UPI Fraud Detection Using Machine Learning (Secure UPI )

Dr D Godfrey  
Department of Electronics and Communication  
Dayananda Sagar UniversityBengaluru , India  
godfrey-ece@dsu.edu.in

Dr Pushpamala S  
Department of Electronics and CommunicationBengaluru , India  
pushpa.mala-ece@dsu.edu.in

Karthik H M  
Department of Electronics and CommunicationBengaluru , India  
eng22ec0025@dsu.edu.in

Abstract: Secure UPI is a leading developer of sophisticated fraud detection devices. We leverage the powerful XG-Boost device learning principles to build sophisticated fraud identity devices. Because of its track record of success across numerous industries and its reputation for handling challenging datasets with efficiency, XG-Boost is a suitable substitute for improving the accuracy of fraud detection models. Our method preprocesses UPI transaction facts to extract relevant records such as transaction quantity, frequency, and locality. In order to benefit from the XG-Boost version's strong prediction abilities and ability to handle imbalanced datasets, this article trains it using a labelled dataset. The key indicators of feasible fraud are found using feature importance evaluation, which aids in the development of a system that is easier to understand and operate. Following training, the model is integrated directly into a real-time UPI transactions tracking system, where it keeps a watch out for any unusual characteristics in incoming transactions. To mitigate the effects of fraudulent behaviour, a 98.2% accurate system is built to give out fast notifications and take preventive action. By showcasing machine learning’s effectiveness in fraud detection, this challenge advances economic eras and enhances the security of UPI transactions.

Keywords—UPI, Fraud Detection, SMOTE, PCA, XG-Boost, Imbalanced Datasets, Dimensionality Reduction, Machine Learning.

# Introduction

In a world where virtual transactions dominate, the unified payments interface [UPI] has emerged as a crucial platform for streamlining and speeding up monetary exchanges. However, the rapid growth of digital payments brings the challenge pf combating sophisticated fraudulent activities. By developing an advanced UPI Fraud Detection System and utilizing contemporary and algorithm selection techniques, this initiative addresses critical issue.

The primary goal of the UPI Fraud Detection System is to enhance the security and reliability of virtual transactions within the UPI framework. The system aims to detect and mitigate fraudulent activities using modern techniques, ensuring the safety and trustworthiness of the UPI environment. During the design and deployment of the UPI Fraud Detection System, the protection and privacy of sensitive financial data are paramount. Strict procedures that comply with legal standards and industry best practices are implemented to safeguard and anonymize the collected data. The system is designed to operate within privacy and data security regulations fostering user confidence and legal compliance.



Figure 1

As shown in Figure 1, the interface of the secure UPI website is designed for continuous improvement and is not static. The system architecture includes continue learning, allowing it to adapt to evolving fraud techniques and customer behavior.

By automatically retraining the model, the system uses the latest data to enhance its predictive capabilities. Incorporating continuous learning ensures that the UPI Fraud Detection System remains reliable defender against new threats and maintains its effectiveness over time.

To address the inherent class imbalance in UPI transaction datasets, the Synthetic Minority Over Sampling Technique -SMOTE is employed during pre-processing. Since fraudulent instances are typically rare, SMOTE generates synthetic examples to balance the classes, enabling the algorithm to train on a more representative dataset. This improves the model’s ability to identify fraudulent patterns and eliminates bias towards the majority class. Principal Component Analysis (PCA) is also used to reduce dimensionality while capturing the most significant information in the dataset, helping to identify patterns that contribute to data variation and enhancing the development of reliable fraud detection models. The UPI Fraud Detection System operates in real-time, continuously monitoring transactions as they occur.

By leveraging the speed and efficiency of the XG Boost algorithm, transactions are swiftly compared to learned patterns allowing for the immediate detection of potentially fraudulent activity. Real-time monitoring ensures that any suspicious activity is promptly addressed, reducing the negative impact of fraudulent transactions on both users and financial institutions.

