Review. *Iraqi Journal of Science*, 65(4).
12. Suleiman, A. B., Luka, S., & Ibrahim, M. (2023). Cardiovascular disease prediction using random forest machine learning algorithm. *Fudma Journal of Sciences*, 7(6), 282-289.
13. Sultana, I., Maheen, S. M., Kshetri, N., & Zim, M. N. F. (2025). detectGNN: Harnessing Graph Neural Networks for Enhanced Fraud Detection in Credit Card Transactions. arXiv preprint arXiv:2503.22681. <https://arxiv.org/abs/2503.22681>
14. Yu, C., Xu, Y., Cao, J., Zhang, Y., Jin, Y., & Zhu, M. (2024). Credit Card Fraud Detection Using Advanced Transformer Model. arXiv preprint arXiv:2406.03733. <https://arxiv.org/abs/2406.03733>
15. Zhu, M., Zhang, Y., Gong, Y., Xu, C., & Xiang, Y. (2024). Enhancing Credit Card Fraud Detection: A Neural Network and SMOTE Integrated Approach. arXiv preprint arXiv:2405.00026. <https://arxiv.org/abs/2405.00026>