International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 663-689

https://doi.org/10.70917/ijcisim-2025-0041
Article

Trustworthiness Validations of Machine Learning Models for Fraud Detection in Mobile Based Business Transactions Using XAI

D. Jeya Mala *, Dev Krishna Jhawar and Chaitanya Bhude

Vellore Institute of Technology, Chennai 600127, Tamil Nadu, India
* Corresponding author: jeyamala.d@vit.ac.in

Received date: 19 November 2024; Accepted date: 23 November 2025; Published online: 31 December 2025

Abstract: The increasing usage of mobile-based transactions makes the business operations more in need of strong fraud detection systems. However, due to the black-box nature of several of the advanced machine learning models, trust in such automated fraud detection has been limited because stakeholders lack insight into how the model provided the decisions. This paper discusses the trustworthiness validation of Machine Learning (ML) models such as Random Forest (RF) and MLP Classifier (Multilayer Perceptron Model) and XGBoost (Extreme Gradient Boosting) applied for the detection of fraudulent transactions within the mobile-based business transactions by using Explainable Artificial Intelligence (XAI) techniques. Some of the most widely used XAI models applied in this work are LIME, SHAP, Anchor Explanations, Surrogate Models, and Explainable Boosting Machine (EBM). These XAI models assess the trustworthiness of the ML models using the most crucial metric such as fidelity and the applied XAI models are validated using unambiguity and interpretability size. The implication of the findings is a framework through which the reliability of models enhancement is done in order to work with increased confidence in ML-driven fraud detection systems within the mobile payment sector.

Keywords: fraudulent transactions detection; mobile payments; digital banking; machine learning; explainable AI; transparency; interpretability; adversarial examples

1. Introduction

Payment fraud refers to using somebody else's payment details in carrying out unauthorized purchases or transactions. Normally, one may detect such unauthorized practices and bring a dispute complaint, as one who is suffering victimization from the activities is the legitimate card owner or account owner. These people can steal someone else's credit card and use it, draw false checks, or con-duct unauthorized electronic fund transfers. All these forms of fraudulent trans-actions waste billions of dollars in businesses and erode customer trust.

Machine learning models are actually the most critical component in payment fraud detection. It provides tremendous tools for the real-time identification and blocking of suspicious transactions. Beyond their ability to predict suspicious transactions, their success will also be founded on the credibility and interpretability they provide along with their capacity to generalize over different fraud patterns. Without proper validation of the ML models, they might result in misleading or biased outcomes that classify fraudulent activities or worse still flag some legitimate transactions incorrectly. It is, therefore, critical to validate these models for trustworthiness so they can function correctly and reliably in high-stakes fraud detection applications, especially where user interactions in mobile business transactions are rapid and sensitive to delays.

To this end, Explainable AI (XAI) frameworks have emerged as invaluable tools. XAI models form

the bridge between the complex world of ML and that of human interpretability by providing stakeholders with the understanding of how the logic of a machine learning prediction is realized. Using XAI techniques such as SHAP, LIME, or Anchors, researchers can investigate the machine learning or the deep learning model behavior, unearth biases, and ensure adherence to business and ethical expectations of predictions. It also increases the level of trust and improves the resilience of fraud detection systems as ML models validated by XAI offer transparency critical for such high-risk sectors as financial transactions.

2. Background

2.1. Traditional Methods of Detecting Fraudulent Payments

Historically, fraud detection systems have been very rule based. Such systems work by a set of predefined rules and criteria established by experts with known fraud patterns from historical data. For example, a rule may detect transactions over a set amount within a short period or those coming from high-risk regions. Rule-based systems are simple, simple to implement and economically inexpensive, which makes them a favorite among many businesses.

The primary advantage of rule-based systems lies in their simplicity. They are inexpensive to design, implement, and understand; it can be relatively straightforward for an organization to develop simple mechanisms for detecting fraud without much cost.

However, though these benefits give rule-based systems an upper hand over more conventional methods, these still have some downsides that increase its ineffectiveness as the fraudulent methods always take ever-changing manipulative properties. Among its worst disadvantages is that it is a static system; once rules are established, those do not get updated to recognize emerging patterns of fraud and evolving ones unless the rules are updated manually. It is a kind of cumbersome process, as it needs updating lags behind the emerging patterns. In addition, these rules are general in nature and specifically intended to cover as much as possible of fraudulent attempts, which subsequently is reflected in a large number of false positives-legitimate transactions flagged as fraudulent. This causes unwanted interruptions and dissatisfaction among customers in this context. This calls for more adaptive and dynamic solutions in the detection of frauds.

2.2. Supervised Learning

The most fundamental approach in machine learning is supervised learning, where an algorithm is trained on a labeled dataset, so every input data point is paired with the corresponding correct output. The goal is that the algorithm learns the relationship between inputs and outputs so well that it could accurately predict outcomes for new data. This method has several advantages: it achieves very high predictions, easy performance metrics such as accuracy and recall can be used, and flexibility across different types of problems, including classification and regression. However, there are also some disadvantages of supervised learning. It heavily relies on the availability of labelling extensive datasets, which can be very costly as well as time-consuming. In addition, there is overfitting, where the model will have high accuracy on the training data but poorer performance on new data unless properly regulated. Furthermore, successful supervised learning models might not adapt to new or unseen patterns, especially if it does not occur during the training of the models, and thus implies declining effectiveness with time.

2.3. Unsupervised Learning

Unsupervised learning is one of the types of machine learning approaches that work on unlabeled data to attempt to find the natural structure and patterns in a dataset. It differs from the supervised learning where it utilizes known outputs for training models, seeking only to identify patterns and relationships directly from the input data. This technique comprises multiple methods such as clustering algorithms and anomaly detection, as well as finding new patterns with respect to fraud-commits; particularly the most critical applications are in the field of financial fraud prevention.

For instance, clustering algorithms cluster similar data points according to their features without any predefined group definitions. These definitions are made so that the inherent structures in the data can be revealed. Other methods include K-means clustering, where the data is partitioned into 'K' clusters by assigning each data point to the cluster with the nearest mean, aiming to minimize variance within each cluster. Hierarchical clustering forms a tree of clusters by an agglomerative or divisive strategy where one starts with individual objects and merges them into bigger clusters. Density-based clustering, on its part, identifies clusters as areas of higher density by comparing the relative density of data points in a particular area; thus, it is able to discover clusters in an arbitrary shape and is robust in the presence of

noise.

Anomaly detection emphasizes on identifying unusual items or occurrences that are substantially different from the vast majority of the data. Most application of anomaly detection is within fraud spots, network security improvement, and quality control. Since unsupervised learning methods do not depend on pre-labeled information, at which they excel in discovering hidden relationships and patterns existing in data, they can identify unsuspected fraud patterns that may not have been achievable by other conventional methods.

3. Literature Survey

Machine learning models have gradually become part of financial fraud detection. These are advanced mechanisms used for real-time fraud detection. In this context, Choi and Lee conducted an in-depth survey of AI approaches in the sphere of IoTs, explaining in detail the challenges and opportunities toward the implementation of such technologies in fraud prevention [1]. The effort by them highlights the scope of AI toward making fraud detection systems even more accurate and reliable.

Certain ML models have sparked interest in their applications for fraud detection. Delecourt and Guo proposed a mobile payment fraud detection system based on adversarial examples to make ML models robust against adversarial attacks [2]. This alone emphatically reveals the necessity of robust models that adapt to hold accuracy even under adversarial conditions.

Among these, Random Forest (RF) and XGBoost have particularly found great applications for fraud detection. Jabeen et al. have used the hybrid approach of RF with SMOTE to counter the imbalance issues in the detection of credit card fraud [3]. Their study as shown in the same exhibits how RF manages imbalanced datasets that are a typical problem also in fraud detection problems. More importantly, Zhang et al. applied an XGBoost model for unmasking customer transaction fraud whereby it emerged to be the best model for solving binary classification problems [4].

On a broader level, Afriyie et al. highlighted the employment of super-vised ML algorithms in uncovering fraudulent credit card transactions using credit cards, which further reinforced the significance of ML models in the detection of financial frauds [5].

Traditional ML models have been known to perform with a very high degree of efficacy in fraud detection. Their "black-box" nature forms the root of many problems concerning the explanation of what these models are doing. As an antidote to this, XAI (Explainable AI) has emerged as a sought-after solution for deepening understanding over the decisions of the model and providing a foundation for eliciting stakeholder trust.

Vihurskyi applied LIME for credit card fraud detection and stated that for all stakeholders to be comfortable, the models have to be explainable [6]. Kotrachai et al. used SHapley Additive exPlanations (SHAP) for comparison and evaluation of the models developed in order to identify credit card fraud thereby providing evidence that XAI can be applied toward enhancing transparency in fraud detection systems [7].

The article by Embarak provides a clearer view of XAI by referring to its applications in many fields, such as finance [8]. The review stresses the need to fill this gap currently existing between typically complex AI models and their end-users by making AI systems more interpretable and account-able.

In addition, XAI techniques play an important role in ensuring regulatory compliance. In this regard, Mishra et al. brought the issue of the adoption of XAI in the financial and accounting decision-making process and underlined the requirement of XAI in enhancing transparency and ensuring conformity with regulatory stipulations [9].

The mobile payment system spread resulted in new fraud vectors. Concerning those vectors, a particular detection method needs to be devised. Muqattash and Kharbat demonstrated that fraud is detected within the mobile payment system using ML methods. The authors successfully demonstrated that it is possible to apply the created ML model in real-time conditions [10]. Hajek et al. proposed a framework for the XGBoost-based fraud detection in a mobile payment system: a proof of concept of such an approach with tough fraud patterns [11].

To this end, support can also be drawn from Thar and Wai, who looked at predictive models that enable identification of fraud in digital banking [12]. According to them, the advancement of computational models has proven to enhance the accuracy and reliability associated with fraud detection systems.

Advanced AI and ML techniques, with techniques such as Federated Learning and XAI, make this complex technique simple for fraud detection. According to Awosika et al., this study presents the role of Explainable AI and Federated Learning toward the improvement of transparency and privacy in detecting financial fraud, along with a future direction to this topic [13].

Further, new ML models and methods are being developed and have been demonstrated by Nijwala et al. using the Extreme Gradient Boost (XGBoost) classifier in detecting credit card fraud an interesting

way forward in this line of research [14]. Raiter's work regarding supervised machine learning algorithms used in combating fraud in anti-money laundering further validates the importance of this area of research, with further developments reported [15].

Bello et al. took a broader view of the techniques within machine learning that is used in fraud detection in finance [16]. Their research highlights the relevance of the techniques in preventing ever-evolving fraud risks. The authors split their research into supervised, unsupervised, and deep learning methodologies and go on to shed light on the strengths and applications of those techniques in finance. This research contributes well towards the comprehension of the role of machine learning in enriching fraud prevention systems relevantly.

To get more insights on the application of various ML models in financial fraud detection, the extensive literature review conducted in 2024 is analyzed [17]. From this survey, it is identified that, majority of the research works were conducted on credit card fraud detection and especially credit card based loan fraud detection using ML models. In addition, it is inferred that, all these models were applied on real-time datasets and only a very less percentage was applied on synthetic dataset.

Jayanthkumar et al. [18] have applied Machine learning models to provide predictions on the changing scenario of monetary exchanges, and the investigation of most recent patterns utilized by financial organizations. They did an extensive review on the application of ML models to detect and prevent fake transactions in banking applications.

Chen et al. [19] have conducted a systematic review on the research works conducted on the application of deep learning models for financial fraud detection. Their review identified key challenges in data imbalance, automation, and explainability and privacy concerns. They also introduced a framework to find the sector specific constraints and regulatory requirements based fraud detection systems development. This work highlighted the importance of explainability in order to get the insights on how the model gives its predictions.

4. Summary of Findings & Motivation for Research

4.1. Summary

We put much emphasis on interpretability in mobile payment fraud detection in this study using XAI models. We start the research with training three machine learning models that were tuned up, Random Forest, XGBoost, and MLP, on a large dataset. Finally, we found the Random Forest model gave the highest performance to be the accuracy of 0.9982. This can be attributed to the ensemble nature of Random Forest, which combines several decision trees, hence reducing variance and maximally catching those intricate fraud patterns other models would otherwise miss. For that reason, precision and recall metrics further confirmed that Random Forest was the best performing model for fraud detection, a good balance between fraudulent transactions being detected and false positives being minimized.

However, the trustworthiness of the Machine Learning models must be further validated by means of the interpretations derived from the models is the need of the hour. Hence, we applied several XAI methods to interpret the predictions of this high-performing model—SHAP, LIME, Anchor, Global Surrogate, and EBM—to ensure this model can be trusted to deploy.

SHAP gives global and local feature importance, showing which feature influences the model's decisions on which data points. LIME allows one to generate explanations for individual predictions and boasts impressive model interpretability. Anchor simply computes rules with high fidelity that often lead to model decisions. The Global Surrogate provided a much reduced but still interpretable version of the Random Forest model. EBM offered an additive, interpretable model that performs similarly well.

Our analysis further finds that different XAI methods play complementary roles: they might be best at explaining certain predictions, yet may offer broader insights as to which features were indicative for the whole model. Hence, these findings underscore the role of XAI in AI-based fraud detection systems-that is, those making such systems more accountable and easier to understand for stakeholders and, therefore, more transparent and trustworthy.

