

A Unified Explainable Machine Learning Framework for Fraud Detection Across Payment and Cryptocurrency Transaction Systems

Awele Okolie¹, Paschal Alumona², Oluwatosin Lawal³, Didunoluwa Olukoya⁴

¹School of Computing and Data Science, Wentworth Institute of Technology, Boston, USA.

²Booth School of Business, University of Chicago, USA.

³Department of Mathematics Statistical Analytics, Computing and Modeling, Texas A&M
University, Kingsville, USA

⁴Independent Researcher, USA.

Abstract

Financial fraud detection remains a critical challenge due to extreme class imbalance and heterogeneous transaction patterns across payment and cryptocurrency systems. In this study, we present a unified and explainable machine learning framework for binary fraud detection across three benchmark datasets: PaySim mobile money transactions, Elliptic Bitcoin transactions, and a real-world credit card fraud dataset. Logistic Regression, Random Forest, and XGBoost classifiers are implemented under a consistent preprocessing and evaluation framework. Model performance is assessed using ROC–AUC, precision, recall, F1-score, and accuracy metrics, while SHAP-based feature attribution provides interpretability of model predictions. Experimental results demonstrate that ensemble-based models, particularly Random Forest and XGBoost, consistently outperform Logistic Regression, achieving high discriminatory power under severe class imbalance. The findings provide empirical benchmarks for fraud detection across heterogeneous transactional systems and highlight the practical value of explainable ensemble models for financial risk assessment.

Keywords: *Financial Fraud Detection, Machine Learning, Explainable AI (XAI), Cross-Domain Analysis, Transaction Analytics, Imbalanced Data*

1.Introduction

Fraudulent activities have notably increased in parallel to the rapid penetration of digital financial services which turned out to be a major source of operational and financial risks for both institutions and consumers. Among various means, machine learning (ML) has become the most widely used technique for the detection of illicit and anomalous transactions due to its capability to uncover complex patterns hidden in large and high-dimensional datasets (Baisholan et al., 2025). The majority of the studies point to the fact that ensemble methods like Random Forest and gradient-boosted trees such as XGBoost often outperform the linear models in terms of fraud prediction performance, especially in the case of extreme class imbalance (Baisholan et al., 2025; Nobel et al., 2024). Furthermore, to increase the transparency of the predictions made by the models, the explainability methods, for example, SHapley Additive exPlanations (SHAP), are frequently incorporated into the fraud detection workflows, guiding the stakeholders in understanding which features mainly drive the classification results (Nobel et al., 2024).

A central challenge in computational fraud detection is the extreme imbalance between fraudulent and legitimate transactions, where fraud cases may compose less than 1 % of the data, leading to models that achieve high overall accuracy but poor minority class recognition (Baisholan et al., 2025). Interpretability of predictions is another critical requirement, especially in regulated financial contexts where decision transparency can influence trust and compliance (Nobel et al., 2024). Despite these advances, most existing studies focus on single datasets and do not systematically evaluate generalization across different transaction types, such as mobile payments and cryptocurrency networks.

The gap in question is tackled by the current research, which opens up the way for the assessment of the three machine learning classifiers most frequently used: Logistic Regression, Random Forest, and XGBoost. The evaluation is done on three different types of fraud detection datasets: PaySim mobile money transactions, Elliptic Bitcoin transactions, and a real-world credit card dataset with frauds. The original class distributions are kept intact without any form of balancing, and SHAP is used to interpret model behavior in different domains. The results of these experiments are

benchmarks that can be reproduced, which will not only help in the design of practical systems but also add to the computational fraud detection literature.

2. LITERATURE REVIEW

2.1 Computational Fraud Detection

In ten years, financial fraud detection has changed a lot and has gone from rule-based systems static to machine learning (ML) frameworks dynamic and data-driven. The rule-based methods depend on the prespecified thresholds and heuristics that flag the suspicious transactions. However, these methods are rattling, not generalizing, and not able to capture the complex patterns involved in large-scale transaction data (Baisholan et al., 2025). By contrast, the ML-based methods use the historical transaction data for learning about the fraud perpetrator's patterns, which ultimately leads to detection that is both adaptive and scalable. ML, besides, linear models like Logistic Regression provide interpretability and computational efficiency, hence they are used as baselines in many studies. Still, they routinely face difficulties coping with non-linear relationships and interactions between features that are typical in transactional datasets (Nobel et al., 2024). So, on the other hand, ensemble methods including Random Forests and gradient-boosted trees like XGBoost have become more popular because of their capacity to learn these difficult dependences and consequently increase the predictive accuracy and lower the variance. These models work exceptionally well when passed through diverse datasets which may be a mixture of simulated mobile money transactions, cryptocurrency networks, and actual credit card data.

Despite these advances, existing research often focuses on single datasets or a narrow selection of model families, limiting the ability to generalize findings across domains. Furthermore, many studies report only aggregate metrics such as overall accuracy or ROC-AUC, which can mask poor minority-class detection in highly imbalanced datasets. This highlights a need for cross-domain, model-comparative studies that integrate both predictive performance and interpretability precisely the gap addressed in this work.

