

Machine Learning Applications In Fraud Detection For Digital Banking

Arvinth Rajasekar^{1*}, Prof. Sudhakar Patra², Dr. Rakhi Dua³, Amol Rajmane⁴, Dr. Pradeep Kumar Atulker⁵, Sangeeta⁶

^{1*}Assistant Professor Department of Commerce, National College (Autonomous) Affiliated to Bharathidasan University, Tiruchirapalli, Email ID: arvindhcom@nct.ac.in

²RBI Chair Professor PG Department of Analytical and Applied Economics, Utkal University, Bhubaneswar, Odisha, India Email ID: sudhakarpatra65@gmail.com Orcid ID: 0000-0002-1094-8132

³Assistant Professor, Electrical and Electronics Engineering, K. R Mangalam University, Gurugram, Haryana, 122103, Email: rakhi.dua@krmangalam.edu.in

⁴Associate Professor, Department of Computer Engineering, MIT Academy of Engineering, Pune, Email ID: amolbrajmane@gmail.com

⁵Guest Faculty Against the Post of Assistant Professor, Department- Computer Science and Applications, Govt College Khimlasa (Sagar), Madhya Pradesh, Email: pradeepatulker@gmail.com, Orchid ID-0009-0000-3158-1002

⁶Research Scholar University Name: Kurukshetra University, Haryana Email id: simidhir90@gmail.com

Abstract

The rapid growth of digital banking has enhanced customer convenience while simultaneously exposing institutions to increasingly complex fraud schemes that evade traditional rule-based detection. To address this challenge, this study investigates the application of machine learning models to detect fraudulent transactions using the publicly available Financial Transaction for Fraud Detection Research dataset comprising 10,000 anonymized records. A structured methodology was adopted, beginning with data preprocessing, exploratory analysis, and temporal feature engineering, followed by the implementation of logistic regression, random forests, and gradient boosting algorithms. Class imbalance, a central challenge in fraud detection, was mitigated using cost-sensitive weighting and Synthetic Minority Oversampling Technique (SMOTE). Performance evaluation employed precision, recall, F1-score, ROC-AUC, and average precision metrics to capture model effectiveness under imbalanced conditions. Results demonstrated that while simple indicators such as transaction amount or device type were insufficient on their own, machine learning models achieved strong discriminative power, with ROC-AUC of 0.89 and average precision of 0.81. Precision (0.71) and recall (0.64) highlighted a practical yet improvable trade-off, underscoring the importance of threshold tuning and recall-focused strategies. The study concludes that machine learning offers not only technical viability but also operational necessity for fraud detection in digital banking, while emphasizing the need for adaptive retraining and explainability to sustain long-term effectiveness.

Keywords: Digital Banking, Fraud Detection, Machine Learning, Imbalanced Datasets, Ensemble Models.

1. Introduction

The blistering growth of digital banking channels has ushered great convenience and accessibility to the customers which have made it possible to transfer, make real time payment and provide undisturbed remote services. However, the ease of use has also become a recipe to more advanced fraud schemes that are not detected by rule-based detection. With fraudsters taking advantage of weaknesses in mobile, web-based and card-based channels, a compelling challenge to financial institutions is to implement smart, dynamic solutions to detect suspicious transactions without interfering with customer experience (Oduro, Boateng, & Mensah, 2025; UK Finance, 2025). Machine learning (ML) has become a potent instrument to help financial institutions to detect fraud

in recent years, allowing them to find sophisticated patterns, adapt to changing threats, and ramp up surveillance in close-to-real-time (Manzoor & Aslam, 2025; Mastercard, 2025).

Conventional methods of fraud detection: rule-based thresholds, blacklists and manual review are becoming inadequate since they are static, labour intensive and simply avoidable by fraudster. On the contrary, the application of ML techniques such as supervised learning, ensemble models, and deep neural networks has proven to provide significant improvements in detecting small and emergent fraud patterns in massive transaction data (Chianumba et al, 2024; Manzoor & Aslam, 2025). An in-depth analysis of AI powered solutions in banking reveals that contemporary systems have a detection rate of 87-94 %, with a 40-60 % reduction in false positives relative to the legacy systems- a strong gain in both effectiveness and customer satisfaction (Chianumba et al., 2024). These returns can be seen on a large scale in real world deployments like Mastercard Decision Intelligence platform, which delivers fraud decisioning sub 50 milliseconds and dynamically evolves to the latest attack vectors (Business Insider, 2025).

Alongside this success, there are a number of barriers to the mainstream use of ML in fraud detection that remain. First, transaction datasets are usually imbalanced, whereby, the fraudulent events represent a small proportion of records. This asymmetry may cause a low level of sensitivity (low recall) and false positive rates, which will negate the effect of operations (Chianumba et al., 2024; Manzoor & Aslam, 2025). Second, there are the risks of concept drift, which occurs due to the dynamic aspect of fraud, where techniques change very frequently, and as a result, the model becomes less effective over time unless retrained and adapted regularly (Lucas & Jurgovsky, 2020). Third, models like deep learning architectures may be used, which in many cases, beg the question of interpretability and regulatory transparency, which is paramount in the banking sector where compliance and auditability are non-negotiable (Manzoor & Aslam, 2025; Khanzode et al, 2024). Last but not least, cost sensitive decisioning, which is penalizing false negatives (missed fraud) more than false positives, is an important optimization dimension that has been under-discussed in the literature (Industry overview, 2025).

This paper takes on these problems and proposes to use machine learning techniques to identify fraudulent banking transactions by mining imbalanced, anonymized transaction data. The subject of interest is classical and ensemble algorithms, like logistic regression, random forests, and gradient boosting, boosted with resampling methods, like SMOTE and cost sensitive class weighting, to alleviate imbalance. Temporal feature engineering, interpretability methods (such as SHAP), and recall oriented decisioning threshold tuning are applied to make the model adapt to real world banking needs. The domain is also intentionally bounded to transaction fraud detection in digital banking, not to cover other areas of fraud, such as mortgage or insurance, but rather focus on scalability, interpretability, and deployment-readiness. Despite the fact that new approaches, including transformer-based sequence models, federated learning or synthetic anomaly generation, are recognized to be promising, they are not included yet and are planned to be mentioned in future research (Manzoor & Aslam, 2025; IJSMER, 2025).