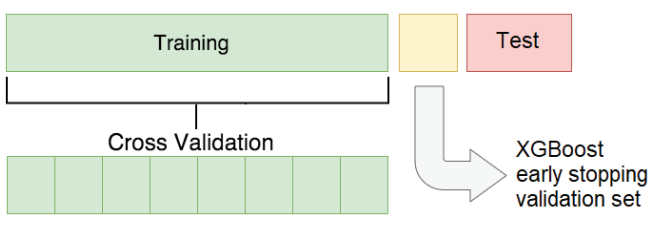


Figure 2

The data splitting procedure, which splits the data into two pieces, is shown in Figure 2. XG Boost algorithm's optimal parameter combination is found by K-fold cross-validation. By assessing the model's performance on a validation set at each iteration, early stopping is used to prevent overfitting. This indicates that while the number of trees is optimised during training, overfitting must be prevented and the accuracy of the cross-validation score must be guaranteed by using a different validation set.

XG-Boost, an ensemble learning algorithm renowned for its remarkable performance and flexibility, serves as the main classification model. Because of its capacity to manage intricate linkages, nonlinearities, and interactions within the data, it is especially well-suited for fraud detection. An extremely precise and adaptable classification procedure is guaranteed by the ensemble of decision trees, which are trained successively to fix prior mistakes. The project includes extensive hyperparameter tuning to maximize XG-Boost predictive power. Systematic adjustments to parameters such as learning rate, maximum tree depth, and the number of boosting rounds are made. This optimization enhances the algorithm’s performance, achieving the right balance between precision and recall -crucial metrics for fraud detection.

In-depth reporting and assessment tools offer information on the performance, efficacy, and abnormalities found in the system. Stakeholders may evaluate flagged transactions, discover factors that contribute to risk ratings, and obtain a comprehensive picture of the fraud detection environment through interactive dashboards and visualisations. This reporting feature supports ongoing system improvement, informed decision-making, and compliance reporting. A notable development in financial security is the UPI Fraud Detection System. By employing machine learning, anomaly detection, and behavioural analysis, the initiative seeks to create a strong barrier against fraudulent UPI transactions. This all-encompassing method not only identifies and stops fraudulent activity, but it also continually improves security protocols in the dynamic world of online banking. As a result of the project's increased customer confidence in UPI transactions, financial security may be considerably enhanced.

# 

# Litterature survey

M.A. Ibrahim [1] talks about common fraud strategies, how to spot them, and new developments in the industry. After balancing their dataset using the SMOTE technique, the researchers discovered that models including K-Nearest Neighbour, Decision Tree, Random Forest, and Neural Network functioned effectively when trained on balanced data. Users are able to select the model that they desire using the system. When trained using unbalanced transaction datasets, the Random Forest model demonstrated reduced efficiency despite achieving an accuracy of 93.58%.

The goal of P. Boulieris [2] was to provide a publicly accessible anonymised dataset for internet fraud detection. To measure model performance in real-world scenarios, the study recommends supplementing typical evaluation measures with both online and offline tests. All metrics were enhanced by the addition of anomaly detection features, with the exception of online detection. In comparison to a previous study, the addition of NLP-based features dramatically improved performance while utilising fewer traditional features.

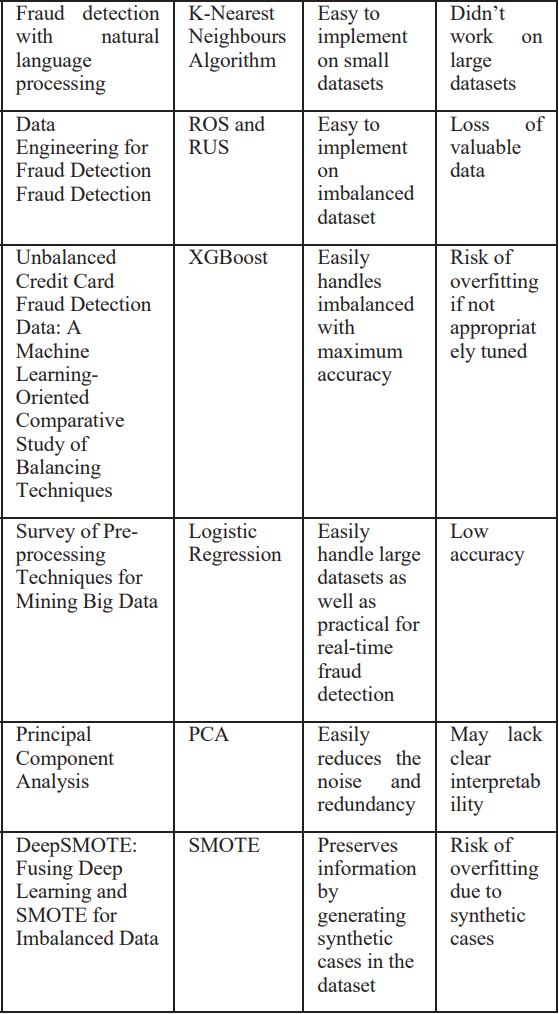
The significance of information technology in fraud detection is emphasised by B. Baesens [9], who points out that clever data design can increase analytical efficiency more successfully than complex new analytical approaches. The study showed that features design and request design both greatly increase the performance of analytical models using payment transaction data from a prominent European bank. If the data is well designed, even basic analytical techniques like logistic regression and classification tress can produce useful findings. The methods covered here can also be used in other fraud like e- commerce , insurance , and healthcare.