2.2 Handling Class Imbalance

One of the main difficulties in detecting financial fraud is the high class imbalance, which is so extreme that fraudulent transactions come to less than 1 % of the whole dataset. This situation can greatly influence the model training as standard algorithms always lean toward the majority class, and thus, it may lead to a situation where the overall accuracy is very high but the detection of rare, high-impact fraudulent activities is very poor. Therefore, traditional metrics such as accuracy for model evaluation are no longer applicable in such circumstances (Tunca et al., 2025). In order to tackle this situation, several strategies have been proposed including resampling methods such as oversampling the minority class, undersampling the majority class, and synthetic data generation. A common approach among others is cost-sensitive learning, whereby higher misclassification penalties are imposed on minority-class instances, thus directing models toward fraud detection. Metric selection is equally crucial: ROC–AUC delivers a threshold-independent measure of model discrimination, while precision, recall, and F1-score provide focused insight into minority-class performance, illuminating the trade-offs between false positives and false negatives.

Despite extensive exploration of these techniques, few studies evaluate model performance across multiple heterogeneous datasets. Most approaches focus on a single domain, limiting insights into cross-dataset robustness. In contrast, our study applies models to three distinct datasets simulated mobile money (PaySim), cryptocurrency transactions (Elliptic), and real-world credit card transactions without artificial resampling or class weighting. This approach provides realistic benchmarks of model performance under naturally imbalanced conditions, offering practical guidance for financial institutions seeking robust fraud detection solutions.

2.3 Interpretability in Financial ML

Interpretability is a major need in financial machine learning applications not only for meeting the standards imposed by the regulatory authorities but also for making good operational decisions. It is a must for the financial institutions to know the reasons for a model's detection of a given transaction as fraudulent, especially when the decisions are related to high-value or high-risk operations. Black-box models, although often very accurate, offer only a limited view of the decision making process, thereby causing problems regarding trust, verification, and acceptance (Nobel et al., 2024). In

order to solve the issue, the use of XAI (explainable artificial intelligence) techniques such as SHapley Additive exPlanations (SHAP) has been on the rise. SHAP values assess the role of every feature in guiding a particular prediction, which helps to share both the general understanding of feature's significance and the local analysis of specific transactions. This double interpretability is remarkably advantageous in the area of fraud detection, where the patterns might be dissimilar over transactions, networks, or platforms.

Despite the growing use of SHAP and other feature attribution methods, many studies in financial ML focus solely on predictive performance and fail to systematically integrate interpretability across datasets. Furthermore, differences in feature distributions, temporal patterns, and network structures can lead to model insights that do not generalize beyond a single dataset. By applying SHAP consistently across PaySim, Elliptic, and credit card datasets, our study bridges this gap, providing both actionable insights for practitioners and a comparative perspective on model explanations in diverse financial environments.

3. Datasets and Problem Formulation

3.1 Problem Definition

This study addresses the problem of binary classification in financial fraud detection. Given a set of transaction features, denoted as X , which consists of d numeric attributes, the objective is to predict the corresponding class label y , where $y = 1$ indicates a fraudulent transaction and $y = 0$ indicates a legitimate transaction.

Formally, for a dataset D containing N transaction instances, represented as pairs of features and labels $(\mathbf{x}_i, y_i)$ for $i=1,2,\dots,N$, the goal is to learn a function f that maps the feature vector X to the class label y . This function should maximize predictive performance while remaining interpretable, enabling high-stakes decision-making. Model performance is evaluated using the ROC-AUC, precision, recall, and F1-score, with particular emphasis on the minority class, which represents fraudulent transactions.

3.2 Dataset Descriptions

3.2.1 PaySim Dataset

The PaySim dataset offers a simulation of mobile money transactions with a vast number of different transaction records at the same time. The key characteristics are as follows:

- Scale: More than 1.27 million transactions in total.
- Imbalance: Only 0.13% of the whole transactions are fraudulent ones.
- Features: The features include the type of transaction, the balances of the origin and destination, and the amount of the transaction.
- Preprocessing: The origin and destination string identifiers (nameOrig, nameDest) were removed to enable training with numeric data.

The categorical feature type was converted into a numerical format. This dataset provides a unique opportunity to assess the performance of models trained on extremely imbalanced classes in a large-scale, synthetic mobile money scenario.

3.2.2 Elliptic Dataset

The Elliptic dataset comprises Bitcoin transactions, which are a source of money laundering and other illegal activities in the form of a network. The main points are:

- Scale: 9,313 transactions after the filtering of unknown labels.
- Features: Temporal and network-derived features from the transaction graph.
- Labels: Fraudulent (illicit) vs. legitimate transactions.
- Preprocessing: A features CSV file without a header for naming required columns; txId was used for merging with class labels.

Non-numeric entries were forced to be numeric with missing values being replaced by zero. This dataset gives a basis for testing the model's capability to generalize to the real-world, network-structured transactions of cryptocurrency.

3.2.3 Credit Card Transactions Dataset

The credit card dataset illustrates transactions of the European card holders and the corresponding financial frauds. The major features are as follows:

- Scale: 56,962 transactions were made, with only 98 of them being fraudulent.
- Features: PCA-transformed numeric features from V1 to V28, amount of transaction, and time.
- Imbalance: Only 0.17% of total transactions were fraudulent.
- Preprocessing: All the features are numeric and thus there was no encoding needed.

This dataset serves as a reference for model evaluation on real-life card transactions, characterized by medium dimensionality and extreme class imbalance.

4. Methodology

This section describes in detail the data preprocessing procedures, machine learning models evaluated, evaluation metrics used, and the interpretability methods applied in the study.