4.2. Motivation

Since the days of civilization, human societies were based on transactions-transactions that, if you will, mean the exchange of goods and value for something between parties. During prehistoric times, people used the barter system a lot; one could get a piece of cloth for a bundle of grain. However, during the onset of currency, the transaction became uncomplicated, and the volume of payments started to increase. Jump to the age of comfort delivered by mobile payments. Today, millions of transactions are conducted daily with just a few taps on a smartphone. This wave of digital payments has brought unprecedented ease and speed into our lives, but it also created an enormous challenge: identification of fraudulent activities buried within that vast sea of legitimate transactions.

As the volume of digital payments increases with time, so does the scope of fraud. The large scale of generated data makes it imperative not only to store and manage such transactions but also to develop sophisticated methods for detecting fraudulent payments. This work is motivated by the urgent need to enhance the security and trustworthiness of mobile payment systems. Improving fraud detection models to be more interpretable toward empowering businesses, consumers by focusing on making fraud detection models more interpretive help businesses and consumers to understand the perception behind decisions deemed fraudulent equally and create in the process a safer, more transparent financial ecosystem.

With reference to the GDPR and PSD2 regulations on consumer data in order to protect personal information by personalized control and robust data management, this research work taken it as one of the crucial factors on using synthetically generated data instead of the original data. This synthetic data is generated from the original data using a simulator software without any personal information. Further, the decision-making by the models are assessed using XAI methods, which provides insights on the trustworthiness of the model's decision. Hence, this work satisfies the regulations of GDPR and PSD2 without compromising the quality of the research outcome.

5. Methodology

The proposed methodology is given in the form of a flow diagram as shown in Figure 1.

5.1. Explanation of Flow Diagram Components

Data Preprocessing cleans raw transaction data, normalizes it, and adds engineered features to the model (errorBalanceOrig, errorBalanceDest) to avoid inaccuracy in the model. 1: Fraud and 0: Non-fraud

Train/Test Split splits the given dataset into training and test datasets with stratified representation of fraud and non-fraud case. Thus, no data imbalance would be encountered at the time of the model-training phase.

Training ML Models: Three Models-Random Forest, XGBoost, and MLP-have been trained for fraud detection. All the models are fine-tuned with optimum performance. Among all three models, the Random Forest becomes the best model owing to its high degree of accuracy, resistance towards overfitting, and efficiency in handling huge datasets.

Model Evaluation compares the performance of all trained models that emerge with Random Forest for better performance metric. Accuracy, Precision, and Recall values are measured to judge model success.

XAI Integration provides multiple explainability methods of the Random Forest model to enhance transparency in the decision-making process. Each XAI method provides different perspectives of feature influences on fraud predictions. SHAP provides global and local feature importance, LIME gives localized explanations, while Anchor offers rule-based insights. Global Surrogate and EBM give global model behavior explanations.

Fraud Detection Output produces an interpretable prediction in which each transaction classification as fraud or not fraud is also accompanied by an explanation, making the system more transparent and trustworthy for its end-users.

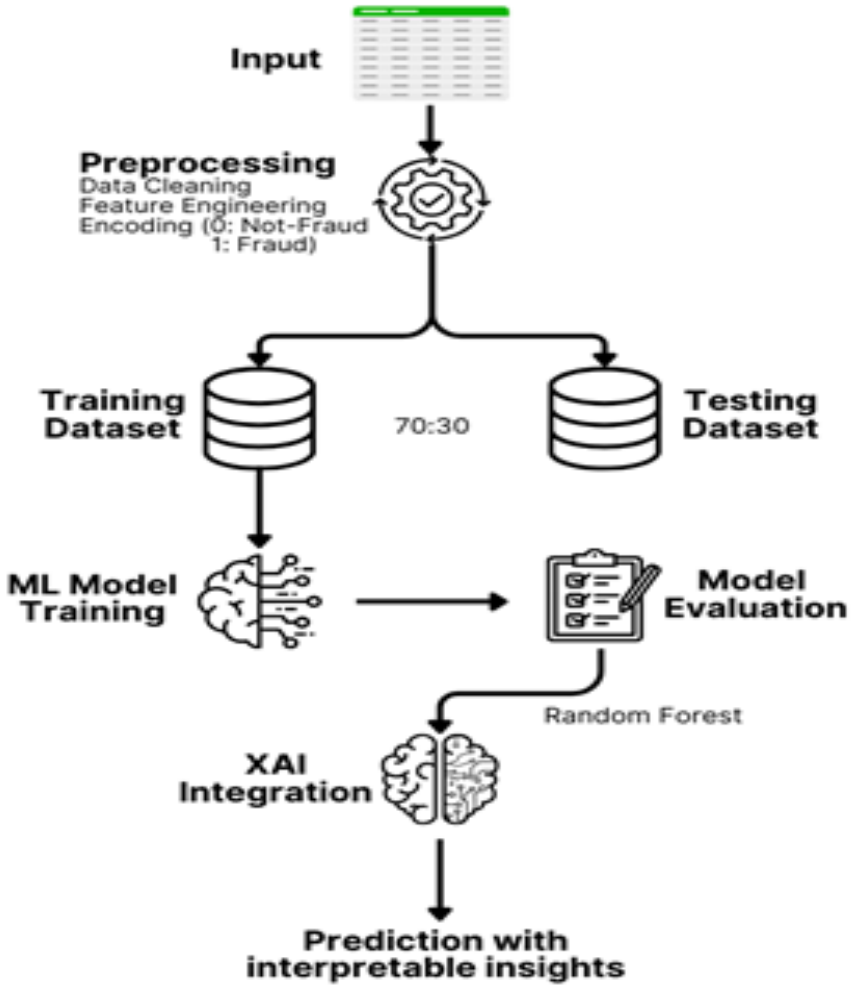


Figure 1. Proposed Methodology for Validating ML models with XAI.

5.2. Trustworthiness Validation Metrics

We applied multiple XAI models to interpret the all three models that we are applying in mobile payment fraud detection. The techniques varied in terms of scope, stage, type, and output formats as follows:

SHAP (Scope: Global & Local, Stage: Post-hoc, Type: Classification, Output: Feature importance values)

SHAP gave global along with local interpretability by providing importance scores of features for every prediction. It offered complete insights into how each feature influenced model outputs at both the instance and dataset level. The formula given in equation (1) is used to validate the ML model

$$\text{Fidelity} = \frac{(FP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

Further, the XAI models validation are evaluated using metrics given in eqn. (2) and (3).

$$\text{Unambiguity} - \frac{1}{1 + Var(Feature\ Weights)} \quad (2)$$

$$\text{Interpretability Size} - \frac{1}{N}\sum_{i=1}^{N} Count\left(F_i \neq 0\right) \quad (3)$$

LIME (Scope: Local, Stage: Post-hoc, Type: Classification, Output: Local feature importance)

LIME was one form of local interpretability that focused on instance-level predictions. It created surrogate models for specific instances in order to understand the behavior of the model through easy-to-understand explanations but applied only to localized predictions. The following equation eqn. (4) is used to validate the model and eqn. (5) and (6) are used to validate the XAI model.

$$Fidelity - \frac{TP + FP + FN + TN}{TP + FP} \quad (4)$$

$$\text{Unambiguity} - \frac{1}{1 + Var(\ Feature\ Weights)} \quad (5)$$

$$\text{Interpretability Size} - \frac{1}{N}\sum_{i=1}^{N} Count\left(\ F_i > 0\right) \quad (6)$$

Anchor Explanations (Scope: Local, Stage: Post-hoc, Type: Classification, Output: Rule-based explanations)

Anchor explanations were very interpretable at the local level since they provided clear rule-based descriptions of individual instances. That simplicity and lack of ambiguity made it very user-friendly. The metrics were calculated using equations (6), (7) and (8).

$$\text{Fidelity} - \frac{TP}{TP + FP} \quad (6)$$

$$\text{Unambiguity} - \frac{1}{1 + Var(\ Precision)} \quad (7)$$

$$\text{Interpretability Size} - \text{Number of features in the anchor} \quad (8)$$

Global Surrogate Model (Scope: Global, Stage: Post-hoc, Type: Classification, Output: Reduced model)

The Global Surrogate model offered global interpretability through approximation of the overall behavior of Random Forest using a simpler model (such as a decision tree). This can be more holistic in nature as the decisions that are taken at each level of the model could be understood in a more comprehensive way. The metrics were calculated using equations (9), (10) and (11).

$$\text{Fidelity} - \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Unambiguity} - \frac{1}{1 + Var(\ Feature\ Weights)} \quad (10)$$

$$\text{Interpretability Size} - \frac{Number\ of\ \ Nodes}{Maximum\ depth\ of\ \ the\ tree} \quad (11)$$

Explainable Boosting Machine (EBM) (Scope: Global, Stage: Pre-hoc, Type: Classification, Output: Feature-level understanding)

EBM inherently balanced interpretability with accuracy by producing interpretable models that also provided global explanations. The method gave insights into how individual features contributed to predictions at a global scale while still being perfectly adequate for good predictive performance. The metrics were calculated as follows.

$$\text{Fidelity} - \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$\text{Unambiguity} - \frac{1}{1 + Var(\ Feature\ Weights)} \quad (13)$$

$$\text{Interpretability Size} - Count(\ F > 0) \quad (14)$$

6. Experimentation and Result Analysis

6.1. Experimentation Design and Setup

The experimentation setup and the dataset generation methodology are discussed in the subsequent sections. The implementation code and the dataset are available in the GitHub repository [20].

6.1.1. Hardware and Software

The experimentation is done with the experimental set-up of Colab GPU 15GB with a memory size of 112GB (100GB used), with a system configuration of RAM 16GB and hard disk space of 1TB.

The Machine Learning Models have been trained on Colab via the Scikit-Learn Library and the different XAI models implemented, namely LIME, SHAP, Anchor Explanations, Global Surrogate Model and Explainable Boosting Machines (EBM), were all trained via their respective libraries. The metrics for validating the performance of the XAI models, that is, Fidelity score, Unambiguity and the Interpretability of the models are all implemented either manually or using specific libraries.

6.1.2. Dataset Generation using PaySim

For performing the validation of different Machine Learning Models for Mobile Payment Fraud

Detection, we used the Synthetic Financial Datasets for Fraud Detection; a simulated dataset generated using a simulator called PaySim to create a similar environment.

6.1.3. About PaySim

PaySim is a financial transaction simulator tool that simulates the behavior of a real-time mobile money system using synthetic data generated from real transactional logs. It uses aggregated data from the private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behavior to later evaluate the performance of fraud detection methods [21].

The generated data is a simulation of mobile money transactions based on a sample of real transactions extracted from financial logs from a mobile money service. For this research, the dataset has been under-sampled from a total size of 636,000 to a size of about 16,000 samples balance the number of fraud and not fraud transactions.

As we cannot get real-time dataset for financial fraud detection due to privacy and security concerns raised by the financial institutions due to the sensitivity of transactions data, this work has used the simulator to generate synthetic dataset. However, this dataset is not created by any random way or mathematical calculations; rather, the tool used the private dataset as the basis and generate similar data that exactly resembles the actual transactions. And then it injects malicious behaviors in order to evaluate the performance of the various ML models. Hence, there is no limitations in terms of analysis using this synthetic data for this research work. However, the real-time malicious attacks may not be available for real-time analysis.

6.2. Exploratory Data Analysis

A sample generated dataset is given in Table 1. The individual features descriptions with sample data are given in Table 2.

Table 1. Sample data from the PaySim synthetic dataset.

	step	**type**	**amount**	**nameOrig**	**oldbalanceOrg**	**newbalanceOrig**	**nameDest**	**oldbalanceDest**	**newbalanceDest**	**isFraud**	**isFlaggedFraud**
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

Table 2. Features Description.

Feature	**Description**	**Sample**
Step	Time step in hours (1-744)	345
type	Type of transaction	CASH_OUT, TRANSFER, DEBIT, CREDIT, PAYMENT, CASH_IN
amount	Transaction Amount	1548.75
nameOrig	ID of origin account	C840083671
oldbalanceOrg	Initial balance of origin amount	415582
newbalanceOrig	Balance after transaction	4587868
nameDest	ID of destination account	M1948758
oldbalanceDest	Initial balance of Destination account	0.0
newbalanceDest	Balance after Transaction	5487.50
isFraud	1 if transaction is fraudulent	0 or 1
isFlaggedFraud	1 if transaction automatically flagged as fraud	0 or 1

6.2.1. Summary of Data Generated

Dataset Size is 636,000 (approx...)
Time duration : 30 days
Fraudulent Transactions: 0.13% of actual data (approx.)
Null Data – There is no null data identified
Categorical Variables – Payment Type Distribution Plot
The exploratory data analysis based graphs are given from Figures 2–4.

