International Journal of Academic Research and Development
www.allreviewjournal.com
ISSN: 2455-4197
Received: 27-10-2025, Accepted: 26-11-2025, Published: 11-12-2025
Volume 10, Issue 6, 2025, Page No. 57-60

AI-Based secure monitoring and fraud detection for UPI payment transactions

Anoop Sharma
Assistant Professor, Department of Computer Science, Government Degree College, Marheen, U. T of Jammu and Kashmir, India

Abstract
The Unified Payments Interface (UPI) has fundamentally transformed the digital payment landscape in India, enabling billions of transactions every month across urban and rural populations alike. However, this explosive growth has simultaneously attracted a surge in fraudulent activities, including phishing, SIM swapping, social engineering, and account takeover attacks. Traditional rule-based fraud detection systems have proven inadequate against the dynamic and evolving nature of modern payment fraud. This paper proposes and examines an Artificial Intelligence (AI)-based framework for secure real-time monitoring and fraud detection specifically designed for UPI payment transactions in the Indian context. The proposed framework integrates machine learning algorithms including Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) networks to analyze transaction patterns, detect anomalies, and flag suspicious activities with high accuracy and low false-positive rates. The paper also discusses data privacy considerations under India's Digital Personal Data Protection Act, 2023, the role of the National Payments Corporation of India (NPCI) in fraud governance, and the challenges unique to India's diverse digital payment ecosystem. Experimental results and comparative analysis suggest that AI-driven models significantly outperform traditional systems in detection speed, accuracy, and adaptability. The paper further outlines future research directions including federated learning for privacy-preserving fraud detection and the integration of explainable AI (XAI) for regulatory transparency.

Keywords: UPI fraud detection, artificial intelligence, machine learning, digital payments India, anomaly detection, NPCI, real-time monitoring, lstm, federated learning, cyber security

Introduction

India's digital payment revolution has been one of the most remarkable technological transformations of the twenty-first century. The Unified Payments Interface, launched by the National Payments Corporation of India in April 2016, has grown from a modest beginning to a platform processing over 13 billion transactions per month by early 2025. With flagship initiatives like Digital India pushing internet and smartphone penetration into semi-urban and rural areas, UPI has become the backbone of everyday financial transactions for hundreds of millions of Indians. Applications such as PhonePe, Google Pay, Paytm, and BHIM have made sending and receiving money as simple as sending a text message.

However, the same accessibility that makes UPI powerful also makes it vulnerable. As the transaction volume has grown, so has the sophistication and frequency of fraud. According to the Reserve Bank of India's Report on Currency and Finance 2024 [11], digital payment frauds have increased significantly over the past three years, with UPI-related complaints forming a substantial share of all reported financial cyber crimes. The nature of these frauds ranges from relatively simple cases of social engineering — where victims are tricked into sharing OTPs — to highly sophisticated automated attacks that exploit API vulnerabilities and behavioral patterns.

Conventional fraud detection mechanisms employed by banks and payment service providers have largely depended on static rule-based systems. These systems work by flagging transactions that match predefined patterns, such as unusually large transfer amounts or transactions from unknown devices. While effective against known fraud patterns, they are inherently reactive. They fail to adapt to new fraud strategies and often produce high rates of false positives, which inconvenience legitimate users and erode trust in digital payment systems. In a country where digital financial inclusion is a national policy priority, such friction can have real consequences on adoption rates, particularly among first-time users in smaller towns and villages.

Artificial Intelligence and machine learning offer a fundamentally different approach. Rather than relying on fixed rules, AI models learn from historical transaction data to identify subtle behavioral patterns that distinguish legitimate transactions from fraudulent ones. They can operate in real time, adapt continuously to new fraud patterns, and handle the enormous scale that UPI transactions demand. This paper presents a comprehensive examination of how AI-based systems can be designed and implemented for UPI fraud detection, with specific attention to the Indian regulatory environment, data diversity, and infrastructure realities.

The rest of this paper is organized as follows. Section 2 reviews existing literature on payment fraud detection and AI applications. Section 3 describes the proposed AI-based monitoring framework. Section 4 discusses the methodology and algorithmic approach. Section 5 presents results and analysis. Section 6 addresses privacy, ethics, and regulatory considerations. Section 7 outlines challenges and future directions, followed by the conclusion in Section 8.

Literature Review

Research on fraud detection in financial systems has a rich history, but the specific application of AI to UPI and real-time mobile payment fraud is relatively recent and growing rapidly. Early work in financial fraud detection focused on credit card transactions. Bhattacharyya *et al.* (2011) [2]

demonstrated that Support Vector Machines and Random Forest classifiers could significantly outperform logistic regression on imbalanced fraud datasets, a finding that has remained relevant across payment platforms. Dal Pozzolo *et al*. (2015) [3] introduced the concept of adaptive machine learning models that could be retrained on streaming transaction data, which laid important groundwork for real-time systems.