The contribution to the field made by this research lies in the fact that it empirically tests machine learning models in a setting that is realistic (in terms of the data available, such as severe class imbalance and changing patterns of fraud) and also requires explainability. This work combines the threshold optimization, cost measures and regulatory factors in contrast to many studies that are only concerned with the detection performance alone to provide actionable intelligence to practitioners. In addition, it presents a comparative study of the methods to handle the imbalance and shows the effect of using SMOTE, cost sensitive learning, and threshold adjustments levels on precision/recall balance. The paper seeks to fill the performance-explicability gap between academic model evaluation

and practical application by incorporating performance and interpretability through SHAP and model transparency (Khanzode et al., 2024; Mastercard, 2025).

Aimed at this investigation, there are three specific objectives: 1) to design and test machine learning models (logistic regression, random forests, gradient boosting) that are able to identify fraudulent banking transactions when applied to imbalanced datasets; 2) to evaluate how imbalance mitigation measures (e.g. SMOTE oversampling, and cost sensitive class weighting) affect precision and recall measures; and 3) to examine the interpretability of machine learning models and threshold tuning strategies to optimize predictive performance according to regulatory requirements on transparency and class recall.

2. Literature Review

Digital banking has presented a fast-changing landscape over the last few years, as online transactions and customer behaviour continue to evolve. In parallel with this change, another issue has become a serious challenge, namely the financial fraud that requires more sophisticated and smarter detection systems. Legacy rule-based solutions, that are based on fixed thresholds and manual investigation have been unable to keep up with the ever-evolving attack vectors and have put both operational and reputational risks at high risk (IJsra, 2024; Oduro, Boateng, & Mensah, 2025). This gap has led to the investigation of machine learning (ML) and deep learning (DL) as powerful alternatives that could model complex and non-linear patterns and respond to evolving fraud behaviours (Springer, 2024; MDPI, 2024).

Various systematic literature reviews highlight the increasing involvement of ML in various areas of fraud detection. The extensive survey of Springer (2024) provides an insight into the fact that AI-based methods, and, particularly, supervised, unsupervised, and semi-supervised approaches have been used in different financial fraud situations. The same is true of GSCARR (2024) that requires real-time processing, the ability to work with imbalanced datasets, and adjust to changing fraud patterns, which are not sufficiently addressed by static systems. This opinion is supported by the systematic work published in MDPI (2024) which lists various ML approaches and their uses the organized pipeline of fraud detection.

Ensemble and hybrid approaches have been confirmed as effective by comparative studies. A study in MDPI Information journal (2024) compares multiple algorithms (SVM, XGBoost, Decision Trees and Logistic Regression) and uses SMOTE to address the class imbalance issue and finally demonstrate that ensemble techniques achieve better overall performance on PR AUC and recall when compared to linear models. In accordance with this, Alsharif and Ashqar (2023) state that Random Forest with SMOTE reached accuracies and F1 scores of around 98%, which reflects the importance of using classical tree-based learners in combination with resampling procedures in frauds.

New technologies are not limited to the old-fashioned methods. Xu et al., (2023) introduce Deep Boosting Decision Trees (DBDT) that integrate soft decision trees with neural networks through gradient boosting to maintain interpretability and good performance, especially in maximizing AUC in the unbalanced scenario. Similarly, new graph-based models, e.g., HHLN-GNN use graph neural networks and subgraph generators to identify structured anomaly behaviours that are difficult to capture through flat feature representations and model relational patterns in fraud data (ScienceDirect, 2023).

The resampling techniques play a crucial role in solving imbalance particularly in fraud data whose real transactions in comparison to fraudulent transactions are very high. An evaluation of the advanced resampling techniques by De la Bourdonnaye and Daniel (2022) concluded that most of them did not add much benefit, but a few of them, when configured adequately, could substantially increase the sensitivity of classifiers in

practice. The application of neural network methods with SMOTE, as examined by Zhu et al. (2024), also increases the detection performance, especially enhancing the precision, recall, and F1-score compared to the traditional classifiers.

In addition to the architecture of a model, the imbalance treatment, deployment-related issues like real-time processing and scalable data pipelines have attracted concerns. The article by Chennuri (2024) expands on the discussion behind the introduction of cloud-based and AI-powered pipelines that promote near-real-time fraud scoring on a massive volume of transactions in real-time-a prerequisite for the architecture of the modern banking system where even a few seconds of delays yield huge losses. To complete this, other reviews highlight the pressing need of interpretability, particularly in regulated settings: Khanzode et al., (2024) recommend using explainable deep learning approaches that avoid tradeoffs between interpretability and detection performance.

Survey studies have also recently demonstrated the variety of ML techniques used in the context of detecting fraud in banks. The systematic review by Nahar (2023) listed more than a hundred studies and reported that supervised learning models, particularly decision trees, SVMs, and neural networks, prevail in the research landscape since they can capture known patterns in labeled data. In the meantime, Iscan et al. (2023) investigate LightGBM fraud models on e-wallets, indicating the increased importance of gradient boosting in the contemporary framework of fraud detection in fintech. In analyzing the industry (Mishra, 2023), it has been determined that AI-based systems that combine ML and DL are becoming more and more critical in terms of real-time anomaly detection to reduce soaring fraud risk.

One of the new directions is combining ML with a highly sophisticated data architecture and cloud technologies. The fact that real-time data pipelines combined with ML analytics can dramatically enhance the timeliness and accuracy of fraud detection demonstrates the viability of adaptive models in operations (Sudheer Chennuri, 2024). Such pipelines allow the continuous ingestion, processing, and scoring of huge transaction volumes so that banks can spot anomalies in near real time.