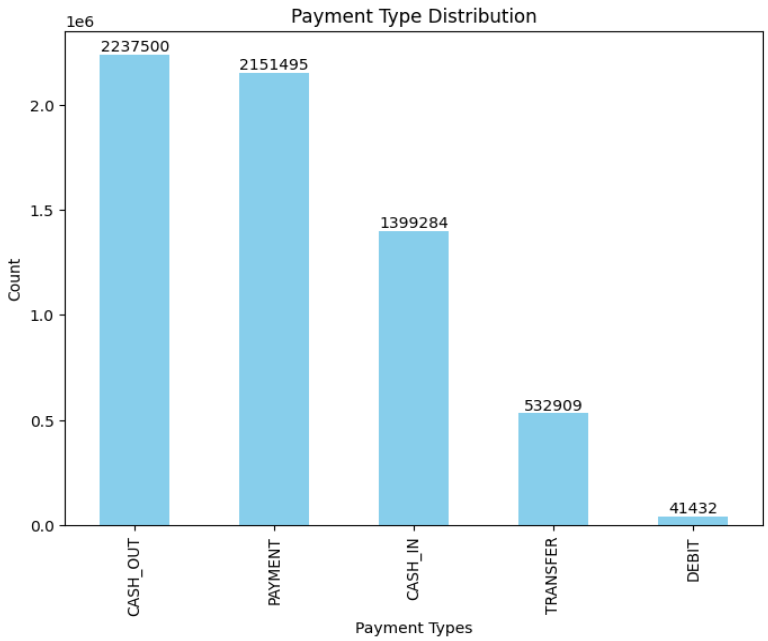


Figure 2. Exploratory Analysis of Data – Payment type distribution graph.

Number of Transactions by Type Vs. Fraud Status

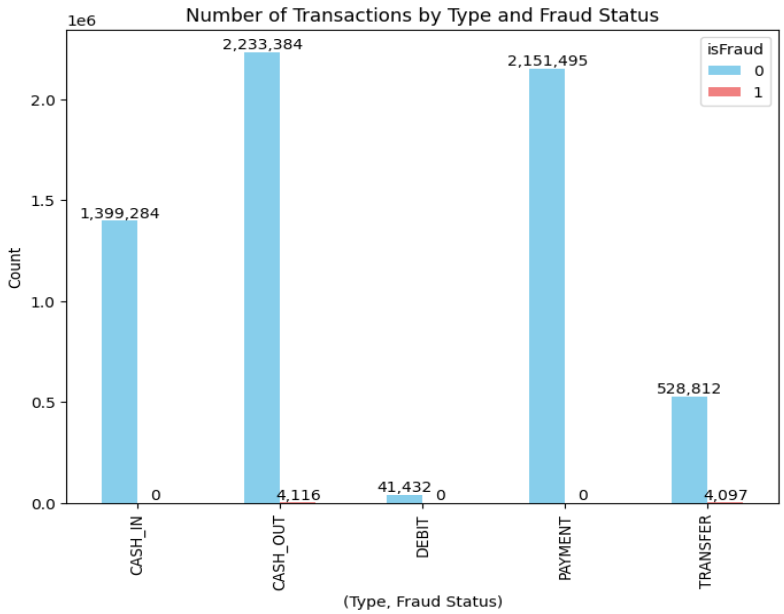


Figure 3. Exploratory Analysis of Data – Number of transactions by type.

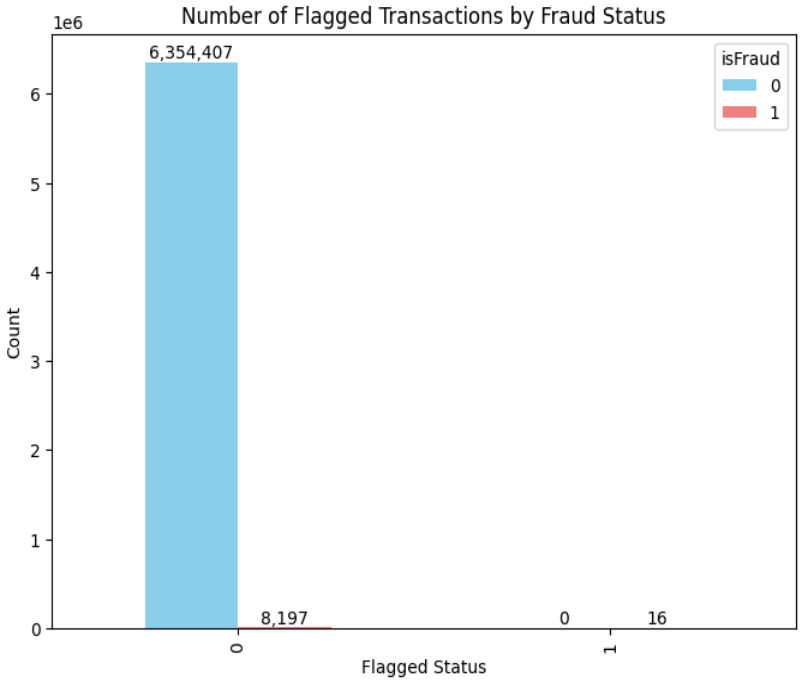


Figure 4. Exploratory Analysis of Data- Number of flagged transactions Vs. Fraud Status.

6.2.2. Exploratory Data Analysis on Dependent Variable

The exploratory data analysis result from the dataset based on dependent variables is given in Figure 5. Here the dependent variable is 'Fraud Status'.

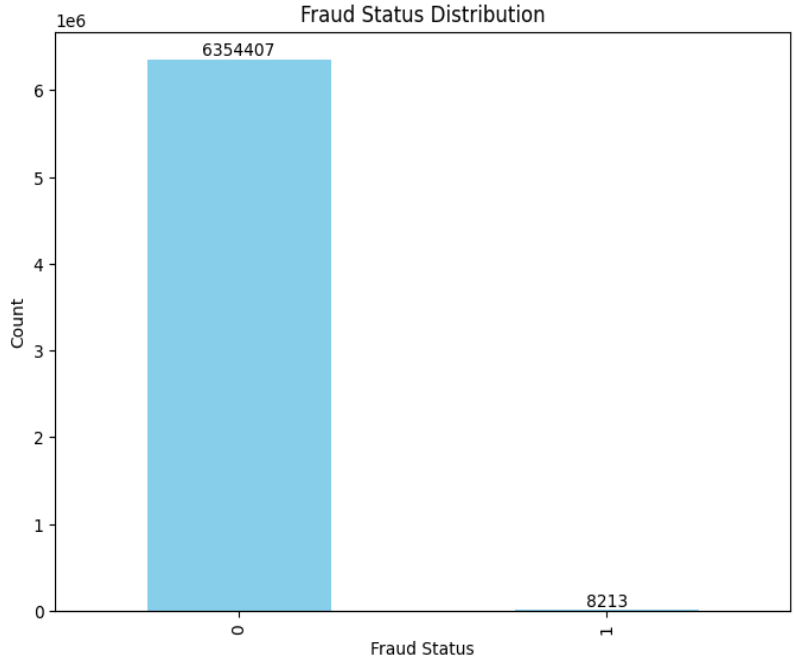


Figure 5. Exploratory Data Analysis on Dependent Variable – Fraud status distribution

The dataset generated for one hour time step distribution is given in Figures 6 and 7 for fraud as well as non-fraud categories.

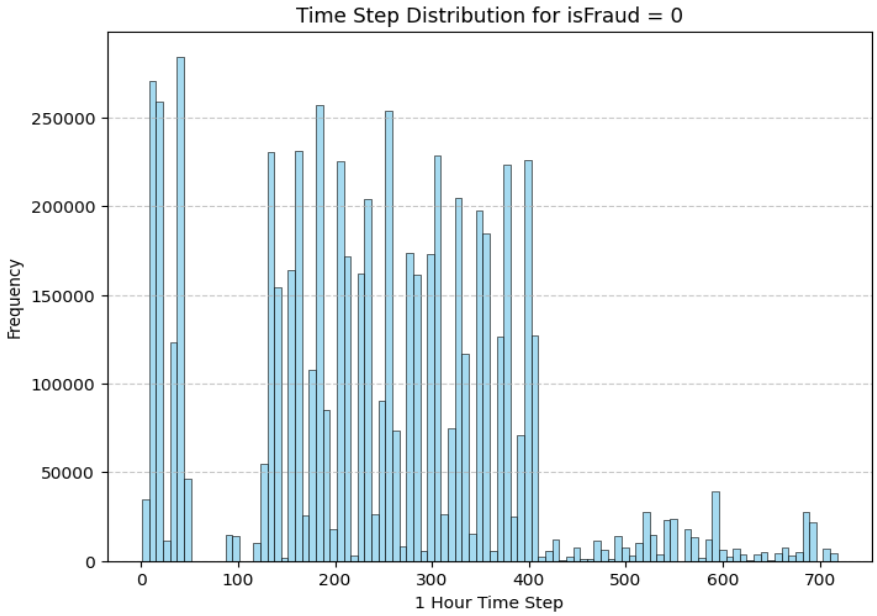


Figure 6. Time step distribution of data generated for isFraud=0.

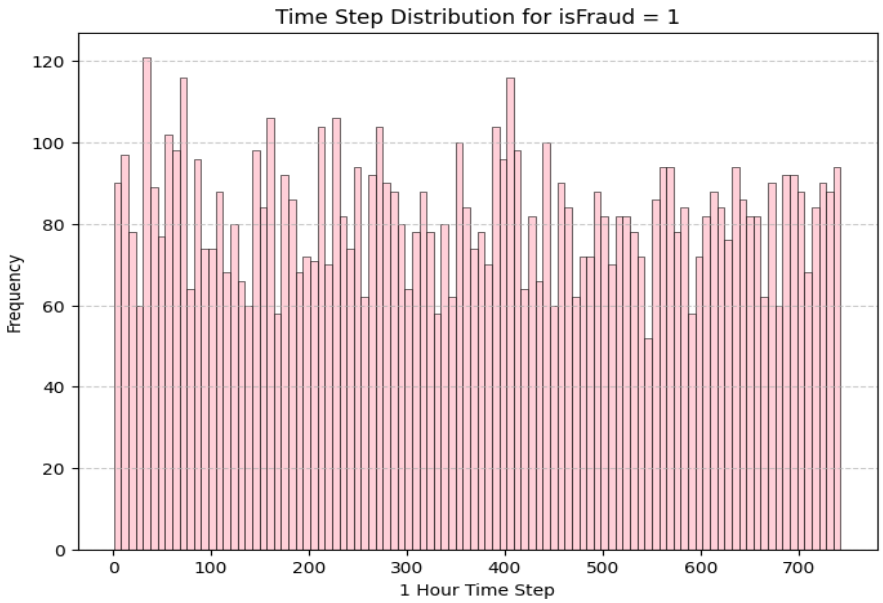


Figure 7. Time step distribution of data generated for isFraud=1.

Based on the different features values, the fraud status is classified. The various intendent features such as Transaction Amount, OldbalanceOrigcust, OldbalanceOrigrec vs. the dependent variable fraud status is depicted in Figure 8. A box plot on dependent and independent features in the dataset is given in Figure 9.

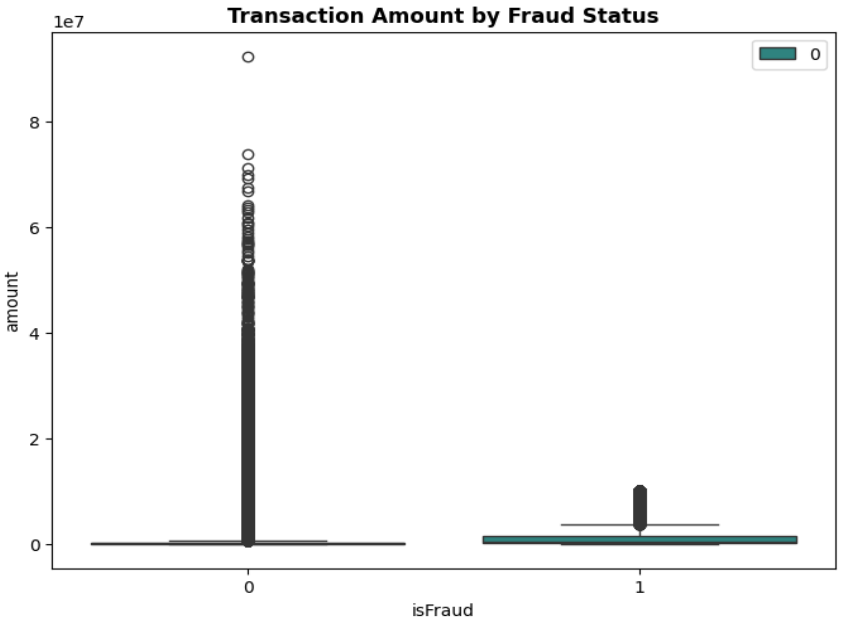


Figure 8. Transaction Amount by Fraud Status Distribution.