In the Indian context, Jain and Kaur (2020) studied the behavioral patterns of UPI users and identified key features such as transaction time, device fingerprinting, geolocation anomalies, and frequency of transactions as strong predictors of fraudulent activity. Their work highlighted that fraud patterns in India are shaped by unique socioeconomic factors, including low digital literacy among new users and the widespread use of shared devices in rural households. Sharma and Mishra (2021) [12] proposed a hybrid model combining rule-based filters with a neural network backend for UPI fraud detection and reported improved precision over standalone rule-based systems.

More recently, the use of deep learning architectures has gained attention. Agarwal *et al*. (2022) [1] applied LSTM networks to sequential UPI transaction data and demonstrated that temporal patterns in transaction history contain significant fraud signal, particularly for account takeover scenarios where a fraudster gradually escalates transaction amounts after gaining access to a compromised account. Their model achieved an F1score of 0.91 on a dataset of over two million transactions.

Graph-based approaches have also been explored. Rao and Venkatesh (2023) [9] modeled UPI transaction networks as directed graphs where nodes represent users and edges represent money flows. By applying Graph Neural Networks (GNNs), they were able to detect fraud rings — coordinated groups of accounts engaged in layered money movement to obscure the origins of fraudulent funds. This approach is particularly relevant in India where organized fraud syndicates operating from specific geographic clusters have been documented by the Indian Cyber Crime Coordination Centre (I4C).

Federated learning has emerged as a promising solution to the data-sharing challenge among competing payment service providers. Mehta and Bose (2024) [7] demonstrated that a federated learning framework across multiple UPI member banks could improve fraud detection accuracy by 17% compared to models trained on isolated bank-level data, without requiring any raw transaction data to leave the originating institution. This is especially relevant given the constraints of India's Digital Personal Data Protection Act, 2023.

Despite these advances, several gaps remain in the literature. Most existing studies use synthetic or proprietary datasets that are not publicly available, making benchmarking difficult. There is also limited work on explainability — understanding why a model flags a transaction — which is crucial for regulatory compliance and user dispute resolution. This paper attempts to address these gaps within the Indian UPI context.

Proposed Ai-Based Monitoring Framework

The proposed framework is designed as a layered, real-time system that sits between the UPI transaction initiation layer and the final settlement layer. It consists of four primary components: data ingestion and preprocessing, feature engineering, multi-model fraud scoring, and alert and response management.

The data ingestion layer continuously receives transaction metadata from UPI rails through secure API integration with NPCI's systems and participating bank infrastructure. This metadata includes the transaction amount, timestamp, payer and payee Virtual Payment Addresses (VPAs), device ID, IP address, GPS coordinates where available, mobile network operator, and UPI application identifier. Crucially, no sensitive personal financial data such as bank account numbers or card details is processed at this layer, reducing privacy exposure.

The preprocessing module cleans incoming data, handles missing fields through imputation, and normalizes numerical features. It also applies tokenization to VPAs so that identity linking can be done securely without exposing raw account information to the analytics engine. A real-time data stream is maintained using Apache Kafka-style event streaming, which allows the system to process thousands of transactions per second without latency-related bottlenecks.

Feature engineering is the most critical stage of the pipeline. Raw transaction data is transformed into meaningful behavioral and contextual features. These include the velocity of transactions from a given device or VPA within rolling time windows of one minute, fifteen minutes, and twenty-four hours; the deviation of transaction amount from the user's historical average; the geographic distance between the current transaction location and the user's typical location cluster; the novelty of the receiving VPA; and the time since account creation for newly registered users who are statistically more fraud-prone. Additional derived features capture network-level signals such as whether the payer and payee have transacted before and the broader connectivity of both accounts in the UPI transaction graph.

The fraud scoring engine operates through an ensemble of three complementary models. A Random Forest classifier handles high-dimensional feature spaces and provides robust baseline detection for known fraud patterns. A Gradient Boosting model, specifically XGBoost, is applied for its superior performance on tabular data and its ability to capture complex non-linear interactions between features. An LSTM network processes the sequential transaction history of each user as a time series, enabling detection of gradual behavioral drift — a hallmark of account takeover fraud. The outputs of all three models are combined through a weighted average ensemble, with weights dynamically adjusted based on recent model performance metrics. This ensemble approach ensures that no single model's failure mode can cause widespread missed detections.

The alert and response layer translate fraud scores into actionable outcomes. Transactions scoring below a low-risk threshold are passed through immediately. Those in a medium-risk band trigger additional authentication challenges such as a second OTP or a biometric prompt through the UPI application. Highrisk transactions are held for a brief review window of up to thirty seconds, during which an automated review agent attempts to gather additional confirming or disconfirming signals. Transactions that remain high-risk after this window are flagged for blocking and the user is notified. This tiered approach minimizes friction for legitimate users while maintaining strong protection against fraud.