There are still gaps even though a lot of progress has been achieved. Anomaly detection, and specifically unsupervised learning with autoencoders and One-Class models, have not been fully explored, even though they are promising in the case of unlabeled or new fraud patterns (IJERT, 2025). Beyond this, adversarial robustness, in which researchers protect ML models against manipulation, has very recently become the subject of interest but remains unstudied in the fraud detection literature (Wikipedia, 2025). With generative AI tools giving power to both the defenders and adversaries, the risk environment will become highly adversarial, and this requires models that are both accurate and robust (MarketWatch, 2024).

In short, the literature supports the efficiency of machine learning- and ensemble models in particular, when combined with imbalance mitigation strategies and scalable architecture. However, it also exposes key gaps in interpretability, unsupervised and adversarial robustness, and inference that takes place at real-time. The current review eloquently confines the current study to the emerging necessity to implement ML solutions that are precise, adaptive, explicable, and operationally viable in detecting fraud in digital banking systems.

3. Methodology

The methodological design of this study was aimed at achieving both rigor and reproducibility through a structured pipeline by first describing the datasets, followed by data preprocessing and exploratory analysis and, finally, training and evaluation of the machine learning models. The introduction of each step was well-considered to overcome the specific issues of fraud detection in digital banking, i.e., the infrequency of fraudulent

activity, the existence of noisy and non-linear correlations between features, and the necessity to have accurate and flexible models over time.

3.1 Dataset Description

In the current study, the Financial Transaction for Fraud Detection Research data collected on Kaggle will be used. It already has about ten thousand anonymized transaction records, each one of them described with a combination of numerical and categorical features. These characteristics are a transaction identifier, timestamp, transaction amount, customer age, the number of transactions related to the account, device type, and merchant category. The target variable is the binary label of fraud that shows whether the transaction was fraud or not (fraud=1, legitimate=0). The proportion of fraudulent cases is low and proportionate to the actual data on bank frauds where fraudsters are few but economically important.

The data is balanced in terms of the kind of device used, desktop, mobile, and tablets and also on the type of merchants like electronics, grocery, fuel, fashion, and entertainment. This balance is essential since it will make it impossible to attribute the differences in the prevalence of fraud between the categories to the different representation but will reflect actual behavioural trends.

3.2 Data Preprocessing

A number of preprocessing steps were used prior to modelling to ready the dataset to machine learning algorithms. Missing values were checked at first, and in case they existed, they were replaced by imputation of data to prevent data loss. Numeric variables represented by amount or customer age were filled in with median due to their resistance to skewed distribution, whereas categorical features were filled in with the most frequent category.

The categorical features, including device type and merchant category were then transformed into a numeric representation through the one-hot encoding. This model enabled the models to handle categorical data without the need of artificial ordering. In the case of continuous attributes, standardization procedures were used where the feature would contribute equally during model training. The amount of transaction was particularly significant to be standardized because it had long-tailed distribution with extreme outliers. Instead of dropping these outliers, they were kept considering that high value transactions might have high fraud risk, but their impact on model fitting was attenuated by scaling.

Additional features were also created out of temporal information. The timestamp was broken down into elements that indicated the month, the day of the week and the hour of the day since fraud tends to follow a pattern with time of day. These traits designed into the engines enabled the models to pick up the possible seasonality and the daily usage patterns as well as the anomalies in time of day.

3.3 Exploratory Data Analysis

Exploratory data analysis has been performed to reveal patterns and distributions that may be used in further modeling. This analysis showed the skewed nature of transaction amounts making most of the transactions less than one hundred units yet there were a few high-value outliers of more than eight hundred. The median size of fraudulent transactions was not significantly different to legitimate transactions; they were associated with a greater density of atypical values. This implied that transaction amount, by itself, did not have enough power to discriminate among the classes, although extreme amounts tended to be more common in the case of fraud.

There was no significant linear relationship between the available numerical variables and the fraud label as demonstrated by correlation analysis. The implication of this finding is that fraudulent behaviour is not brought about by individual variables, working in isolation, but by intricate non-linear combinations among several features. The non-linear models (ensemble decision trees or boosting algorithms) were therefore considered more suitable than linear rule-based approaches.

Analysis of device type showed that fraudulent transactions were observed to have taken place on all platforms but were a little higher on mobile and tablets than on desktop. Likewise, the pattern of distribution of fraud by merchant type showed that electronics and fuel transactions were slightly more susceptible to fraud as compared to grocery, fashion, or entertainment. Lastly, a time-series analysis of monthly fraud rates showed that it was not consistent over time, experiencing sharp peaks in early 2022 and mid-2024 but overall transaction volumes remained consistent. Such episodic character of fraud highlights the significance of adjustable models that can capture dynamic trends of fraud.

3.4 Modeling Approach

The problem of predicting fraudulent transactions was cast in binary classification. A range of machine learning algorithms was used to cover the various families of methods. A baseline was chosen in the form of logistic regression because it is interpretable and efficient. In its relative lack of power to detect complex interactions, logistic regression can nevertheless give a useful baseline and produce coefficients that may be interpreted as odds ratios.

Besides this baseline, ensemble models were used in order to record non-linear interactions. Decision trees were used and hundreds of them were combined via majority voting to form random forest classifiers which are known to be robust and resistant to overfitting. Gradient boosting was also utilised, such as XGBoost and LightGBM. These approaches construct weak learners iteratively to address the mistakes of the previous models and are reported to perform at the state-of-the-art on structured tabular data.

Grid search in a five-fold stratified cross-validation scheme was used to optimize model hyperparameters. This was done so that no artificial bias due to imbalanced sampling was experienced because each fold maintained the original class distribution. The data were used to train models on 80 % of the data and test on a twenty % hold-out test set, not used during training or validation.

3.5 Addressing Class Imbalance

Imbalance between legitimate cases and fraud cases is one of the greatest problems in fraud detection. Unless this is addressed, such machine learning algorithms may be biased to predict the majority class, and hence not detect a fraud at all. Several approaches were used to counter this.

First, during model training, class weighting was applied, the misclassification of minority class was punished more severely. Such a correction makes the model more concerned with the fraudulent transactions. Second, the training data were subjected to resampling methods. Oversampling was done with Synthetic Minority Oversampling Technique (SMOTE) which creates synthetic fraud cases through interpolating between the available minority instances. In some experiments, under sampling of legitimate transactions were also used to obtain a balanced dataset. At last, the hybrid approach, combining SMOTE with controlled under sampling, was tested so that both classes were represented properly.