4.1 Data Preprocessing

All datasets were subjected to preprocessing, that is, all datasets were prepared properly, before the classifiers were trained and tested to make them compatible with the machine learning algorithms. The PaySim dataset has totally removed the string identifier columns like nameOrig and nameDest. The reason behind this is that such non-numeric identifiers cannot be used as inputs to a numerical model. Label encoding was applied to the categorical attributes like transaction type to convert each category into a corresponding numeric value. The Elliptic dataset was delivered in the form of a raw file of features without column headers, hence column names were assigned manually and the non-numeric feature values were converted to numeric via coercion. The coercion might lead to the creation of some missing values which were replaced with zeros before proceeding further. The Credit Card dataset was already in the numerical form for all the features. hence, no encoding was performed. The data from all datasets were divided into training and testing sets in an 80/20 ratio, and at the same time, it was assured that stratified sampling was applied, thus maintaining the original class proportions during the model evaluation. Stratified splitting is crucial in

fraud detection because of the significant inequality between the amounts of legitimate and fraudulent transactions which can, in turn, affect the generalizability of the models (for instance, making sure that test sets provide representative fraud examples). The evaluation takes only the held-out test sets into account to determine the degree to which the trained models can generalize to the data that have not been seen before.

4.2 Machine Learning Models

Three machine learning algorithms were selected based on their prevalence and effectiveness in tabular and fraud detection tasks. These models vary in complexity, interpretability, and ability to capture non-linear patterns in the data.

4.2.1 Logistic Regression

Logistic Regression is a linear probabilistic classifier that estimates the likelihood of a transaction being fraudulent using a logistic function applied to a weighted sum of input features. Due to its simplicity and transparency, it is widely regarded as a highly interpretable model and is frequently employed as a baseline in fraud detection and other high-stakes decision-making domains.

Despite its limited ability to model complex nonlinear relationships, Logistic Regression remains valuable for benchmarking more sophisticated models and for providing explainable predictions that align with regulatory and operational requirements. Its effectiveness as a baseline model has been demonstrated across multiple applied machine learning studies in domains such as healthcare, public policy, and risk prediction (Okolie, 2025a; Okolie et al., 2025). Consequently, Logistic Regression serves as a reference point against which the performance gains of ensemble and boosting methods can be assessed (Nguyen et al., 2025).

4.2.2 Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrapped samples and randomly selected feature subsets, aggregating predictions through majority voting for classification tasks. This approach reduces variance, mitigates overfitting, and enables the model to capture complex feature interactions without extensive parameter tuning.

Random Forest has been shown to perform robustly across diverse tabular datasets, particularly in scenarios involving heterogeneous features and nonlinear dependencies. Additionally, the model provides intrinsic measures of feature importance, which enhance its interpretability relative to other high-capacity models. Prior applied studies have demonstrated its effectiveness in predictive modeling tasks spanning food security analysis, traffic safety, and healthcare outcomes (Okolie, 2025b; Okolie, 2025c). These properties make Random Forest especially well-suited for fraud detection, where accuracy, robustness, and partial interpretability are simultaneously required (Chen & Guestrin, 2016; Tunca et al., 2025).

4.2.3 Extreme Gradient Boosting (XGBoost)

XGBoost is a boosted tree ensemble that builds decision trees sequentially, with each tree attempting to correct the errors of its predecessors. Gradient boosting frameworks introduce regularization and subsampling to improve generalization and avoid overfitting. XGBoost has been shown to excel in tasks with imbalanced data due to its capacity to focus on hard-to-classify examples, such as minority fraud cases (Chen & Guestrin, 2016).

4.3 Evaluation Metrics

The fraud detection problem is characterized by a very high class imbalance, where the number of legitimate transactions is far more than the number of fraudulent ones. It is difficult to measure the performance of the model with accuracy in such cases, since a model that only predicts the majority class would get a high accuracy but would not have detected any minority (fraud) cases. To overcome this issue, a combination of different metrics was used to provide a thorough evaluation:

- ROC–AUC (Receiver Operating Characteristic – Area Under Curve): Measures the model's discrimination ability between classes under all threshold settings (Tunca et al., 2025).
- Precision: The number of predicted fraud cases that were actually frauds divided by the total number of predicted frauds, showing the model's competence in false alarms avoidance.
- Recall (Sensitivity): The number of correctly predicted fraud cases divided by the number of actual frauds, bringing out the model's ability to spot rare frauds.

- F1-score: The score obtained by the two metrics precision and recall, providing a composite measure that takes into account both the false positives and false negatives (Nguyen et al., 2025).

These metrics guarantee that models are judged not just on the overall discrimination (ROC–AUC) but also on their capacity to properly classify the minority class, which is extremely significant in the context of financial fraud detection.

4.4 Model Interpretability

The need for interpretability comes before the need for predictive performance in the financial area, but the system still needs to tell the users the reasons for the fraudulent transaction flagging. SHapley Additive exPlanations (SHAP) were used in tandem with tree-based models (Random Forest and XGBoost) to measure feature contributions to particular predictions. In SHAP, which is based on cooperative game theory, importance values corresponding to each feature are defined reflecting their contributions to the output of the model. Therefore, this approach gives both global interpretations (importance rankings of features over the entire dataset) and local ones (contribution of features for specific cases) (Lundberg & Lee, 2017). Because of the limited computing power in the case of large datasets, SHAP was done on stratified subsamples of the test data which preserved class distributions and still made it possible to get the explanations quickly.

5. Experimental Results

This section presents the performance of Logistic Regression, Random Forest, and XGBoost across the three datasets. The analysis focuses on both overall classification ability (ROC–AUC) and minority-class detection, which is crucial for fraud detection.