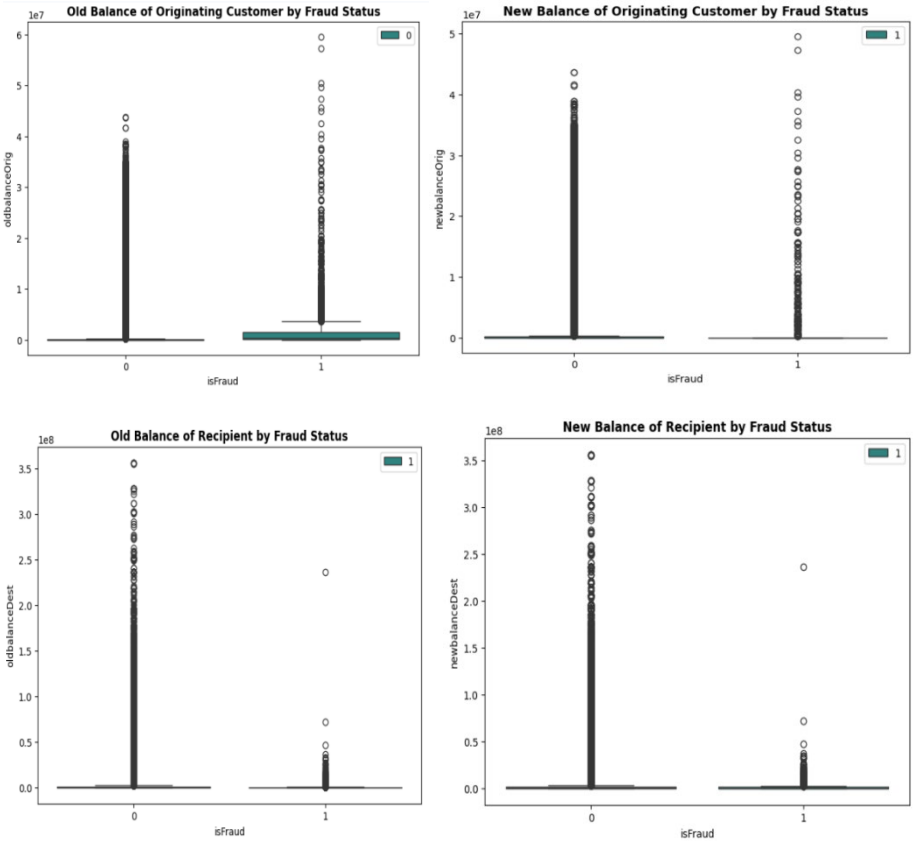


Figure 9. Box plot of independent variables Vs. Dependent variable.

The correlation heatmap visualization of the features in the dataset is given in Figure 10. From this visualization, it will be easier to find the correlation between the features and their complex patterns. As the diagonal cells provides overall 1 as the value, it indicates a strong correlation between the features.

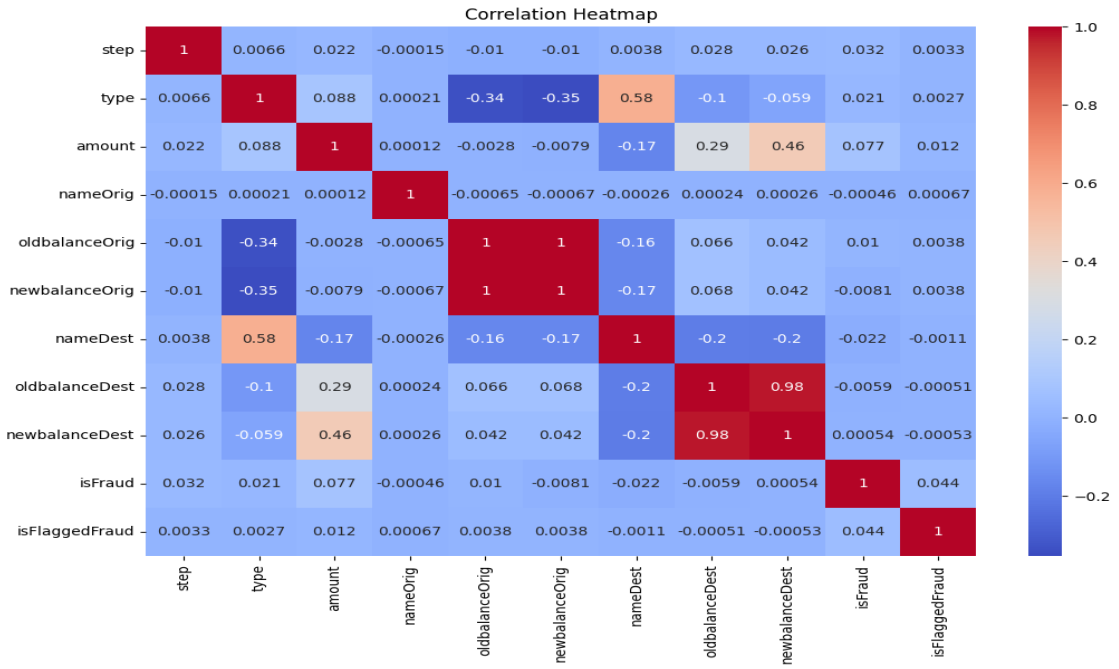


Figure 10. Correlation Heatmap of the features in the dataset.

6.3. Data Leakage and Caveats related to Dataset

The PaySim dataset though it is widely used for several research works it has certain known caveats and data leakage risks that is acknowledged in this work. Some of the issues are: (i) label leakage by means of balance features - for fraudulent transactions, the destination account balances with feature names as oldbalanceDest and newbalanaceDest, their values often remain 0 or inconsistent as the fraudulent transactions are artificially interrupted during simulation; (ii) temporal and account identity leakage – as the dataset does not anonymize the data, random splitting of data may allow some users information may appear during training or testing; and (iii) imbalanced label caveat – fraudulent transactions make up less than 13% transactions as fraudulent and hence modes achieve higher accuracy. However, these can be resolved by means of F1 Score and AUC values apart from accuracy.
Hence, this research work has calculated F1 Score and ROC-AUC curve based metrics to assess the model's performance.

6.4. Machine Learning Models Implementation

The dataset was varied to include features relevant to mobile payment fraud detection such as the amount of transaction and balance before and after transactions and engineered features such as errorBalanceOrig and errorBalanceDest. Stratified sampling was used to split data into training and testing sets in a 70:30 ratio so as not to induce stratification through sampling. This consisted of 5749 fraud cases and 5749 non-fraud cases for the training set, while the test set consisted of 2464 fraud cases and 2464 non-fraud cases.

To that end, we tested three machine learning models-for accuracy and applicability in fraud detection for mobile-based business transactions. The tuned Random Forest model stands out as the most accurate from our results with an accuracy of 0.9982, followed by XGBoost at 0.9976 and MLP at 0.9840. Each model has peculiar strengths and weaknesses. For example, Random Forest holds an ensemble structure that reduces overfitting and generally high predictiveness within imbalanced datasets even though it requires wide-range interpretability support. XGBoost can pick very complex fraud patterns nearly accurately but suffers on the computation side and for interpretability. MLP is strong for a non-linear relationship but relative to accuracy and low for interpretability hence unsuitable for high-trust applications. In general, methods of XAI complement Random Forest and make this model the most reliable one for successful and readable accurate detection of fraud in mobile-based transactions. The Table 3 presents the base models' analysis alongside accuracy, precision, recall, and F1-score for each model.

Table 3. Performance metric findings of base models. (0: Not Fraud; 1: Fraud)

Model	Accuracy	Precision	F1-Score
A1: Random Forest	0.9980	0.9964 (for 0), 0.9996 (for 1)	0.9980
B1: XGBoost	0.9976	0.9968 (for 0), 0.9984 (for 1)	0.9976
C1:Multilayer Perceptron	0.9564	0.9229 (for 0), 0.9956 (for 1)	0.9580

Table 4 shows the results of the hyper-parameter tuned models applied to the dataset for mobile payment fraud detection. The table summarizes key performance metrics including accuracy, precision and F1-score for each model.

Table 4. Performance metric findings of hyper-parameter-tuned models. (0: Not Fraud; 1: Fraud)

Model	Accuracy	Precision	F1-Score
A2: Tuned Random Forest	0.9982	0.9964 (for 0), 1.0000 (for 1)	0.9982
B2: Tuned XGBoost	0.9976	0.9964 (for 0), 0.9988 (for 1)	0.9976
C2: Tuned MLP	0.9840	0.9826 (for 0), 0.9853 (for 1)	0.9840

6.5. Machine Learning Models Comparative Performance Analysis

The performance analysis of all the ML models applied is given in Figure 11. It indicates tuned XGBoost provided the ROC curve with AUC value 0.9976, which is making the model to be suitable for further applications. The feature importance graph generated from the tuned XGBoost is given in Figure 12.

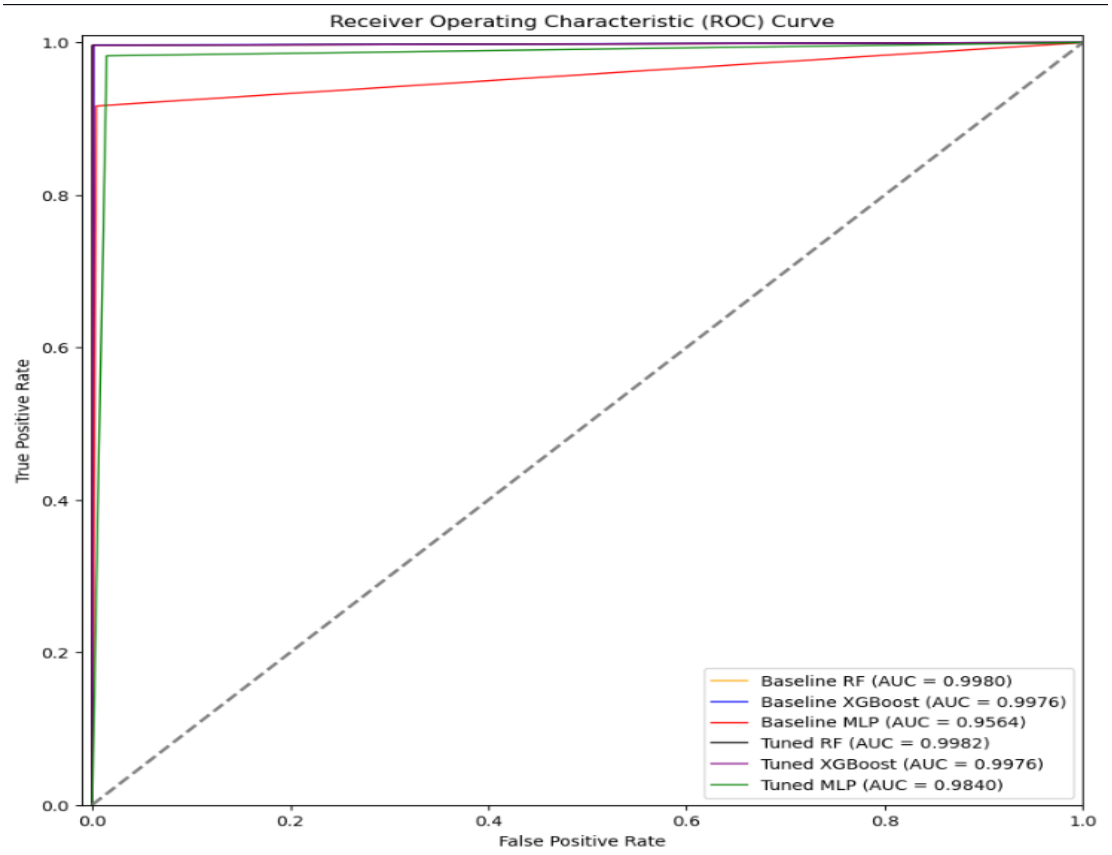


Figure 11. Comparison of ROC Curves of ML Models.