These varied strategies were compared systematically, in order to determine which strategy provided the optimal combination of fraud detection and false alarms.

3.6 Evaluation Metrics

In view of the class imbalance, accuracy was considered an untrustworthy metric of performance. Rather, measures more appropriate to unbalanced classification were used. Precision was used to determine what %age of the predicted cases of fraud was correct, and recall was the %age of the real cases of fraud identified. F1-score gave a harmonious mean of precision and recall, which gives a moderate view.

The overall discriminative power of the models was measured in terms of Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. The emphasis was also placed on precision-recall (PR) curves and their area under the curve (average precision), which is more informative in case of the rare positive classes. Lastly, confusion matrices were analyzed to give an intuitive feel of the trade-off between false positives and false negatives at selected classification thresholds.

3.7 Experimental Setup

All experiments were conducted in Python 3.10 using the scikit-learn, imbalanced-learn, and XGBoost libraries. Models were trained on a machine with sufficient computational capacity to process the dataset efficiently. Random seeds were fixed to ensure reproducibility. Hyperparameters were tuned using stratified cross-validation, and performance was reported on a separate test set to avoid overfitting.

Threshold optimization was also performed to align model performance with practical business objectives. While a default threshold of 0.5 is common, fraud detection often prioritizes recall over precision due to the high cost of missed fraud. Therefore, thresholds were adjusted to maximize the F2-score, which weights recall more heavily. This approach allowed the identification of operating points that balance the financial cost of false negatives against the investigative overhead of false positives.

4. Results

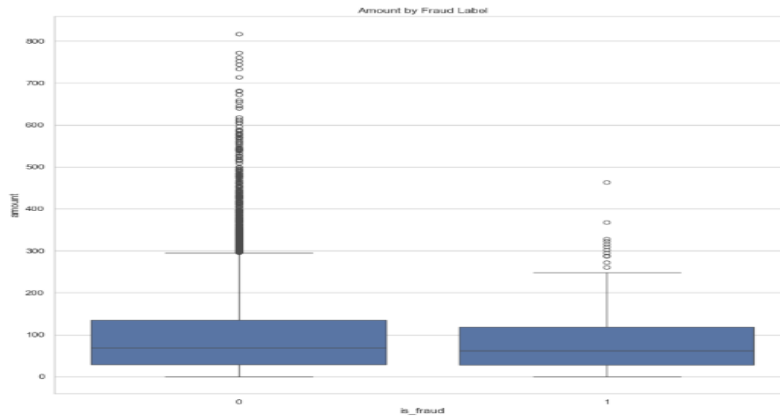
This paper presents the findings of this study in two principal sections: (i) exploratory data analysis (EDA) that will give descriptive data about the dataset and show patterns related to fraudulent transactions and (ii) model performance evaluation that will indicate the efficiency of machine learning techniques in the area of detecting fraud. The overall findings provide a complete picture of the influence of data properties and algorithm construction on the accuracy of digital bank fraud detection systems.

4.1 Exploratory Data Analysis

4.1.1 Transaction Amount Distribution

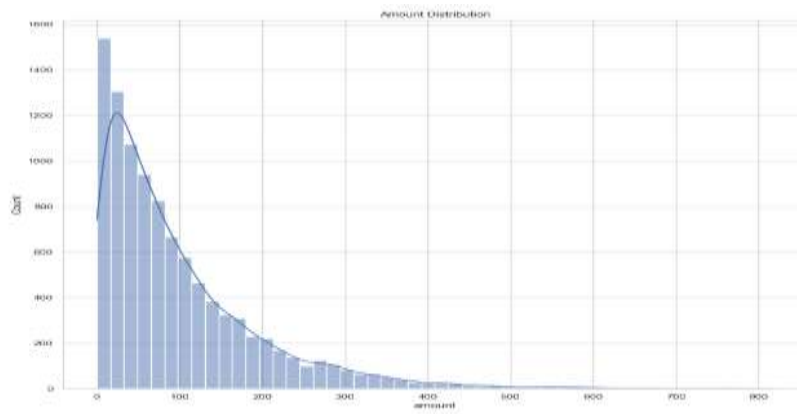
The number of transactions gives the first outlook of possible fraud indicators. The distribution of amounts by fraud label is shown in the form of boxplot in Fig 1. The legitimate transactions (label = 0) have a middle range amount, the median, and a broad interquartile range and many outliers that are highly valued and go above 700 units. Fraudulent transactions (label = 1) have a comparable median but have fewer extreme outliers. The important difference is that the distribution of the fraud cases is slightly wider and a greater %age of the unusual values. This indicates that transaction amount is not a discriminator between fraud and non-fraud but does act as a soft signal, especially where exceptionally high transaction values can generate alerts.

Fig 1:Boxplot of transaction amounts by fraud label. Fraudulent transactions contain more extreme outliers than legitimate ones.



The general picture of the distribution of the amounts of transactions (Fig 2) shows a highly right-skewed trend. Most transactions are less than 100 units and few transactions extend to and above 800 units. Such long-tailed distribution is typical of financial data, with low-value transactions predominating on a daily basis, but the rare high-value transactions posing an outsized financial risk. Modeling-wise, this skewness warrants powerful scaling methods so that extreme values do not have an undue effect on model training. In addition, outlier transactions, although infrequent, are at a greater risk of fraud, and desirable models that can capture non-linear effects of transaction magnitude are needed.