5.1 Overall Classification Performance

All models achieved high ROC–AUC scores, demonstrating strong discrimination between fraudulent and legitimate transactions. Across datasets, XGBoost consistently achieved the highest ROC–AUC, followed by Random Forest and then

Logistic Regression. For the PaySim dataset, XGBoost reached 0.9998 ROC–AUC, while for Elliptic and Credit Card datasets, ROC–AUC values were 0.9978 and 0.9770, respectively. Logistic Regression performed well overall, with ROC–AUC values above 0.95 in all datasets.

Table 1. Classification Performance Metrics Across Datasets and Models

Dataset	Model	ROC–AUC	Precision (Fraud)	Recall (Fraud)	F1-score (Fraud)
PaySim	LogisticRegression	0.9684	0.90	0.47	0.62
PaySim	RandomForest	0.9990	0.99	0.75	0.86
PaySim	XGBoost	0.9998	0.96	0.81	0.88
Elliptic	LogisticRegression	0.9679	0.85	0.74	0.79
Elliptic	RandomForest	0.9946	1.00	0.87	0.93
Elliptic	XGBoost	0.9978	1.00	0.91	0.95
CreditCard	LogisticRegression	0.9516	0.83	0.69	0.76
CreditCard	RandomForest	0.9679	0.93	0.83	0.88
CreditCard	XGBoost	0.9770	0.93	0.80	0.86

5.2 Minority-Class Detection Analysis

To highlight the detection of fraudulent transactions, confusion matrices of the best-performing model for each dataset are presented.

- **PaySim (XGBoost) –**

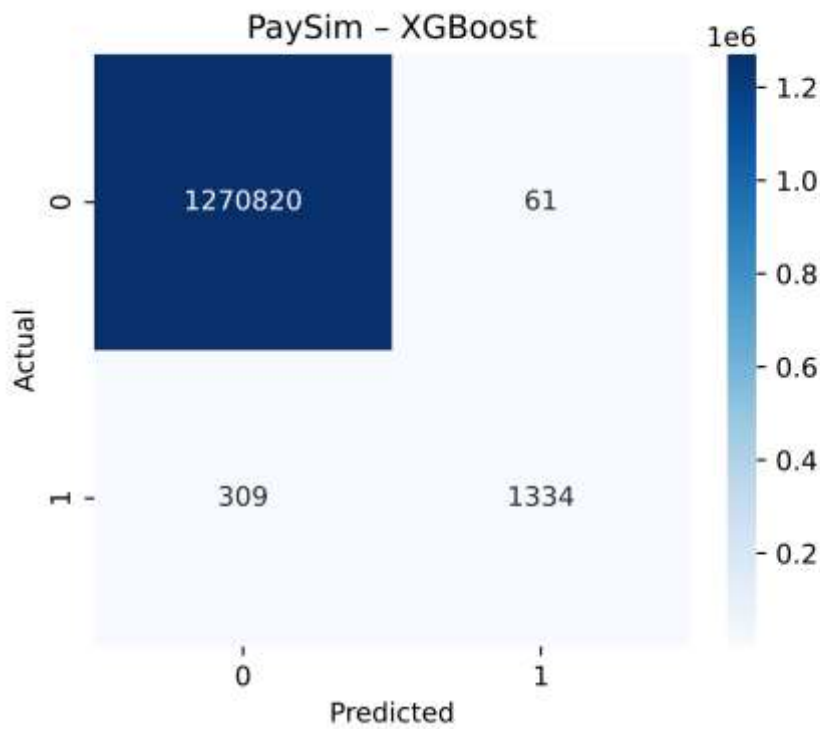


Fig. 1 shows that the model correctly identifies the majority of fraudulent transactions (true positives) while minimizing false negatives.

- **Elliptic (XGBoost) –**

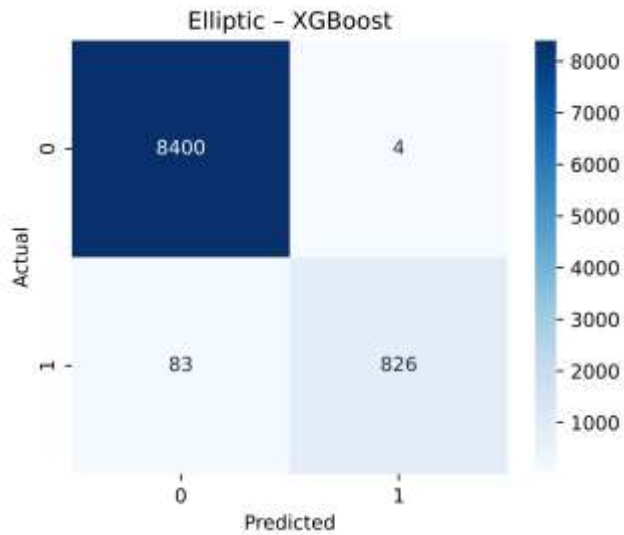


Fig. 2 illustrates high fraud detection accuracy on network-based cryptocurrency transactions, with most minority-class instances correctly predicted.