The XGBOOST classifier first outperformed four other classification algorithms, according to S. Alam [12]. Three data balancing strategies were evaluated in order to improve performance; ROS performed the best, but RUS and SMOTE did not. Accuracy, precision, recall, and F1 scores of the XG-Boost classifier were all markedly improved by using Random over sampling. Future investigations should look into more practical methods and techniques to deal with the shortcomings this study found.

S. Mohanavalli [13] focuses on data cleaning processes that address data quality issues such as noise, outliers, inconsistent data, duplicates, and missing values. Traditional techniques primarily handle these issues during data collection, but hybrid error correction methods need to be developed to provide noise-free samples for data mining and analysis. Different denoising techniques can create varied databases, with the median being used for homogeneous datasets and an expanded range of values for heterogeneous datasets. Future work will investigate hybrid methods for cleaning homogeneous data and reducing noise in heterogeneous datasets using a voting system that combines multiple techniques. The impact of these denoising techniques on classification performance and mining efficiency should also be explored for large datasets.

Principal Component Analysis (PCA) is a popular multivariate analysis technique for comprehending and analysing data, according to M. Greenacre [21]. The efficacy of PCA in extracting crucial information from intricate datasets has been demonstrated across several disciplines. Large datasets of various kinds may now be analysed using PCA thanks to recent breakthroughs, and this statistical technique should continue to be improved in the future. PCA is regarded as a basic tool in data science, along with its extensions and variations.

In order to deal with uneven data, D. Dablain [22] presents DEEPSMOTE, a unique model that blends deep learning with the SMOTE method. To ensure that the training set is balanced and that deep classifiers are trained without bias, DEEPSMOTE creates synthetic instances. It can interact with raw photos, generate high-quality synthetic images, and effectively create low-dimensional embeddings. Experimental studies demonstrate that DEEPSMOTE works better than conventional oversampling methods and is robust against various imbalance ratios. Subsequent efforts will concentrate on improving DEEPSMOTE by tackling issues at the class and instance levels, optimising the loss function, and broadening its scope to support dynamic class ratios and extra data modalities including text and graphs.



1.1 Comparison of the previous related research

1. Proposed Methodology

The general method we are using with our dataset is depicted in figure 3 below. Following the division of the dataset into training and test sets, where the test set comprises the final 20% of the original data and the training data set comprises 80% of the original data. After that, we’ll apply all of these methods to our training set, and finally, we’ll utilise the test set to forecast the veracity of the transactions.

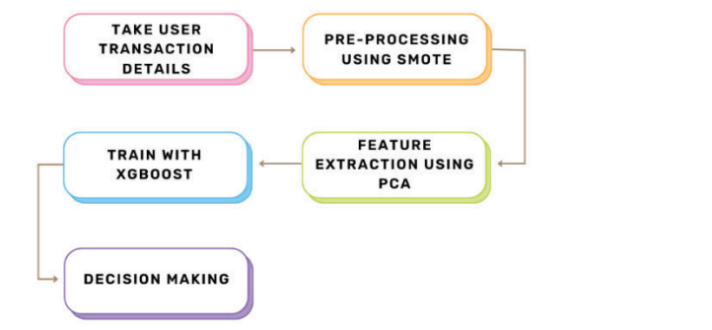


Fig 3. Procedure for detection of UPI fraud

1.**Data Collection** (Taking user transaction details)

Description: In the first step, pertinent data regarding user transactions is collected. This contains any data that is judged essential for the fraud detection process, such as transaction amounts, timestamps, user details and other facts.