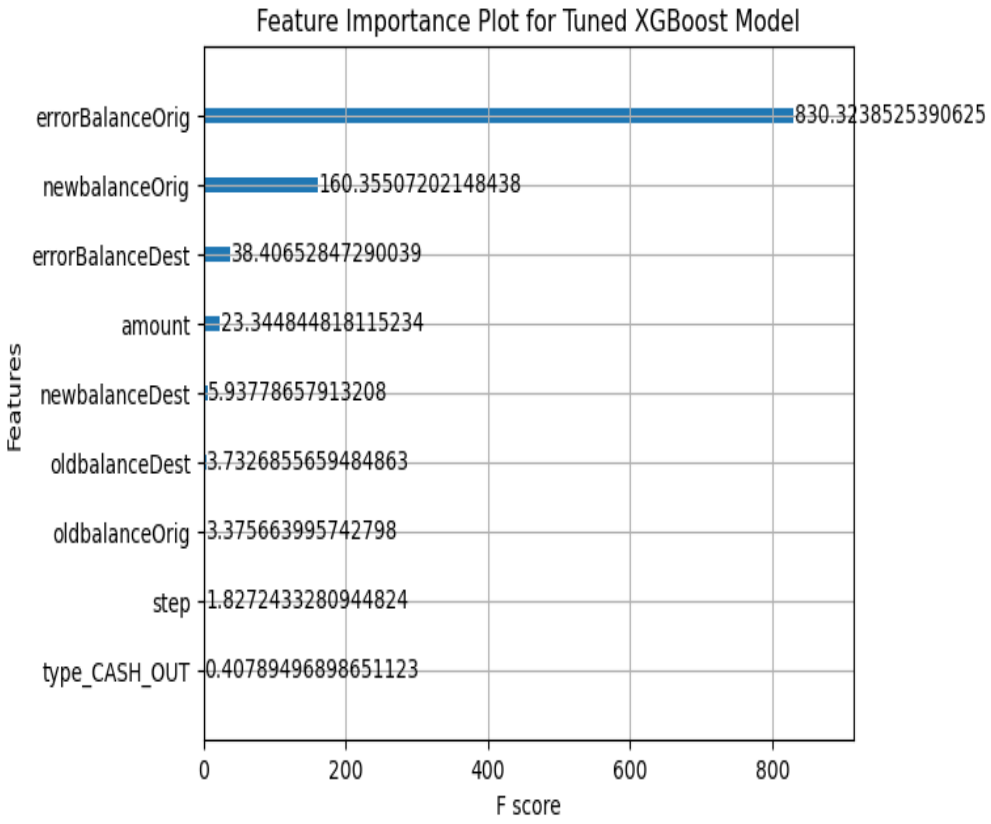
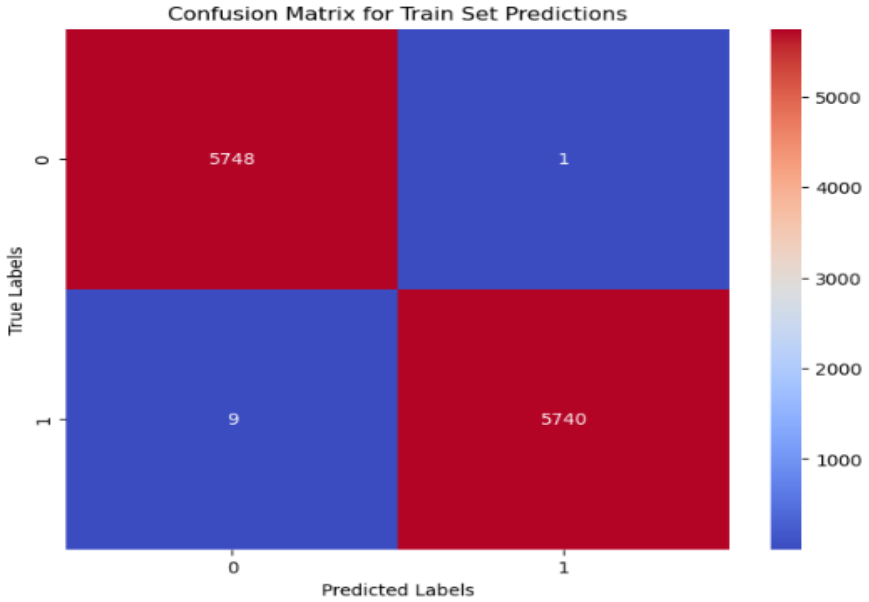


Figure 12. Feature Importance Plot for Tuned XGBoost Model.

6.5.1. Confusion Matrix

The Figures 13–15 provide the confusion matrix visualization of the metrics calculated from the Random Forest, XG Boost and NLP classifier models.



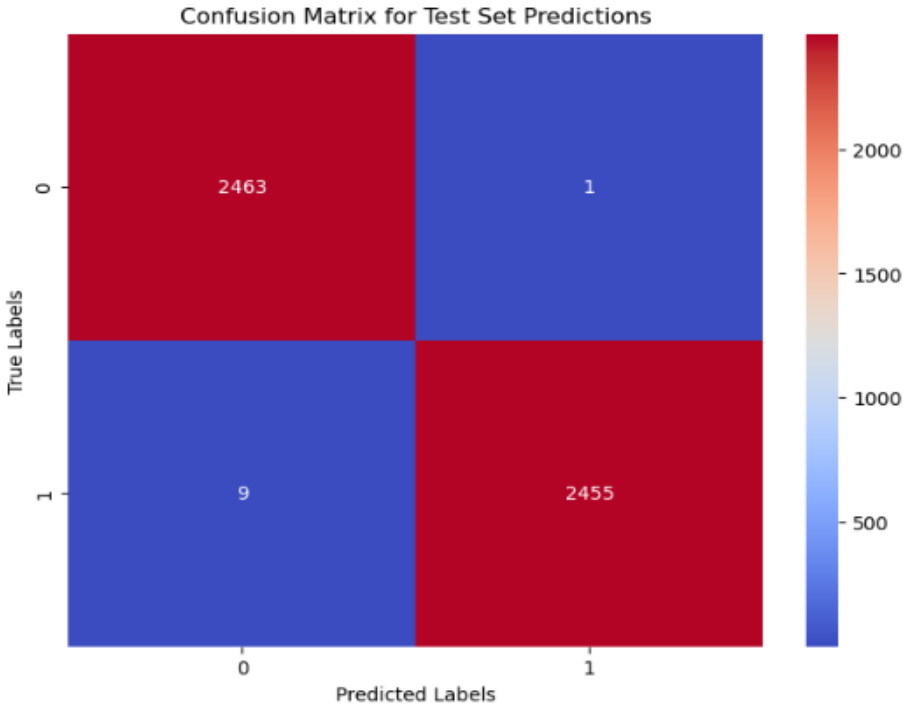
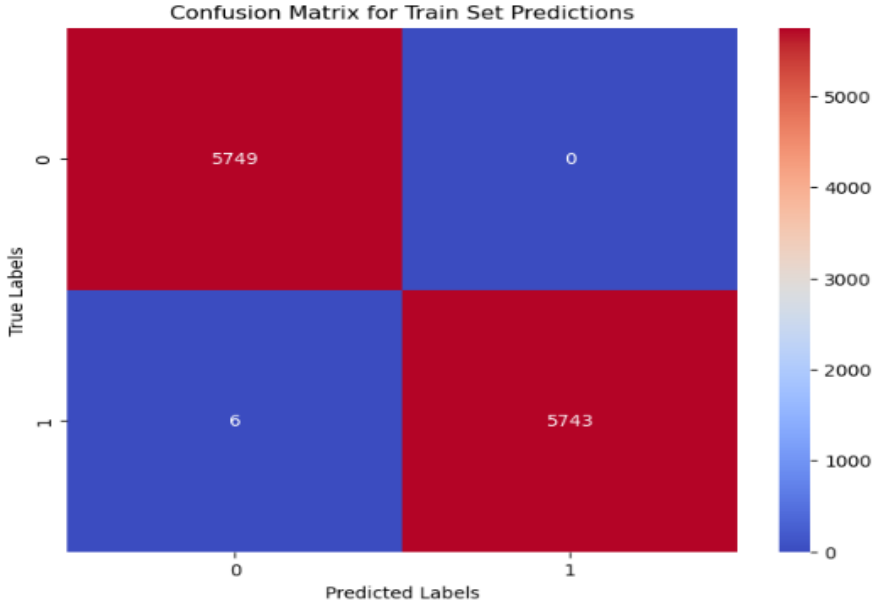


Figure13. Confusion Matrix for training and Testing of Random Forest Classifier.



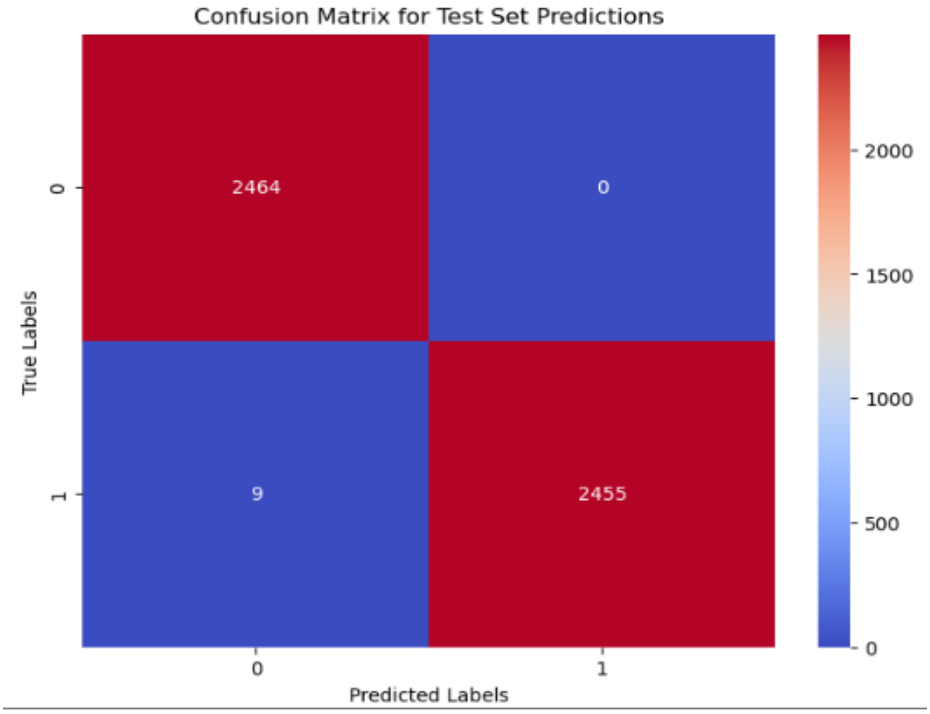
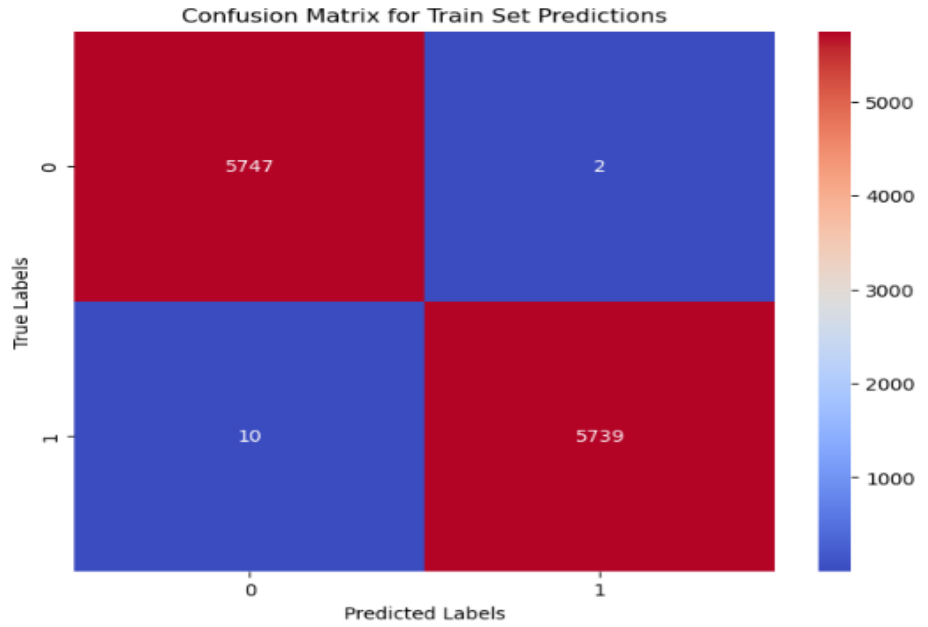


Figure 14. Confusion Matrix for Training and Testing of XGBoost Classifier**.**



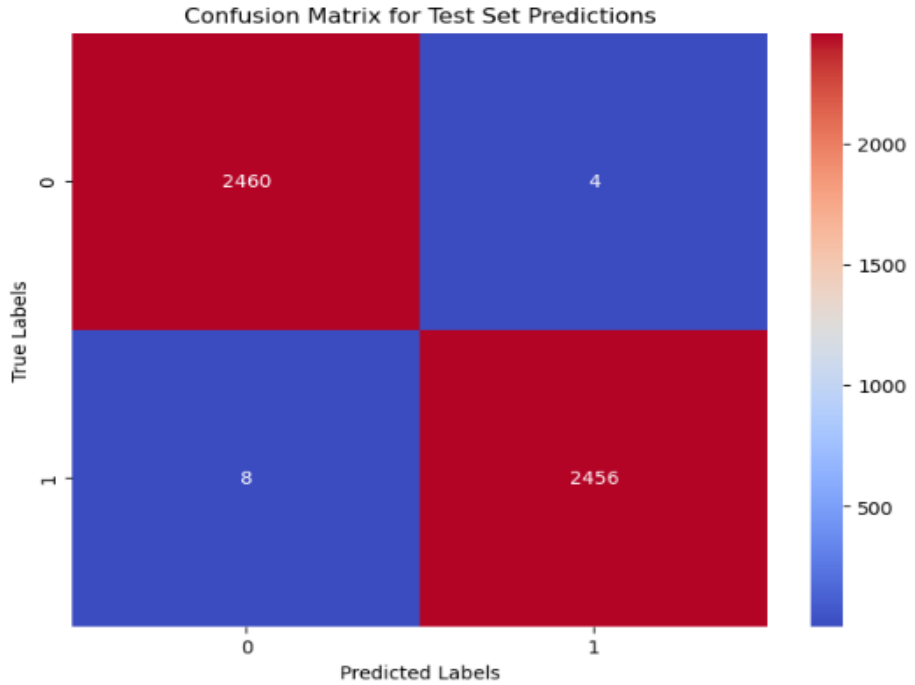


Figure 15. Confusion Matrix for Training and Testing of NLP model.