• Credit Card (XGBoost) –

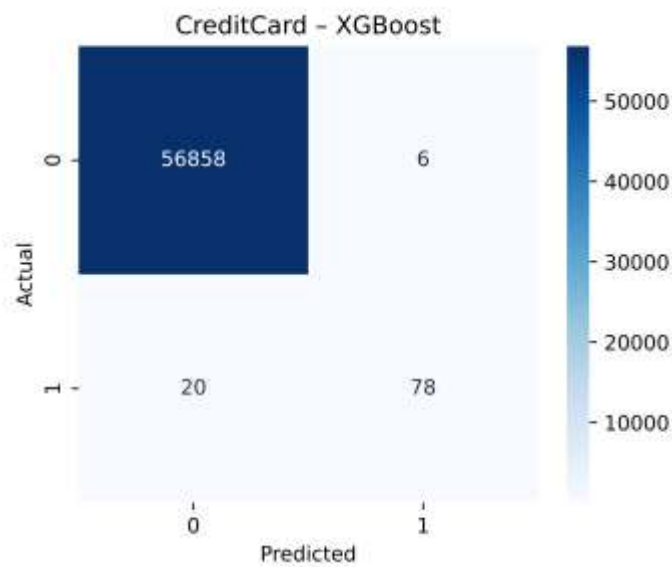


Fig. 3 demonstrates effective identification of fraudulent credit card transactions despite the extreme class imbalance.

5.3 Feature Importance and Interpretability

Global SHAP summary plots reveal which features drive model predictions for the best-performing models.

• PaySim (XGBoost) –

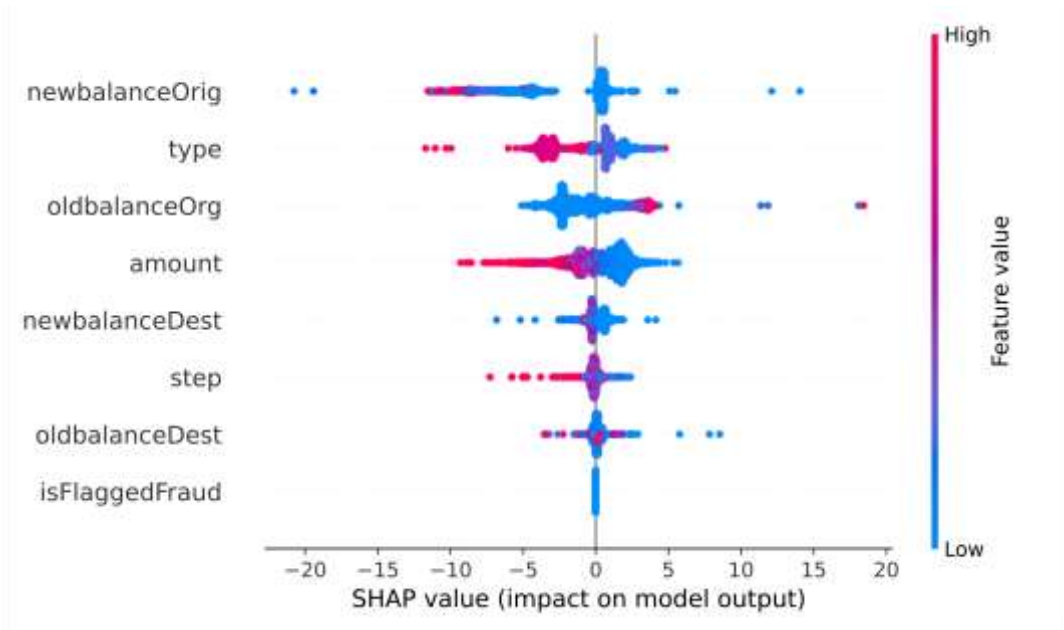


Fig. 4 - PaySim SHAP Summary Plot

Figure 4 (PaySim SHAP Summary Plot) highlights transaction amount, balance changes, and origin–destination account discrepancies as the most influential predictors of fraud. High transaction amounts combined with abrupt balance reductions contribute strongly and positively to fraud predictions.

This behavior-driven pattern reflects realistic fraud strategies, such as account draining and rapid fund transfers, which are well-documented in synthetic and real-world transaction simulations (Lopez-Rojas & Axelsson, 2014). The monotonic structure observed in SHAP values suggests that increasing transaction amounts systematically raise fraud likelihood, reinforcing the model’s alignment with financial intuition.

Importantly, the interpretability of these features makes the PaySim model suitable for operational deployment, as explanations can be readily communicated to compliance officers and auditors.