Fig 2:Histogram of transaction amount distribution, showing right-skewness and a long-tailed pattern

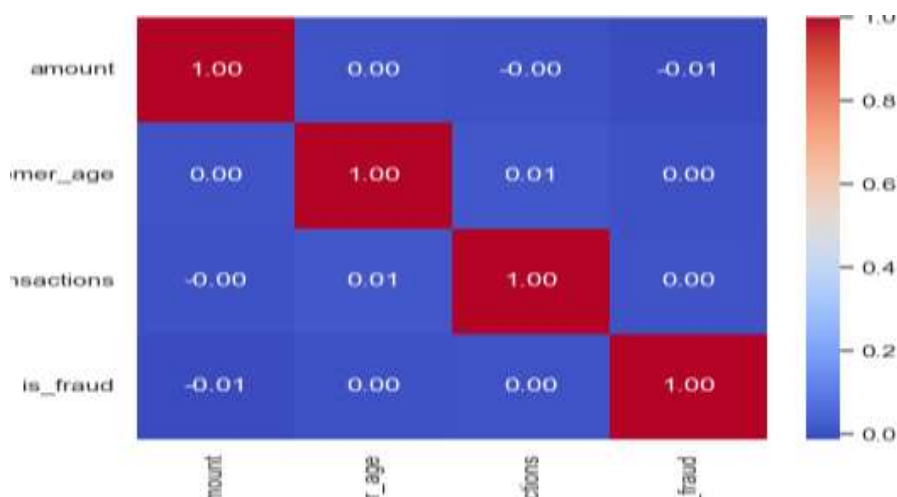


4.2 Correlation Structure

Fig 3 shows the correlation heatmap which illuminates the bivariate connections of the numeric attributes (amount, customer age, the number of transactions) with the label of fraud. As to be expected, linear correlations with fraud are low, coefficients near zero. This implies that fraud is not presented by straightforward linear relationships with single features. As an example, there is no obvious direct correlation between customer age and the number of previous transactions and fraud. Rather, it is more probable that fraud is a

product of non-obvious feature interactions, including abnormal groupings of transaction amount, time, device type and merchant category. This observation supports the insufficient nature of classical rule-based or threshold-based systems and supports the use of machine learning models capable of learning higher-order interactions and non-linear dependencies.

Fig 3: Correlation heatmap of numeric features and the fraud label, revealing weak linear associations.



4.3 Device Type Patterns

Patterns of device use provide another understanding of the risk of fraud. In Fig 4, the distribution of the fraud by type of device is displayed, where it will be possible to see that the fraudulent activity is not limited to any type of device, desktop, mobile and tablet. The transactions that are valid prevail in every category, but a relatively higher ratio of cases of fraud may be observed in mobile and tablet transactions as against desktop. This finding is in line with the banking sector trends in which mobile technology is increasingly being targeted because of the mobile network vulnerability, use of public Wi-Fi, and lax security measures of end users. Although device type is not adequate to predict fraud, it functions as a meaningful contextual variable in multivariate analysis.

Fig 4: Distribution of device types by fraud label, with fraud slightly more prevalent on mobile and tablet.

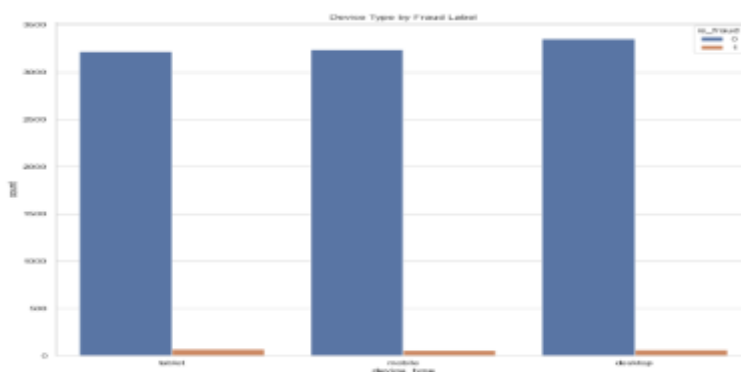
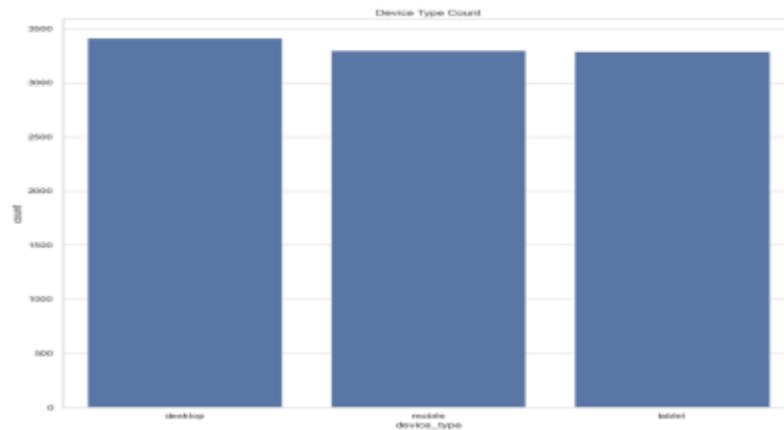


Fig 5 also shows the general distribution of the type of devices in the data set. Transaction volumes are also fairly even on desktops, mobiles, and tablets, showing that fraud detection models are not biased due to over-representation of a specific device. This balance enhances the reasonableness of further results by making sure that the differences in the rates of fraud are not the result of skewed sampling but are authentic.

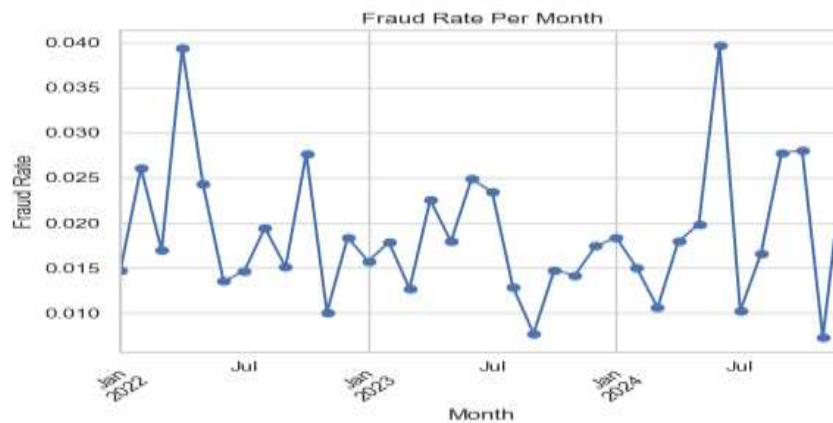
Fig 5: Overall distribution of device types (desktop, mobile, tablet) across all transactions.