The Random Forest classifier has emerged as the best model for mobile payment fraud detection, as evidenced by its performance metrics in Table 2 as well as Figure 3. With an impressive accuracy of 99.82%, it outperformed other models in precision, recall, and F1-score, achieving a precision of 99.64% for non-fraud cases and 100% for fraud cases. Its balanced performance minimizes false positives while reliably identifying fraudulent transactions. Additionally, Random Forest's capability to provide insights into feature importance enhances transparency, making it easier for stakeholders to understand decision-making. Its resilience to overfitting further contributes to its robustness, solidifying its selection as the optimal model for this study.

6.6.2. AUPRC Metric

The metric Area Under the Precision-Recall Curve (AUPRC) summarizes the model performance across all thresholds. It is especially useful when positive class is rare as it focuses on how well positives are separated from negatives.

From the confusion matrix of tuned XGBoost model, class prevalence is computed as 0.5; TP=2455; FN = 9; Precision = 0.9984; FP=4; TN=2464 which leads to accuracy of 0.9976 and Recall=0.9968. Now, AUPRC which is the area under Precision and Recall is 0.9801 (as precision >= .99 and recall >=0.99) and the adjusted AUPRC is 0.996.

It indicates the model can perfectly distinguish between the positive and negative classes. Both Sensitivity and Specificity are balanced and they exhibit the strength of the model.
From all these calculations, it is identified that, the tuned XGBoost classifier is highly reliable and can be safely used for decision making with minimal risk of false positive and false negative cases.

As AUPRC emphasizes the performance on the positive class, the result we got shows the excellent detection of true positives even under higher precision requirements.

6.6. XAI Implementation

XAI models developed within our research were used to increase the interpretability of the Random Forest classifier that proves out as the best overall-performing machine learning model for mobile payment fraud detection. Including XAI techniques will give the stakeholders an understanding of the decision-making process in the "black box" of the Random Forest model itself, the ways through which predictions are made, and what factors influence the outcome.

Analysis of XAI models shows some of the main strengths and weaknesses: LIME, Anchor, and Surrogate models show the highest fidelity, as they are closest to Random Forest predictions, thus increasing the confidence in the decision-making process. LIME and Anchor show the greatest clarity and explanation is the simplest for user understanding. The compact interpretability size of Anchor makes

it suitable when simplicity is required, whereas SHAP and EBM give more detailed, though complex, insights. Therefore, these two are helpful for LIME and Anchor when their explanations are clear and short, but SHAP and EBM can be further applied.

6.6.1. SHAP (Shapely Additive Explanations)

SHAP values give an integrated measure of feature importance as it assigns a score to each feature for a given prediction. This method gives deeper insight into the output given by the model and elevates the contribution of every feature in the decision-making process. The global and local (instance based) explanations generated from SHAP are given in Figures 16 and 17 respectively.

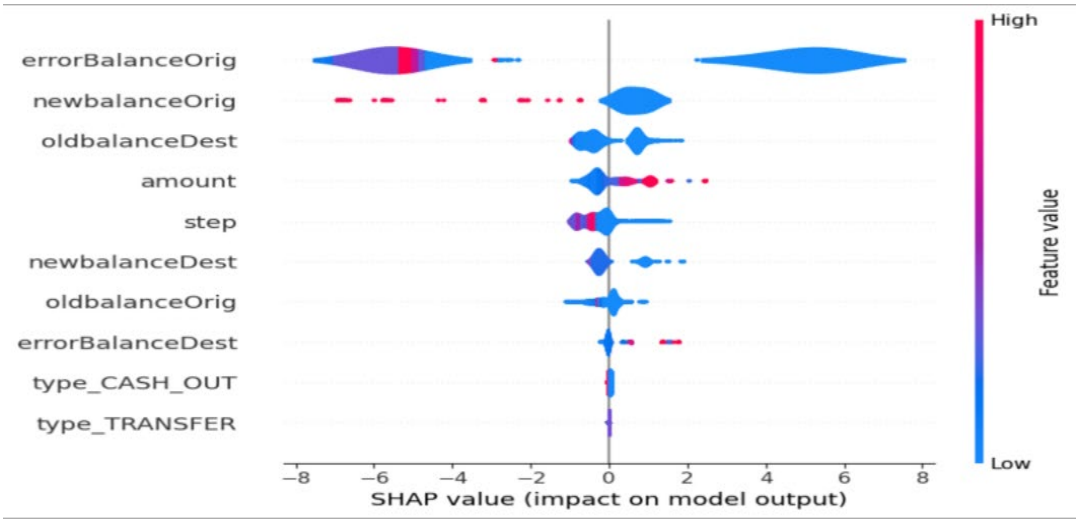


Figure 16. Summary Plot of SHAP for Global Explanations.

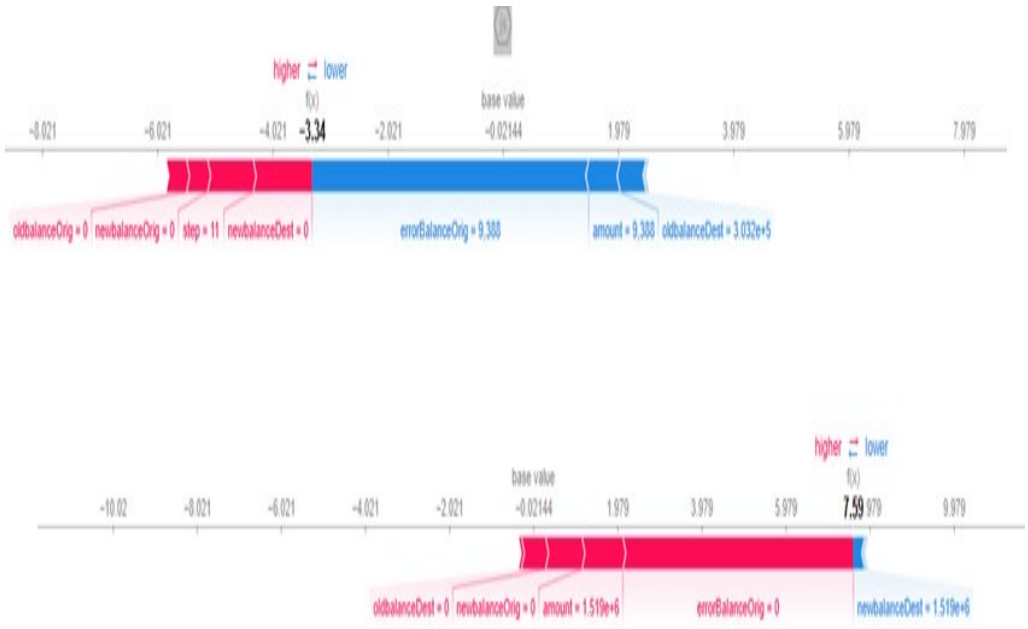


Figure 17. SHAP Plot for Local Explanation of instances (114803 and 6338605).

6.6.2. LIME (Local Interpretable Model-agnostic Explanations)

LIME focuses on explaining individual predictions by approximating the Random Forest model with a simpler, more interpretable model locally around the prediction. It helps identify the most impactful features for specific instances, thus improving transparency. The explanations from local explanations for two different instances from LIME model are given in Figures 18 and 19 respectively.

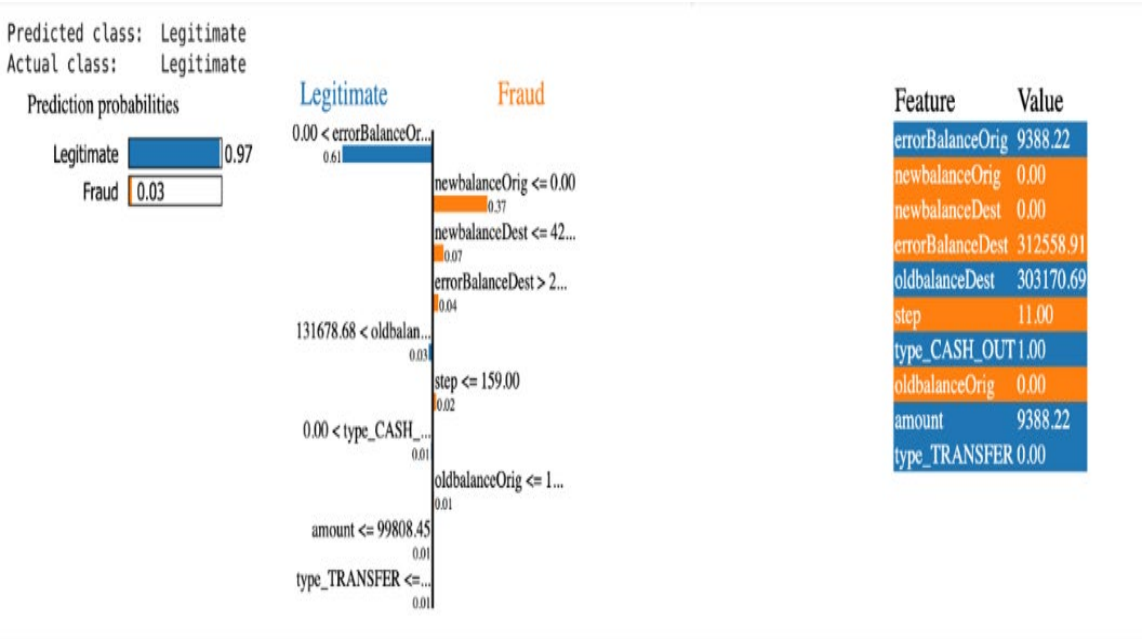


Figure 18. LIME Visualization for local Explanation of instance [114803].

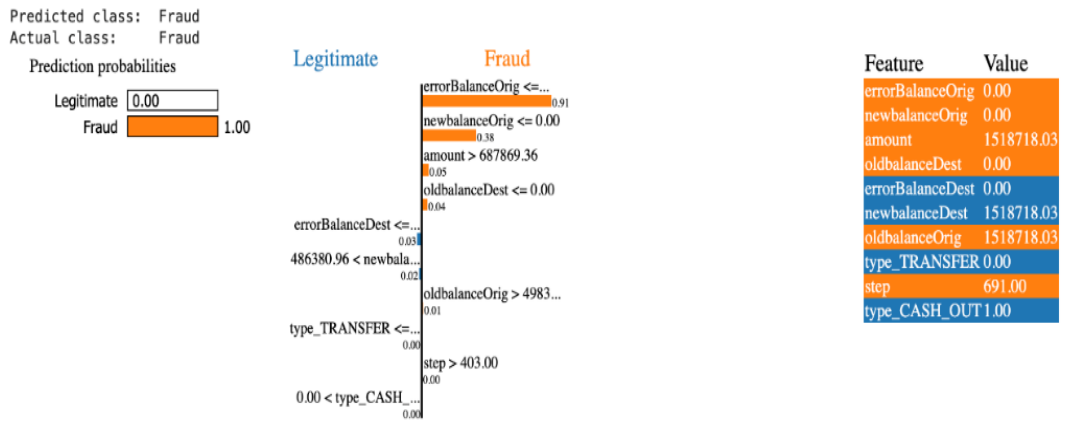


Figure 19. LIME Visualization for local Explanation of instance [6338605].

6.6.3. Anchor Explanations using Alibi

Anchor explanations is a local interpretability method that identifies specific values of features important to a model's prediction. It makes explicit "anchors" for under which conditions a model's prediction is reliable. Anchors are provided as intuitive rules, hence explaining why a certain decision occurred.

This approach not only makes it more transparent but enables users to understand what drives the predictions, which results in increased trust in the model's outputs and makes informed decisions possible. The anchor explanation provided by the alibi library is very useful for interpreting complex machine learning models. The first three instances explanations derived from Anchor Explanation model is given in Figure 20.