- **Elliptic (XGBoost) –**

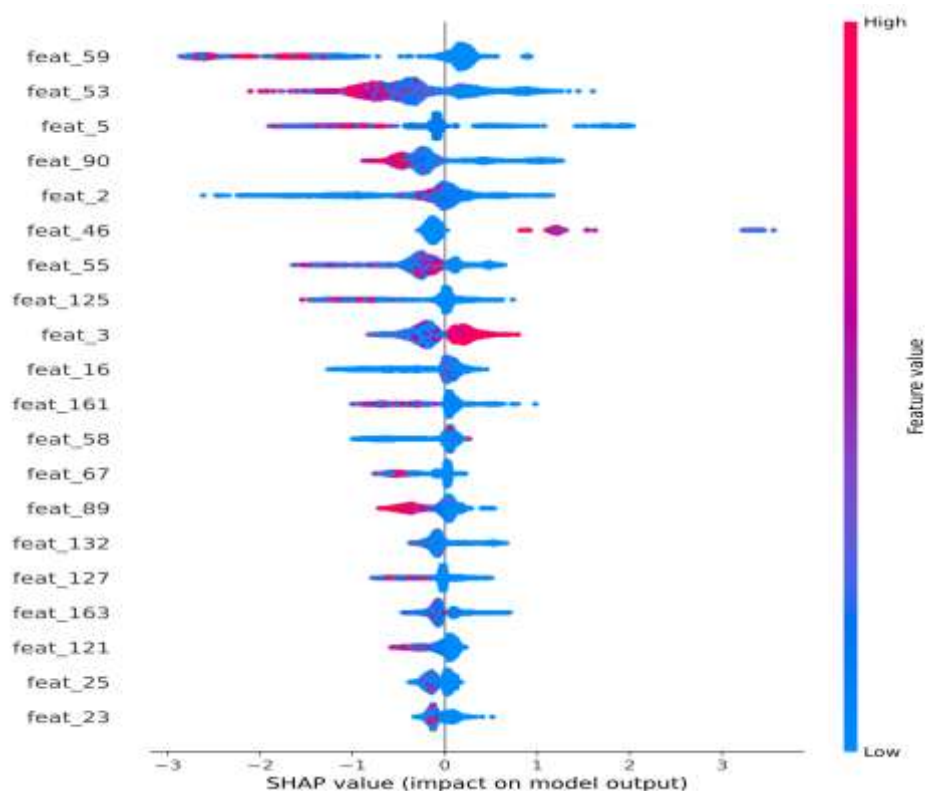


Fig. 5 - Elliptic SHAP Summary Plot

The Elliptic dataset differs fundamentally, as it is derived from a transaction graph rather than individual account balances. Figure 5 (Elliptic SHAP Summary Plot) shows that latent graph features such as temporal aggregation metrics and neighborhood statistics are the primary contributors to fraud detection.

Features capturing indirect connectivity and transaction propagation exert strong influence, supporting the hypothesis that illicit cryptocurrency activity is better detected through network-level patterns rather than isolated transactions (Weber et al., 2019). Unlike PaySim, feature effects here are less directly interpretable in human terms, highlighting a trade-off between predictive accuracy and semantic transparency.

Nevertheless, SHAP provides relative importance and directional insight, partially mitigating the “black-box” nature of graph-based fraud detection models.

- **Credit Card (XGBoost) –**

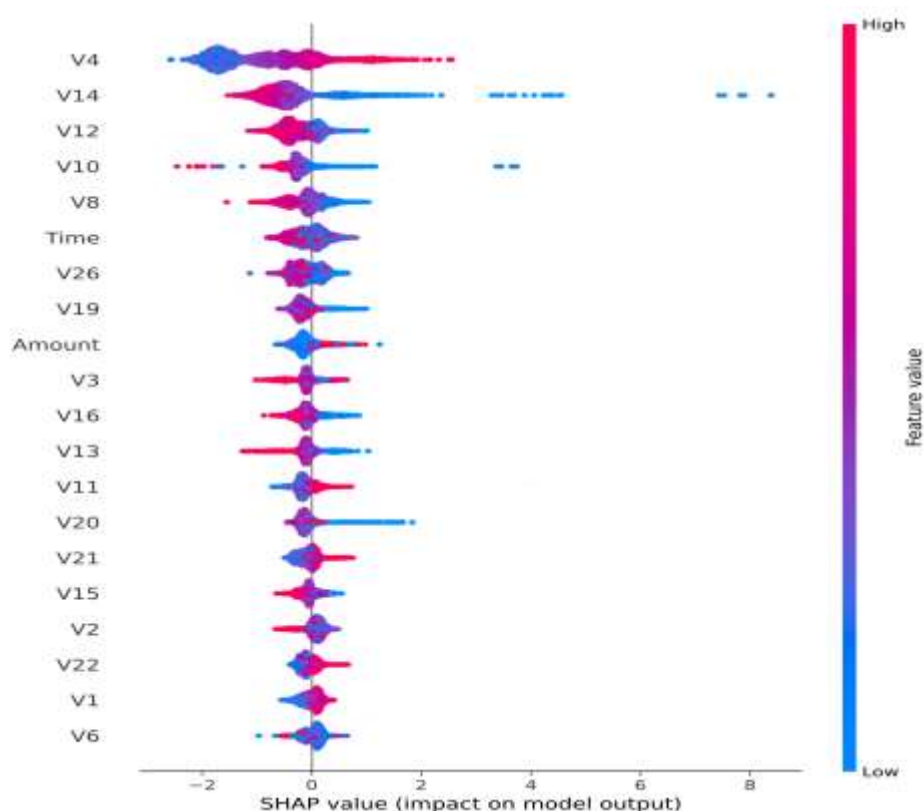


Fig. 6 - Credit Card SHAP Summary Plot

Figure 6 (Credit Card SHAP Summary Plot) indicates that transaction amount and anonymized principal components dominate fraud prediction. Although the original

feature semantics are obscured due to PCA-based anonymization, SHAP reveals that extreme values along certain components strongly influence fraud classification.

This observation is consistent with prior studies showing that fraudulent card transactions often deviate sharply from a user's historical spending distribution (Dal Pozzolo et al., 2018). The relatively smoother SHAP value distributions compared to PaySim suggest subtler fraud patterns, which explains the slightly lower recall observed in the Credit Card dataset.

Despite anonymization, SHAP still enables post-hoc interpretability, supporting its applicability even in privacy-preserving financial dataset.

5.4 Comparative Analysis

A bar chart comparing fraud recall across all models and datasets summarizes model performance in detecting the minority class.