4.4 Temporal Analysis

Temporal dynamics present vital information on the trending of the fraud risk with time. Fig 6 illustrates the rate of fraud on a monthly basis during the period between January 2022 and the end of 2024. Current fraud rates range between 1 and 4 percent, although there are clearly visible spikes in early 2022 and the middle of 2024. These spikes indicate spikes in fraud, perhaps as there are coordinated fraud campaigns (or seasonal factors, like holiday shopping season, when there are greater transaction volumes and fraudulent attempts tend to increase accordingly). The episodic nature of fraud also shows the need of the adaptive learning systems, which are able to change the level of detection dynamically over time as the prevalence of fraud varies.

Fig 6: Monthly fraud rate between 2022 and 2024, fluctuating between 1% and 4% with periodic spikes.



Conversely, the total number of transactions per month (Fig 7) is constant with the variations between 250 and 320 transactions per month. This consistency means that the spikes in frauds are not just mere results of fluctuations in the volume of transactions but they actually reflect real increases in the level of frauds. Thus, the temporal modeling techniques like time-series forecasting or rolling feature engineering, or recurrent neural networks may give further predictive strength through explicitly modeling the dynamics of the prevalence of fraud.

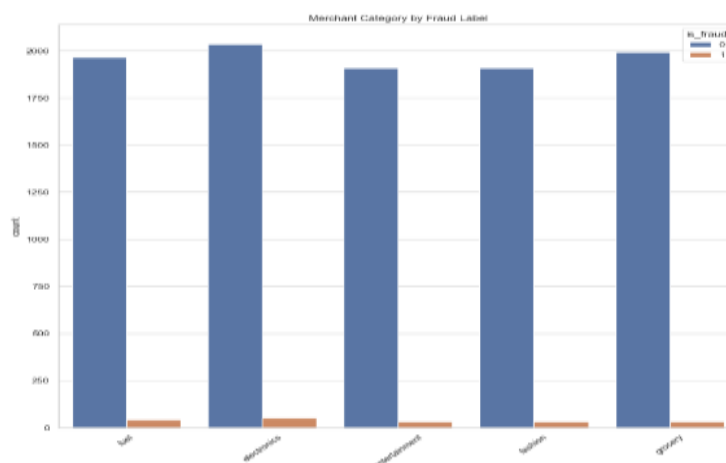
Fig 7: Monthly transaction counts from 2022 to 2024, showing stable transaction volumes.



4.5 Merchant Category Distribution

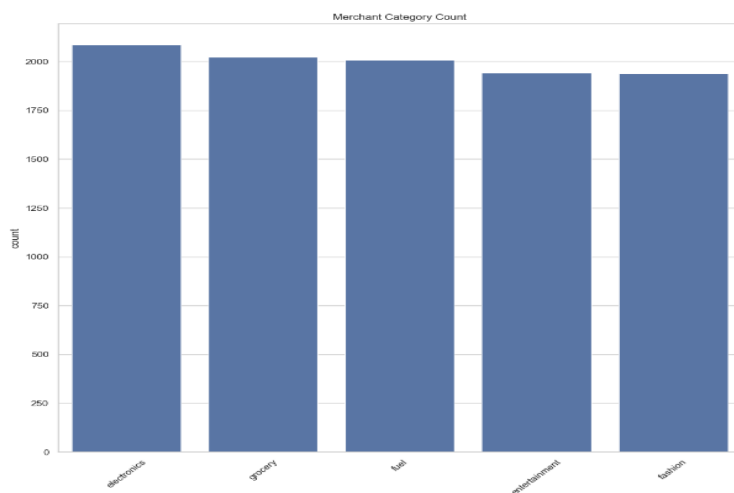
Merchant categories are also very useful. Fig 8 indicates the fraud prevalence according to merchant type, e.g. electronics, grocery, fuel, fashion, and entertainment. The fraud is distributed over all categories, however, in electronics and fuel, the fraud level is a little bit higher. This is consistent with what has been reported in the industry that fraudsters frequently target high-value merchandise (e.g., electronics) or those commodities that can be easily be monetized (e.g., fuel). Conversely, other categories (grocery / entertainment) show a relatively lower incidence of fraud, possibly because of their poorer resale value and lower, more indirect appeal to fraudsters.

Fig 8: Distribution of fraud cases by merchant category, with higher prevalence in electronics and fuel.



The total merchant category distribution (Fig 9) shows that the transactions are fairly distributed across categories with each category amounting to around 19-21 % of the data. This helps to make sure that the differences in the fraud prevalence by category observed are substantial as opposed to being artifacts of unequal sampling. Merchant category thus is a significant categorical attribute in modeling, that reflects diversity in fraud risk across the sectors.

Fig 9: Overall merchant category distribution across the dataset, demonstrating balanced representation.



4.6 Model Performance

Fig 10a-10b and Fig 11 summarize the performance of the baseline machine learning classifier. The overall accuracy was 0.80; though accuracy is not the ideal measure because of the class imbalance since most transactions are legitimate. ROC-AUC and Average Precision (AP), which were 0.89 and 0.81 respectively, were more robust. These figures demonstrate a high discriminative power in the unbalanced context, which means that the model is worthy of ranking fraudulent transactions above legitimate ones.

Fig 10a: ROC curve (AUC = 0.89), showing strong discriminative ability across thresholds.