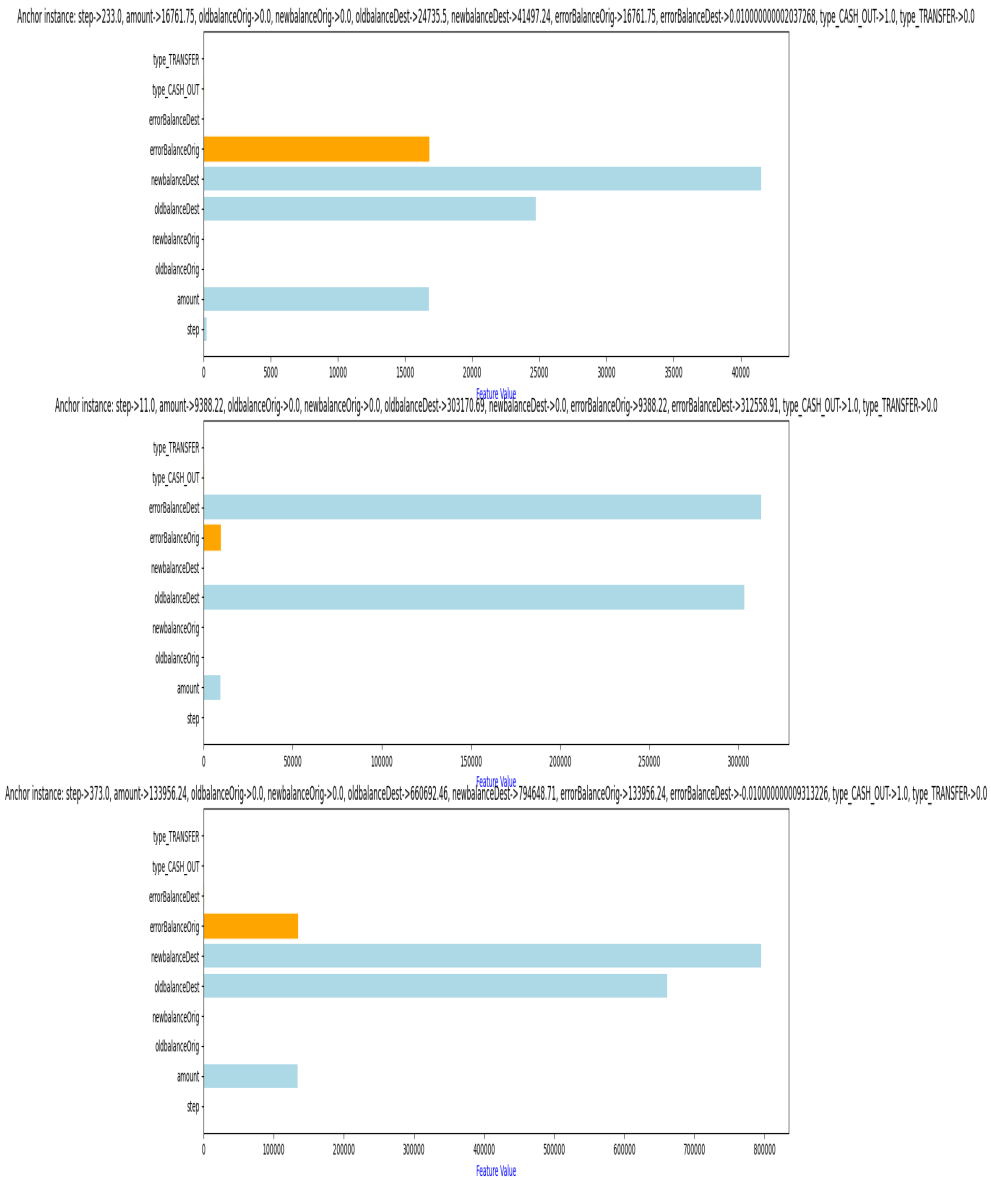


Figure 20. Anchor Explanations for first 3 instances, title is the instance values.

6.6.4. Global Surrogate Model

Global surrogate models serve as understandable approximations of rather complex machine learning models, which provides insights into behavior that the model itself does not provide. This method captures important patterns in data, further making it interpretable with feature importance and decision boundaries, by training a simpler model over a decision tree is used on the predictions of the original. It makes it possible for stakeholders to understand how the complex model makes predictions everywhere in the dataset. It promotes transparency and helps induce trust in machine learning systems. The decision tree's visualization as a global surrogate model is given in Figure 21.

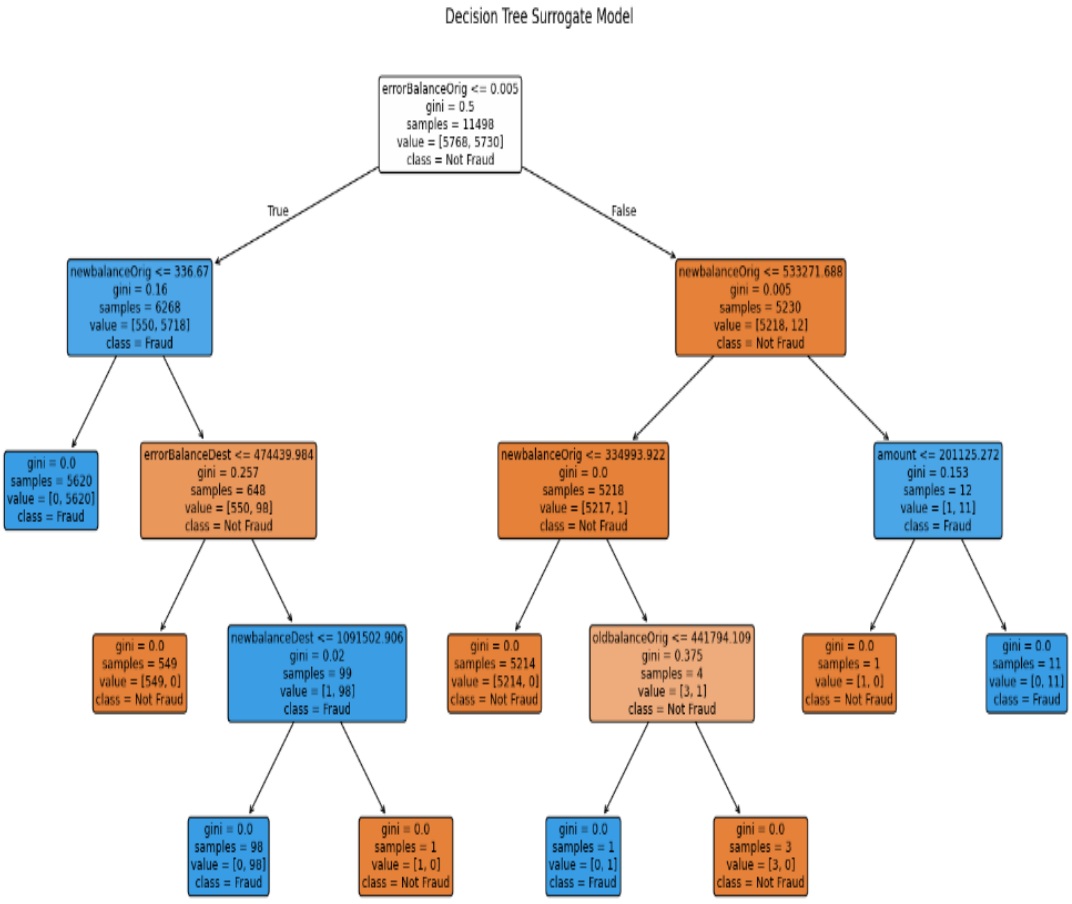


Figure 21. Global Surrogate Model using Decision Tree Classifier.

6.6.5. Explainable Boosting Machine (EBM)

EBM is a class of interpretable models, which retain the power of boosting with interpretability. In an EBM, we have an ensemble of simple models where the contributions of individual features are easy to understand. The component plots of each of these features visualize their contribution toward prediction by clearly showing how changes in values contribute to the outcomes. This clarity gives stakeholders insight into the model's decision-making process and fosters trust and accurate decision-making at high levels of performance. The model's feature-wise comparison between EBM and the best performing model is given in Figure 22.

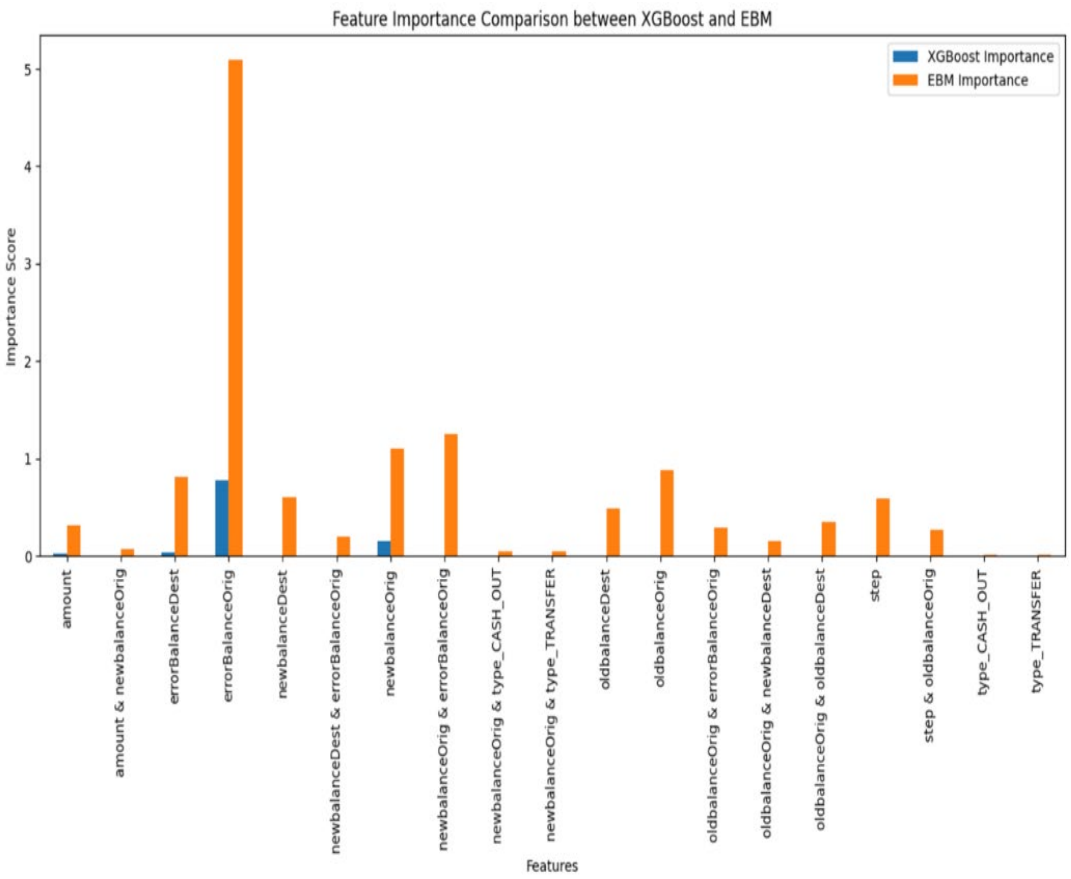


Figure 22. Feature comparison EBM vs XGBoost Best Performing Model.

7. Validation of ML and XAI models

In our study, we applied all the above-mentioned XAI models in order to improve the interpretability of our Random Forest classifier: SHAP to EBM. Each one of these methods has its advantages and uniqueness: SHAP provides a fine-grained feature importance analysis based on the prediction level, Anchor Explanations provide understandable rule-based insights, and Global Surrogates simplify the model behavior at a global level, and EBM balances accuracy with transparency. Through the following comparative analysis, we evaluate each method based on fidelity, unambiguity, and interpretable size, which would help in the identification of the best-suited XAI for our fraud detection model.

The Tables 5, 6, 7, and Figures 23–25 given below report the main evaluation metrics of XGBoost, Random Forest and MLP Classifier Models for each type of XAI, namely fidelity, unambiguity, and interpretable size. From the analysis, it is identified that, the Anchor model provides higher evaluation result when compared to the other XAI models.

Table 5. XAI based Evaluation Metrics for XGBoost Classifier Model.

Model	Fidelity	Unambiguity	Interpretability Size
SHAP	0.9993	0.238564	9.0
LIME	1.0	1.0	10.0
Anchor	1.0	1.0	2.0
Surrogate	0.9997	0.9436	4.2500
EBM	0.9985	0.4505	19.0000

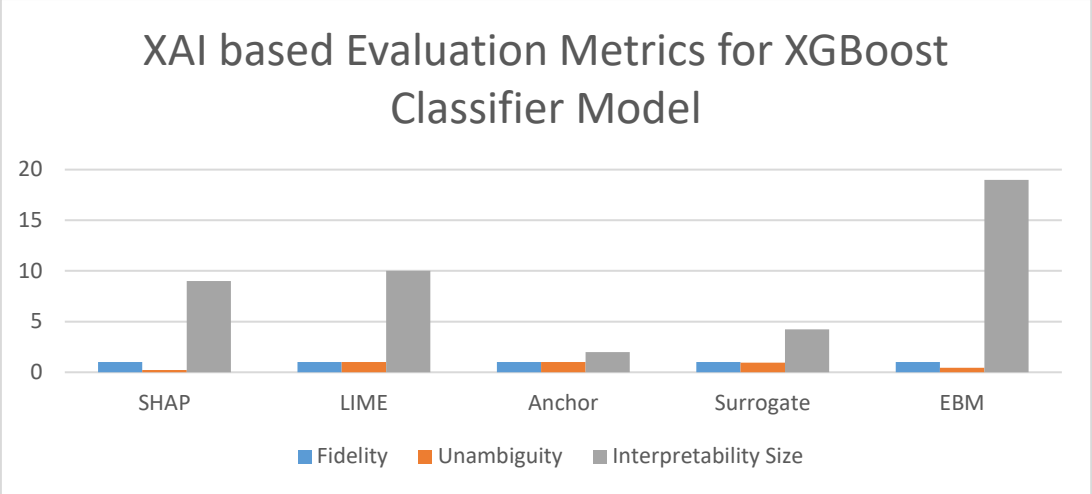


Figure 23. XAI based Evaluation Metrics Analysis for XGBoost Classifier Model.

Table 6. XAI based Evaluation Metrics for Random Forest Model.