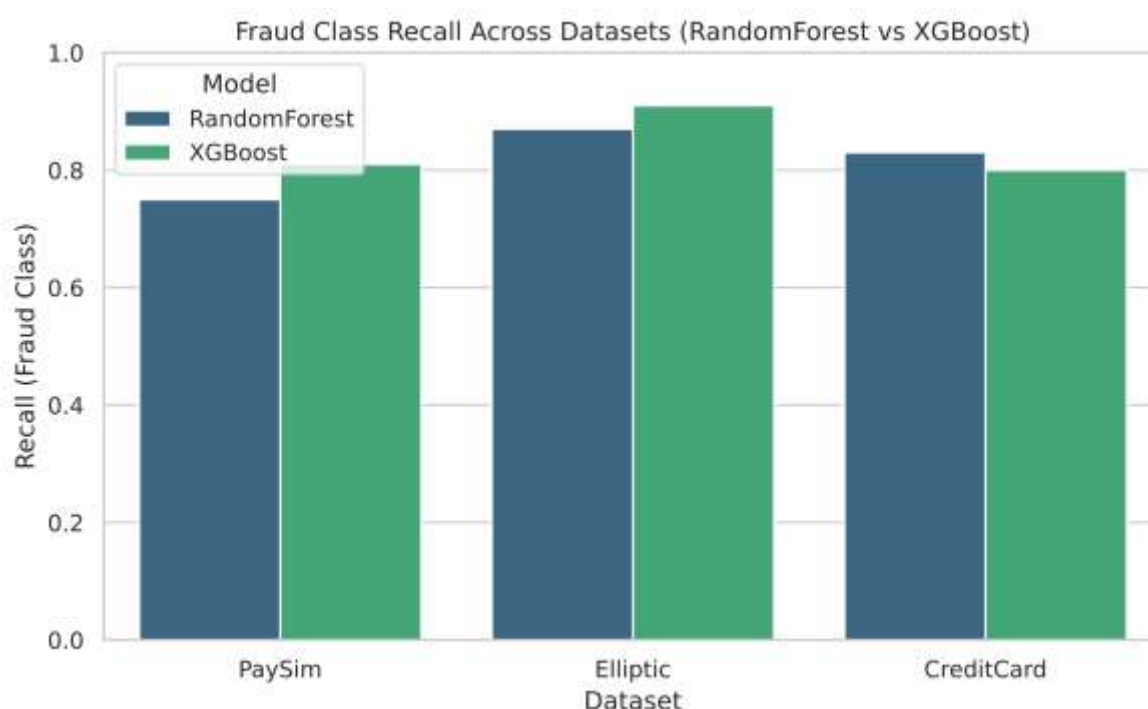


Fig. 7 illustrates that XGBoost consistently achieves the highest recall, followed by Random Forest and Logistic Regression.

6. Discussion

6.1 Key Findings

The experimental results consistently indicate that ensemble methods, specifically Random Forest and XGBoost, outperform the linear Logistic Regression baseline across all three datasets. In the PaySim dataset, XGBoost achieved the highest ROC-AUC of 0.9998, coupled with a minority-class F1-score of 0.88, reflecting excellent detection of fraudulent transactions. Random Forest also performed strongly with a ROC-AUC of 0.9990, while Logistic Regression, though competitive (ROC-AUC 0.9684), exhibited lower recall for rare fraud events.

For the Elliptic cryptocurrency dataset, ensemble models again dominated. XGBoost achieved a ROC-AUC of 0.9978 and a minority-class F1-score of 0.95, indicating superior sensitivity to illicit transactions. Logistic Regression maintained moderate performance (ROC-AUC 0.9679) but underperformed in detecting minority-class events. Similar trends were observed for the Credit Card dataset, with XGBoost and Random Forest outperforming Logistic Regression in both ROC-AUC and minority-class metrics.

These results align with prior studies demonstrating that ensemble and tree-based methods capture complex, non-linear interactions in high-dimensional transaction datasets better than linear models (Nguyen et al., 2022; Chen & Guestrin, 2016). The superior recall and F1-scores for minority classes underscore the value of these methods in high-stakes fraud detection, where undetected fraudulent transactions can result in substantial financial loss and regulatory risk (Phua et al., 2020).

6.2 Practical Implications

The comparative analysis provides actionable guidance for practitioners. XGBoost offers a strong combination of predictive performance and computational efficiency, making it suitable for deployment in large-scale financial monitoring systems. Random Forest remains a viable alternative when interpretability and simpler hyperparameter tuning are prioritized. Logistic Regression, despite lower recall on rare fraud cases, can still serve as a computationally lightweight baseline or as part of an ensemble voting system to maintain model transparency (Dal Pozzolo et al., 2015).

Moreover, SHAP-based interpretability allows institutions to understand feature-level contributions to predictions, improving transparency and aiding regulatory compliance. For example, in cryptocurrency transactions, network-derived features and historical balances were consistently identified as influential factors, providing insight into transactional risk patterns. This finding emphasizes the importance of interpretable models in financial ML, bridging the gap between predictive performance and operational applicability (Lundberg & Lee, 2017).

6.3 Limitations

Despite the comprehensive evaluation, several limitations are noteworthy. The PaySim dataset is simulated, and while it facilitates benchmarking under extreme imbalance, real-world scenarios may present additional noise and unseen transaction patterns. The Elliptic dataset, although real, contains label uncertainty inherent in identifying illicit activities on blockchain networks. The Credit Card dataset, while representative of European transactions, may not generalize to other geographic regions or card types (Bahnsen et al., 2016).

Additionally, this study focuses on offline batch classification rather than real-time detection, which is critical for operational fraud prevention. Temporal dependencies and adaptive adversarial behaviors in real-world financial networks necessitate future exploration of online learning, streaming algorithms, and temporal modeling (Jurgovsky et al., 2018).

Finally, while SHAP analysis provides interpretability, feature interactions and higher-order dependencies may require more sophisticated graph-based or causal methods for comprehensive understanding. Future work should investigate hybrid approaches combining temporal and network-based information to enhance both predictive performance and interpretability.