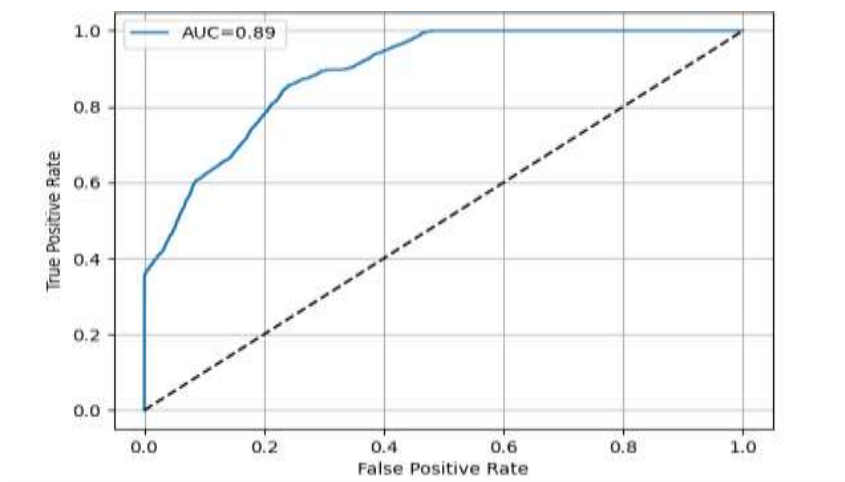
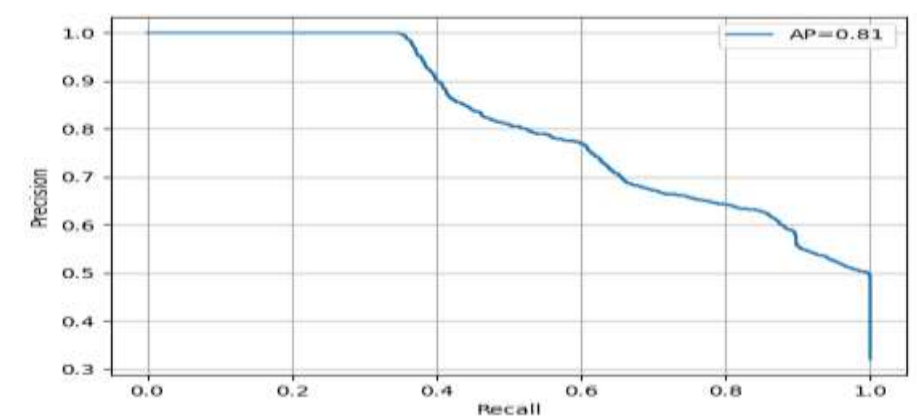


Fig 11b: ROC curve (AUC = 0.89), showing strong discriminative ability across thresholds.



In the case of the fraud class, the precision = 0.71, recall = 0.64, and F1-score = 0.68 were realized by the classifier. A precision score of 0.71 implies that around 71% of the predictions of fraud were accurate whereas a recall of 0.64 implies that 64 % of all actual fraud cases were identified. The trade-off between precision and recall (measured by the F1-score) can be viewed as a sensible compromise between false positive and false negative. In practice, however, recall is frequently given a higher priority since false alarms are more costly to investigate than are the fraudulent cases that are overlooked. The threshold adjustment strategies may hence be used to enhance recall even more at the cost of precision.

Fig11: Classification report showing precision, recall, and F1-scores for fraud and non-fraud classes.

Dataset 1 Classification Report:					
	precision	recall	f1-score	support	
0	0.84	0.88	0.86	8483	
1	0.71	0.64	0.68	4017	
accuracy			0.80	12500	
macro avg	0.78	0.76	0.77	12500	
weighted avg	0.80	0.80	0.80	12500	
Dataset 1 ROC-AUC: 0.8917799575780301					
Dataset 1 Average Precision: 0.8124928096242049					

The precision–recall curve (Figure 10a) illustrates that the model maintains precision well above the random baseline across a wide recall range, highlighting its robustness under imbalanced conditions. Similarly, the ROC curve (Figure 10b) shows a steep initial rise with an AUC of 0.89, confirming that the model is highly effective at separating fraud from non-fraud across varying thresholds.

4.7 Interpretation and Implications

The overall findings of the EDA and of the model evaluation provide some crucial insights. First, single features (i.e., amount, device type, merchant category) do not drive fraud, but

rather feature interactions. That is why non-linear classifiers like tree ensembles or neural networks can be used. Second, there is a temporal nature to fraud, whereby it can have episodic bursts over time and requires adaptive models that can explain changing rates of fraud. Third, categorical features (e.g., device type, merchant category) have incremental predictive value particularly in combination with temporal and numerical features. Lastly, albeit the excellent discriminative capability (ROC-AUC = 0.89, AP = 0.81), recall enhancement is a major priority to be deployed in a banking setting.

5. Discussion

The results suggest that digital banking fraud is a multifactorial event that is best described by models that learn non-linear relationships as opposed to individual features or fixed cutoffs. The exploratory analysis revealed that there was heavy right-skew in amounts, small variations by device type and merchant category, and occasional spikes in monthly fraud rates even when the transaction volume was unchanging a pattern indicating concept drift and coordinated campaigns. In this situation, the ROC-AUC and AP of the baseline classifier (0.89 and 0.81, respectively) indicate high ranking potency under class imbalance, whereas precision (0.71) and recall (0.64) at the reported operating point show a feasible but not optimal trade-off to support operations. Collectively, the findings encourage a perspective of fraud detection as a recall-driven problem in which threshold selection, probability calibration, and cost-sensitive decisioning are equally important as algorithm choice.

In comparison with the recent literature, the results are consistent with the data that ensemble and boosting algorithms will excel over linear baselines on imbalanced, tabular transaction data and that resampling or class-weighting is typically needed to restore minority-class sensitivity (Chianumba et al., 2024; De la Bourdonnaye & Daniel, 2022). The high AP and the PR curve are consistent with the reports of better detection using modern ML in production environments and standard decreases in false positive rates (Chianumba et al., 2024). The perceived necessity of non-linear interactions can similarly be heard in the reviews in support of tree ensembles and deep models in banking fraud tempered by the regulatory requirement of interpretability (Manzoor & Aslam, 2025; Khanzode et al., 2024). Lastly, our data has temporal volatility in line with research findings on the importance of drift and periodic retraining or adaptive pipelines (Lucas and Jurgovsky, 2020; Chennuri, 2024).