Goal: Acquiring a sizable dataset is crucial for training a robust fraud detection model. The subsequent steps of the process are built upon the dataset.

2.**Pre-Processing SMOTE Algorithm**:

Description: The synthetic Minority Over-sampling Technique approach is used to pre-process the dataset inside the second degree. Particularly when there is a significant ratio of genuine to fraudulent transactions, SMOTE is specifically built to handle information built to handle information imbalances.

Purpose: Unbalanced dataset results may be prevented via biassed models that perform poorly on minority training SMOTE generates synthetic samples in order to optimise the distribution of commands and improve the version’s fraud detection skills.

**3.Using PCA for Feature Extraction:**

Description: The 1/three diploma uses Principal Component Analysis (PCA) to extract characteristics. PCA is used to reduce the dimensionally of the dataset while preserving significant information. This helps to overcome the additional computational challenges posed by large-scale statistics.

Goal: By reducing dimensionally with PCA, the dataset becomes more accessible for system learning algorithms. The goal of extracted abilities is to improve model performance by extracting the most important information.

**4.XGBoost Model Training**: In the fourth section, a pre-processed and function-extracted dataset is used to train the XG-Boost system learning model. XG-Boost is a well-known ensemble learning set of politics that, thanks to its amazing traditional performance and performance, excels at handling intricate, non-linear connection in statistics.

Goal: XG-Boost is chosen due to its adaptability in handling datasets that aren’t balanced, noise tolerance, and ability to recognise intricate styles within the recordings. The version

Can identify between legitimate and fraudulent transactions based on the skills that have been processed and modified.

**5. Reak or Fake Transaction:**

Decision: Using the informed version of XG-Boost is the last step in making predictions on new, unseen transactions. The version classifies each transactions as either authentic or fraudulent using the styles it has learned at some point in the future during training.

Goal: This phase specifies how the version must be applied in practical settings. The reason is that it provides financial institutions or rate processors with a tool for creating preference by accurately identifying by accurately identifying and flagging fraudulent transactions in all likelihood. This flowchart outlines a thorough process that starts with data in order to detect UPI fraud.

## 

## Results

To determine if a transaction is legitimate or fraudulent , we must click the “click to detect “ button after entering all the transactional information. Figure 4 will be displayed as an output if the transaction is legitimate, and Figure 5 will be displayed if it is fraudulent.

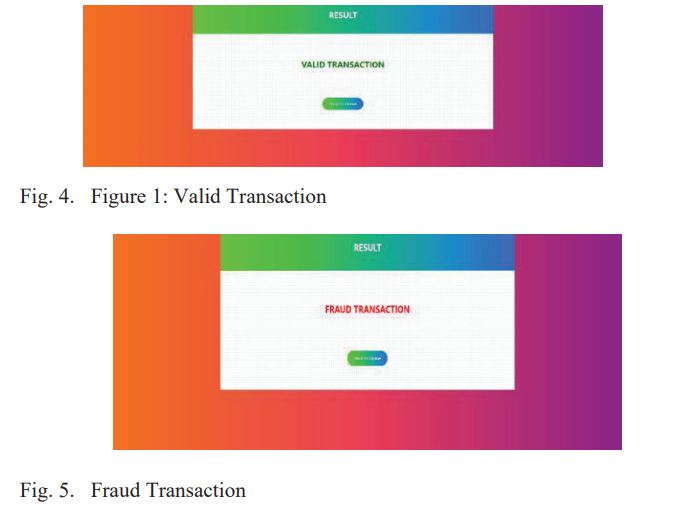


Figure 6 shows a comparison of several methods, with XG-Boost performing better than every approach.