Model	Fidelity	Unambiguity	Interpretability Size
SHAP	1.0	0.9884	10.0
LIME	1.0	1.0	10.0
Anchor	0.9966	1.0	2.0
Surrogate	0.9992	0.9437	4.7500
EBM	0.9996	0.4505	19.0000

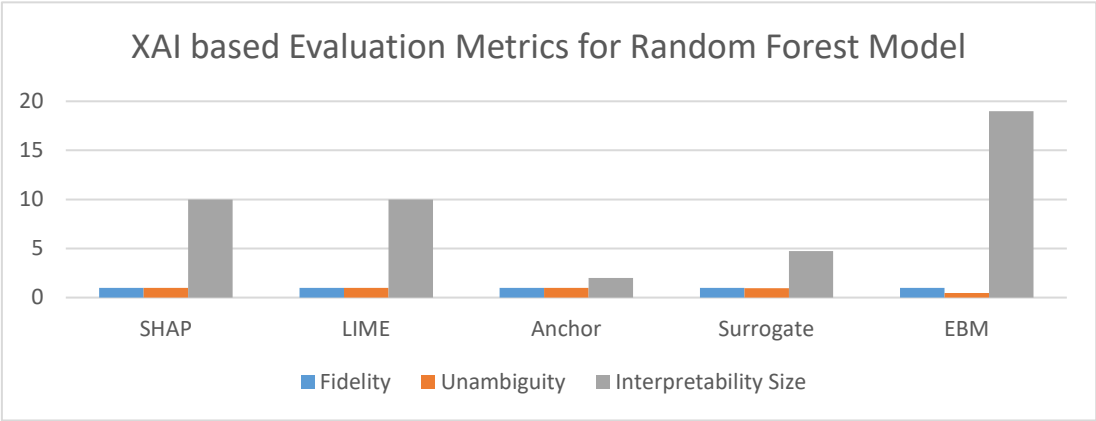


Figure 24. XAI based Evaluation Metrics Analysis for Random Forest Model.

Table 7. XAI based Evaluation Metrics for MLP Classifier Model

Model	Fidelity	Unambiguity	Interpretability Size
SHAP	1.0	0.9781	7.0233
LIME	1.0	1.0	10.0
Anchor	0.9955	1.0	1.0
Surrogate	0.9886	0.9436	6.7500
EBM	0.9852	0.4505	19.0000

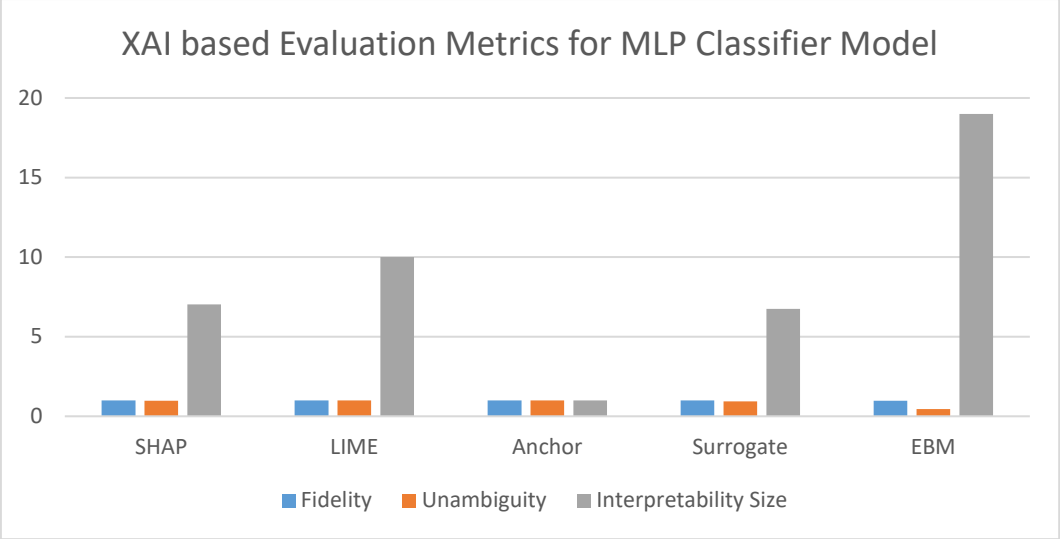


Figure 25. XAI based Evaluation Metrics Analysis for MLP Classifier Model.

8. Results and Discussion

From the extensive experimentations, the following inferences have been made: (i) SHAP provides a fine-grained feature importance analysis based on the prediction level, (ii) Anchor Explanations provide understandable rule-based insights, and Global Surrogates simplify the model behavior at a global level, and EBM balances accuracy with transparency. (iii) From the analysis, it is identified that, the Anchor model provides higher evaluation result when compared to the other XAI models.

8.1. Limitations of the proposed work

As the proposed work is based on PaySim dataset [21], it may suffer from synthetic bias and generalization gap and hence may lead to non-generalization of the model to exact replication of operational financial environments. The proposed work has applied both original and tuned ML models for prediction and so are limited to those models based trustworthiness assessment using XAI methods.

8.1.2. Advantages of the Proposed Work

The PaySim dataset [21] combines the accessibility of synthetic data with behavioral aspects of mobile based financial transaction systems. Due to its large scale and well defined features, it can be easily reproducible as an effective testbed for fraud detection, financial risk modeling and XAI based research to assess the fidelity of the models.

In addition, the application of XAI models for trustworthiness assessment is well addressed in this research work so that, the researchers can use this as a basis to apply other XAI methods for much more insight.

9. Conclusion and Future Work

9.1. Conclusion

We have focused, in this paper on the validation of the trustworthiness of machine learning models for the detection of mobile payment fraud since the tuned Random Forest model with a performance score of 0.9982 provided the best accuracy. While testing the XAI models, excellence of different methods across different metrics was observed. For fidelity, which reflects the alignment of explanations with the original model's predictions, LIME and the surrogate model performed best. The most unambiguous explanation has been given by LIME and Anchor to understand the insights with the minimum level of ambiguity. Interpretability size favors SHAP and EBM, as it offers a very detailed explanation that is helpful in deep analysis.

To ensure the reliability of our models in fraud detection, we tested them through a number of XAI methods to analyze the interpretability, transparency, and accuracy in depicting underlying decision processes. These validations have become essential in establishing trust, especially in high-stakes environments such as fraud detection where model decisions have financial implications and erode customer trust. The results are thus shown to be indicating the best model for short, concise, and transparent explanation for LIME and Anchor but more de-tailed insight with SHAP and EBM, so

stakeholders may opt for the most appropriate based on the needs for interpretability and the complexity of fraud pat-terns. Therefore, this study contributes toward the development of a reliable yet trustworthy framework for ML models to be used in mobile fraud detection, trading off performance with essential interpretability.

9.2. Future Work

Further expanding this work by including real-time model adaptation will further improve the detection ability of fraud in a dynamic environment by this system. Fraud patterns evolve quickly and hence the embedding of adaptive XAI mechanisms into the ML models would allow it for continuous learning. Therefore, adapting the models to keep being high in interpretability even while adapting to new threats would be achieved. Additionally, more testing would also be carried out to understand how these XAI models work when there are evolution cases in fraud patterns, but adapting to different geographical and demographic datasets.

Another promising direction to develop model transparency and reliability involves integrating user feedback. Gathering feedback from stakeholders such as the end-users, fraud analysts, and regulatory bodies could help improve XAI explanations to be more relevant and accessible to non-technical users. Such feedback may also be used for the improvement of model accuracy and interpretability metrics over time to gain trust and accountability.

The final study will involve the combination of mobile fraud detection specific to trustworthiness metrics that addresses model biases, privacy concerns, and regulatory compliance. Such future work would therefore contribute to developing not just high-performance but also ethical and resilient fraud detection solutions to meet the emerging requirements of financial and privacy standards in mobile-based business transactions.

Author Contributions
D.Jeya Mala - Project Mentor, Research work Inception, Problem Analysis, Modelling, Application of ML and XAI models, Results Verification, Paper Review, Evaluation and Submission. Dev Krishna Jhawar - Implementation of ML and XAI models, Analysis of Results, Paper Drafting. Chaitanya Bhude - Implementation of XAI models, Comparative Analysis of the results and Paper Drafting. All authors have read and agreed to the published version of the manuscript.

Funding
This research work didn't have any funding support from any organizations.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Data Availability Statement
Data and Material will be provided on request to the authors.

Acknowledgment
We the authors sincerely acknowledge Vellore Institute of Technology, Chennai, India for providing all the support in executing and completing this research work with all the resources.

List of Abbreviations

1. ML – Machine Learning
2. AI – Artificial Intelligence
3. XAI – Explainable Artificial Intelligence
4. RF – Random Forest
5. MLP – Multi Layer Perceptron
6. LIME – Local Interpretable Model-Agnostic Explanation
7. SHAP – Shapley Additive Explanations
8. EBM – Explainable Boosting Machine
9. TP – True Positive
10. TN – True Negative
11. FP – False Positive
12. FN – False Negative

References

1. Choi, D., & Lee, K. (2018). An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation. Security and Communication Networks, 2018, 5483472. https://doi.org/10.1155/2018/5483472
2. Delecourt, S., & Guo, L. (2019). Building a Robust Mobile Payment Fraud Detection System with Adversarial Examples. 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). doi:10.1109/AIKE.2019.00026
3. Jabeen, U., Singh, K., & Vats, S. (2023). Credit Card Fraud Detection Scheme Using Machine Learning and Synthetic Minority Oversampling Technique (SMOTE). 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA). doi:10.1109/ICIRCA57980.2023.10220646
4. Zhang, Y., Tong, J., Wang, Z., & Gao, F. (2020). Customer Transaction Fraud Detection Using XGBoost Model. 2020 International Conference on Computer Engineering and Application (ICCEA). doi:10.1109/ICCEA50009.2020.00122
5. Afriyie, J., Tawiah, K., Pels, W., Addai-Henne, S., Dwamena, H., Emmanuel, O., Ayeh, S., & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163. doi:10.1016/j.dajour.2023.100163
6. Vihurskyi, B. (2024). Credit Card Fraud Detection with XAI: Improving Interpretability and Trust. 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). doi:10.1109/ICDCECE60827.2024.10548159
7. Kotrachai, C., Chanruangrat, P., Thaipisutikul, T., Kusakunniran, W., Hsu, W.-C., & Sun, Y.-C. (2023). Explainable AI supported Evaluation and Comparison on Credit Card Fraud Detection Models. 2023 7th International Conference on Information Technology (InCIT). doi:10.1109/InCIT60207.2023.10413100
8. Embarak, O. (2023). Decoding the Black Box: A Comprehensive Review of Explainable Artificial Intelligence. 2023 9th International Conference on Information Technology Trends (ITT). doi:10.1109/ITT59889.2023.10184238
9. Mishra, R. K., Srivastava, S., Singh, S., & Upadhyay, M. K. (2024). Exploring the Opportunities of AI Integral with DL and ML Models in Financial and Accounting Systems. 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). doi:10.1109/ICACITE60783.2024.10616847
10. Muqattash, R., & Kharbat, F. (2023). Detecting Mobile Payment Fraud: Leveraging Machine Learning for Rapid Analysis. 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/SNAMS60348.2023.10375448
11. Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. Information Systems Frontiers, 25(1985–2003). doi:10.1007/s10796-023-10387-y
12. Thar, C. S. K., & Wai, A. Y. S. (2024). Digital Banking Fraud Detection Using Predictive Modeling Techniques. 2024 Fourth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT). doi:10.1109/ICAECT54875.2024.10564399
13. Awosika, A., Samuel, E., & Chukwuma, I. (2024). Explainable AI and Federated Learning: Future Directions in Financial Fraud Detection. 2024 Second International Conference on Data Science and Computational Intelligence (DSCI). doi:10.1109/DSCI60692.2024.10564667
14. Nijwala, A., Patel, R., Singh, J., & Gopalan, P. (2023). Application of Extreme Gradient Boost (XGBoost) Classifier for Credit Card Fraud Detection. 2023 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT). doi:10.1109/AIDA-AT58256.2023.00023
15. Raiter, A. (2021). Supervised Machine Learning for Anti-Money Laundering: A Credit Card Fraud Detection Study. 2021 IEEE International Conference on Big Data (Big Data). doi:10.1109/BigData52589.2021.9671414
16. Bello, Oluwabusayo & Folorunso, Adebola & Ejiofor, Oluomachi & Budale, Folake & Adebayo, Kayode & Olayemi, Alex & Babatunde, & Bello, Citation. (2023). Machine Learning Approaches for Enhancing Fraud Prevention in Financial Transactions. 85-108. doi:10.37745/ijmt.2013/vol10n185109
17. Hernandez Aros, L., Bustamante Molano, L.X., Gutierrez-Portela, F. et al. “Financial fraud detection through the application of machine learning techniques: a literature review”, Humanit Soc Sci Commun 11, 1130 (2024). https://doi.org/10.1057/s41599-024-03606-0
18. A. Jayanthkumar et al., "Fraud Detection in Financial Transactions using Machine Learning - A Comparative Study," 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), 2024, pp. 1-4, doi: 10.1109/DELCON64804.2024.10866169.
19. Yisong Chen, Chuqing Zhao, Yixin Xu, Chuanhao Nie, Yixin Zhang, “Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications”, Data Science and Management, 2025, https://doi.org/10.1016/j.dsm.2025.08.002. (In Press).
20. https://github.com/jeyamala/XAI-Finance-Fraud-Detection-Proj
21. PaySim dataset - https://www.kaggle.com/datasets/ealaxi/paysim1