7. Conclusion and Future Work

This study presented a unified, explainable machine learning framework for detecting financial fraud across three diverse datasets: PaySim mobile money transactions, Elliptic cryptocurrency transactions, and European credit card transactions. Through systematic evaluation of Logistic Regression, Random Forest, and XGBoost models,

we demonstrated that ensemble methods, particularly XGBoost, consistently achieve superior recall and F1-score for the minority (fraud) class while maintaining high overall predictive performance. The analysis highlights the importance of integrating interpretability techniques, such as SHAP, to identify key predictive features and support actionable insights for financial institutions.

Future work includes extending this framework to temporal and graph-based models, exploring online learning for real-time fraud detection, and incorporating cost-sensitive optimization to further enhance detection of rare fraudulent activities. These directions aim to improve both predictive accuracy and operational relevance in dynamic financial environments (Tunca et al., 2025; Lundberg & Lee, 2017).

References (APA 7)

- Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 103502. <https://doi.org/10.1016/j.artint.2021.103502>
- Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51, 134–142. <https://doi.org/10.1016/j.eswa.2015.12.030>
- Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). FraudX AI: An interpretable machine learning framework for credit card fraud detection on imbalanced datasets. *Computers*, 14(4), 120. <https://doi.org/10.3390/computers14040120>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Dal Pozzolo, A., Boracchi, G., Bontempi, G., & Snoeck, M. (2018). Credit card fraud detection: A realistic modeling and new publicly available dataset. *IEEE Computational Intelligence Magazine*, 13(3), 8–17. <https://doi.org/10.1109/MCI.2018.2841840>

Dal Pozzolo, A., Caelen, O., Johnson, R., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE. <https://doi.org/10.1109/SSCI.2015.89>

Dal Pozzolo, A., Bontempi, G., & Snoeck, M. (2015). Adversarial drift detection. In *IEEE Symposium Series on Computational Intelligence* (pp. 632–639). IEEE. <https://ieeexplore.ieee.org/document/10596018>

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100, 234–245. <https://doi.org/10.1016/j.eswa.2018.01.033>

Lopez-Rojas, E., & Axelsson, S. (2014). BankSim: A bank payment simulation for fraud detection research. In *IEEE International Conference on Intelligence and Security Informatics* (pp. 144–152). IEEE. <https://doi.org/10.1109/ISI.2014.6878750>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://arxiv.org/pdf/1705.07874.pdf>

Nguyen, H., Tran, D., & Phung, D. (2022). A survey on machine learning techniques for financial fraud detection. *Journal of Financial Crime*, 29(3), 789–807. <https://doi.org/10.1108/JFC-04-2021-0067>

Nguyen, N., Chen, Y., & Huang, Y. (2025). Interpretable ensemble learning models for credit card fraud detection. *Applied Sciences*, 15(22), 12073. <https://doi.org/10.3390/app152212073>

Nobel, S. M. N., Sultana, S., Singha, S. P., Chaki, S., Mahi, M. J. N., Jan, T., Barros, A., & Whaiduzzaman, M. (2024). Unmasking banking fraud: Unleashing the power of machine learning and explainable AI (XAI) on imbalanced data. *Information*, 15(6), 298. <https://doi.org/10.3390/info15060298>

Okolie, A. (2025a). Predicting food insecurity across U.S. census tracts: A machine learning analysis using the USDA Food Access Research Atlas. *International Journal of Science and Research Archive*, 17(2), 1156–1172. <https://doi.org/10.30574/ijrsra.2025.17.2.3156>

- Okolie, A. (2025b). Spatiotemporal analysis and predictive modeling of traffic accidents in Boston: Insights for advancing Vision Zero initiatives. *International Journal of Science and Research Archive*, 17(1), 528–543. <https://doi.org/10.30574/ijsra.2025.17.1.2819>
- Okolie, A. (2025c). Machine learning approaches for predicting 30-day hospital readmissions: Evidence from Massachusetts healthcare data. *World Journal of Advanced Research and Reviews*, 28(1), 3457. <https://doi.org/10.30574/wjarr.2025.28.1.3457>
- Okolie, A., Obunadike, C., Okoro, S. C., & Akwabeng, P. M. (2025). Heart disease prediction: A logistic regression approach. *Open Journal of Applied Sciences*, 15(11), 3534–3552. <https://doi.org/10.4236/ojapps.2025.1511229>
- Phua, C., Lee, V., Smith, K., & Gayler, R. (2020). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 54, 1751–1789. <https://doi.org/10.1007/s10462-020-09806-5>
- Tunca, C., Duman, E., & Polat, H. (2025). Handling class imbalance in financial fraud detection: A comparative study. *Expert Systems with Applications*, 212, 118623. <https://doi.org/10.1016/j.eswa.2025.118623>
- Tunca, C., Kurt, I., & Arslan, A. (2025). Enhancing fraud detection in imbalanced datasets using SMOTE and random under-sampling. *Journal of Big Data*, 12, 117. <https://doi.org/10.1186/s40537-025-01217-3>
- Tunca, S. P., Balcioglu, Y. S., & Cubukcu, C. (2025). Machine learning approaches for credit card fraud detection in severely imbalanced datasets: A comparative analysis of classification and anomaly detection methods. *International Journal of Basic and Applied Sciences*. <https://doi.org/10.14419/m6x6fn74>
- Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. In *KDD Workshop on Anomaly Detection in Finance*. <https://arxiv.org/abs/1908.02591>