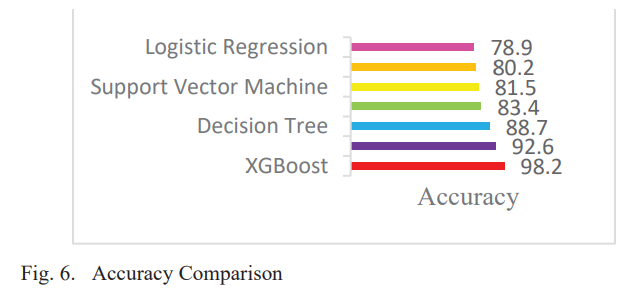
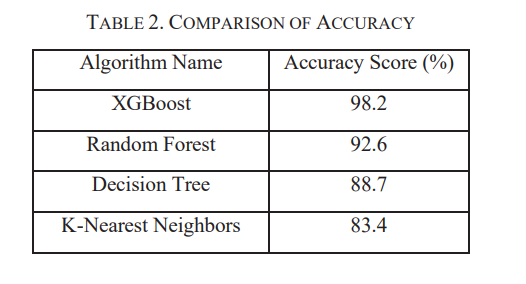
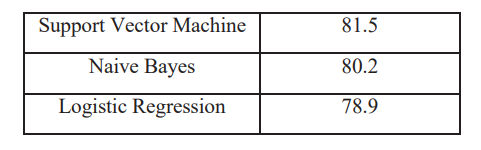


Table 2 – comparision of the algorithms





1. Conclusion

This UPI fraud detection project has applied innovative methodologies to enhance the robustness and efficiency of our fraud detection model.By addressing the imbalanced nature of the dataset with the synthetic Minority over-Sampling Technique SMOTE , we have successfully balanced the representation of fraudulent transactions, boosting the model’s ability to recognize subtle patterns indicative of fraudulet operation.

By identifying the most important parts of the data and reducing multicollinearity, the use of principal component analysis (PCA) for feature extraction has helped to reduce dimensionality. This made it easy to comprehend the underlying fraud tendencies in the representation and simplified the computing process.

We have had success using XG-Boost as the training model because of its ability to handle skewed data, resilience, and high projected accuracy. The ability of XG-Boost to identify complex linkages in the data has greatly improved the model's overall performance in identifying potential UPI fraud cases.

In the future, we can decrease the quantity of transaction details [23] [24] required to determine the validity of a transaction, while also including analytics on user behaviour to monitor patterns in user interactions with the UPI platform. Anomalies in behaviour could serve as early warning indicators of potential fraud.

.

##### References

[1]warning signs of possible fraud. REFERENCES [1] Adekunle, I. M., &Ozoh, P. (2023). Fraud detection model for illegitimate transactions. Kabale University Interdisciplinary Research Journal, 2(2), 21-37 https://doi.org/10.1016/j.future.2015.01.001

[2] Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2023). Fraud detection with natural language processing. Machine Learning,

[3] Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S., & Syerov, Y. (2023). Application of Artificial Intelligence for Fraudulent Banking Operations Recognition. Big Data and Cognitive-Computing,7(2),93. <https://doi.org/10.1016/j.dss.2010.08.008>

[4] Ridwan, R., Abdullah, S., & Yusmita, F. (2022). IMPLEMENTATION OF CASHLESS POLICY STRATEGIES TO MINIMIZE FRAUD IN THE GOVERNMENTSECTOR: SYSTEMIC REVIEW. Jurnal Akuntansi, 12(3), 181-201. [https://doi.org/10.1007/s10994-023- 06354-5](https://doi.org/10.1007/s10994-023-%2006354-5)

[5] Chang, V., Di Stefano, A., Sun, Z., & Fortino, G. (2022). Digital payment fraud detection methods in digital ages and Industry 4.0. Computers and Electrical Engineering,100, 107734. <https://doi.org/10.1145/3394486.3403361>

[6] Bandyopadhyay, S. K., & Dutta, S. (2020). Detection of fraud transactions using recurrent neural network during COVID-19: fraud transaction during COVID-19. Journal of Advanced Research in Medical Science & Technology (ISSN: 2394-6539),7(3),16-21. <https://doi.org/10.1016/j.compeleceng.2022.107734>

[7] Manocha, S., Kejriwal, R., & Upadhyaya, D. A. (2019, September). The impact of demonetization on digital payment transactions: a statistical study. In Proceedings ofInternational Conference on Advancements in Computing & Management (ICACM). https://doi.org/10.1109/TNNLS.2021.3136503