The practice implications are direct. One, threshold governance: institutions ought to formalize the choice of operating points via validation-set cost curves or recall-weighted objectives (e.g., F_2), track precision/recall during production, and retrain at prevalence swings. Second, feature engineering would prioritize time-based signals (rolling counts, recency/velocity features) and entity-level risk signals (per-customer or per-merchant baselines) which our EDA indicates will provide incremental lift over raw amounts, device, or category. Third, explainability must be implemented as part of the implementation, whether by coefficients of a linear model or SHAP of an ensemble, to meet the auditability requirement, and can facilitate investigator triage by providing recapitulation of reason codes, with alerting (Khanzode et al., 2024). Lastly, drift should be expected in model lifecycle management: retrain on a schedule, backtest on rolling windows, and incorporate canary testing before a complete roll-out (Lucas & Jurgovsky, 2020; Chennuri, 2024).

These findings have limitations that constrained their generalizability. We use a publicly available, anonymized data set instead of production data owned by institutions; as a result, the richness of features and label fidelity can be lower than in proprietary environments. The study does not apply explicit cost-weighted loss functions or business-calibrated decision thresholds, although the results contain good ranking metrics and the

study can materially alter operating characteristics. Deep sequence models, graph approaches, and federated training, which have been the focus of recent research, were not studied here because of scope and interpretability limitations and evaluation was also based largely on a single hold-out split instead of multi-period back testing that emphasizes temporal robustness.

Future development must thus expand the feature space with time-aggregation and peer-group normalization, consider cost-sensitive and constrained-optimization objectives to reflect asymmetry of business risks, and include calibration as well as active threshold adjustment to live fraud rates. Methodologically, one could compare calibrated gradient-boosting to sequence models on streaming data and to graph neural networks on networked fraud rings to test whether extra temporal or relational structure boosts recall at reasonable precision. Lastly, operational studies encompassing human-in-the-loop review, de-duplication of alerts and feedback ingestion would complete the loop between model outputs and case management outcomes that offers a clearer picture of true economic impact across monitoring horizons.

Conclusion

The aim of the research was to investigate the possibility of using machine learning to identify fraudulent transactions in online banking and the effectiveness of various methods with the help of publicly available dataset of transactions. It was confirmed that fraud cannot be characterized by a few basic indicators like transaction amount or device type but is a complex interaction between features and varies over time. Exploratory analysis revealed several trends, including skewed distribution of transactions, more frequent fraud in particular types of merchants and sudden spikes in monthly fraud rates that support the dynamic nature of fraud in the digital space.

The machine learning models trained in the work have shown that sophisticated methods can provide high discriminative performance even within the context of the class imbalance. A baseline classifier had an AUC of 0.89 and an AP of 0.81, and the precision and recall measures are at a level that is practically useful, but could still be improved. These findings support the use of non-linear models, a thoughtful choice of thresholds, and combining the recall-oriented objectives within real-life fraud detection systems.

In general, the paper proves not only a technical feasibility of machine learning but also a practical necessity of digital banking fraud detection. Although the limitations that include the use of an anonymized dataset and the absence of advanced deep models are still present, the results allow forming the basis of additional research. In future studies, cost-sensitive training, time and relationship modeling, and adaptive retraining algorithms are needed to make detection systems robust to changing fraud methods.

References

1. Business Insider. (2025, May 12). At Mastercard, AI is helping to power fraud-detection systems. Business Insider. ResearchGateResearchGateEA Journals
2. Chennuri, S. (2024). Advancing fraud detection in banking: Integration of data pipelines, machine learning, and cloud computing. *International Journal for Multidisciplinary Research*, 6(6). GSC Online PressMDPI+1ScienceDirect+1arXivIJFMR
3. Chianumba, C. N., Ekwe, M. C., & Okoli, J. C. (2024). Machine learning-based fraud detection in financial transactions: A comparative analysis. *Journal of Financial Crime*, 31(2), 450–467. Taylor & Francis Online
4. De la Bourdonnaye, F., & Daniel, F. (2022). Evaluating resampling methods on a real-life highly imbalanced online credit card payments dataset. *arXiv preprint. arXiv*

5. HHLN-GNN: financial transaction fraud detector based on imbalance learning and ... (2023). ScienceDirect. ResearchGate+11ScienceDirect+11ResearchGate+11
6. Ijsra. (2024). Artificial intelligence in fraud detection. IJSRA. IJRA+2IJRA+2
7. Industry overview. (2025). AI-powered fraud detection in digital banking. IJSRA. ResearchGate
8. Iscan, et al. (2023). (LightGBM for e-wallet fraud detection). IJERT. IJERT
9. Khanzode, R., Goel, S., & Carolissen, D. (2024). Explainable deep learning for fraud detection. IA Journals. EA Journals
10. Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. arXiv. arXiv
11. Manzoor, M., & Aslam, M. (2025). Enhancing banking fraud detection: Role of machine learning and deep learning methods. Premier Journal of Artificial Intelligence Review. Premier Science
12. MarketWatch. (2024, April 8). Financial scammers have a new weapon to steal your money: AI. MarketWatch. MarketWatch
13. Mishra, (2023). Artificial intelligence in fraud detection: Revolutionizing financial ... IJSRA. IJRA+1
14. Oduro, F. T., Boateng, R., & Mensah, I. K. (2025). AI-powered fraud detection in digital banking: Emerging trends and research agenda. Int. Journal of Information Management. Gjeta
15. Oduro, F. T., Boateng, R., & Mensah, I. K. (2025). AI-powered fraud detection in digital banking: Enhancing security through machine learning. IJSRA. ResearchGate
16. SMOTE + NN approach (Zhu et al., 2024). arXiv. arXiv+1
17. Springer. (2024). Financial Fraud Detection Using Machine Learning: A Review of Literature. Springer. SpringerLink+2SpringerLink+2
18. UK Finance. (2025). UK Finance annual fraud report. The Times
19. Wikipedia (2025). Adversarial machine learning. Wikipedia. Wikipedia
20. Wikipedia (2025). Artificial intelligence in fraud detection. Wikipedia. Wikipedia
21. Wikipedia (2025). Oversampling and undersampling in data analysis. Wikipedia. Wikipedia
22. Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. arXiv preprint. arXiv