Methodology

The methodology for evaluating the proposed framework was built around a simulated transaction dataset constructed to mirror real UPI transaction characteristics as reported in NPCI annual reports and published academic datasets from Indian banking research. The dataset comprised approximately five million synthetic transactions generated using statistical distributions calibrated to match known UPI usage patterns, including transaction amount distributions, peak usage hours, geographic spread across Tier-1, Tier-2, and Tier-3 cities, and the proportion of transactions occurring on feature phones versus smartphones.

Fraud cases were injected into the dataset at a rate of approximately 0.3%, consistent with industry estimates for UPI fraud incidence. Five categories of fraud were represented: social engineering induced transfers, SIM swap attacks, phishing-based account takeovers, merchant fraud, and automated bot-driven micro-transaction probing attacks. Each category was modeled with distinct behavioral signatures based on documented fraud patterns reported by the Cyber Crime cells of various Indian state police departments and RBI advisories published between 2022 and 2024.

The dataset was split into 70% training, 15% validation, and 15% test sets, with stratification to maintain fraud proportions across splits. All models were trained using the training set and hyperparameters were tuned on the validation set using grid search with cross-validation. Final performance was evaluated on the held-out test set. The key evaluation metrics used were Precision, Recall, F1-Score, Area Under the ROC Curve (AUC-ROC), and average detection latency in milliseconds, since real-time performance is a nonnegotiable requirement for a live payment system.

To simulate real-world distribution shift — where fraud patterns evolve over time — the test set was also divided into temporal batches representing successive weeks of transactions, and model performance was tracked across these batches to assess degradation over time and the effectiveness of online retraining.

Results and Analysis

The ensemble model demonstrated strong performance across all evaluation metrics on the test dataset. The Random Forest model alone achieved a Precision of 0.88 and Recall of 0.84, yielding an F1-Score of 0.86 and an AUC-ROC of 0.93. The XGBoost model performed comparably with Precision of 0.89, Recall of 0.85, and AUC-ROC of 0.94. The LSTM model showed particular strength in detecting account takeover fraud, where its Recall on that specific fraud category reached 0.91, owing to its ability to model behavioral sequences. The final ensemble achieved Precision of 0.92, Recall of 0.89, F1-Score of 0.905, and AUCROC of 0.97, outperforming each individual model.

Compared to a baseline rule-based system modeled after typical bank fraud filters — which achieved a Precision of 0.71 and Recall of 0.62 — the AI ensemble represents a substantial improvement, particularly in reducing false negatives, which correspond to fraud cases that go undetected. The false positive rate of the ensemble system was 1.8%, compared to 6.3% for the rule-based baseline, meaning that legitimate users faced unnecessary friction far less often.

In terms of latency, the ensemble model produced fraud scores within an average of 47 milliseconds per transaction, well within the practical real-time threshold required to insert a decision into the UPI transaction flow before settlement confirmation. The LSTM component, being the most computationally intensive, accounted for the majority of this latency. Optimization using model quantization and hardware acceleration reduced LSTM inference time by approximately 35% without meaningful accuracy loss.

The temporal batch analysis showed that model performance degraded modestly over successive weeks as fraud patterns shifted, with F1-Score dropping from 0.905 in the first week to 0.871 by week eight. However, incremental retraining using newly labeled fraud cases from the preceding week restored performance in each subsequent batch, demonstrating the viability of a continuously learning fraud detection system.

Privacy, Ethics, and Regulatory Considerations

Any AI-based fraud detection system operating in India must navigate a carefully defined regulatory landscape. The Digital Personal Data Protection Act, 2023 establishes clear obligations for entities that process personal data of Indian citizens. Transaction metadata, device identifiers, and location information used by the proposed system all potentially qualify as personal data under this Act. The system architecture addresses this through several design choices: data minimization ensuring only necessary fields are collected, tokenization of user identifiers before processing, strict data retention limits with automatic deletion of raw transaction logs after seven days, and access control mechanisms that prevent individual analysts from linking fraud scores back to identifiable users without formal authorization.

The Reserve Bank of India's Master Direction on Digital Payment Security Controls, updated in 2023 [10], mandates that payment system operators maintain audit trails for fraud detection decisions, provide users with a mechanism to dispute blocked transactions, and report significant fraud incidents to the RBI within specified timeframes. The proposed framework includes a decision logging module that records the feature values and model scores associated with every flagged transaction, enabling post-hoc review and audit. A user-facing dispute interface integrated with the UPI application allows users to challenge a blocked transaction, triggering a human review workflow.

The question of explainability deserves particular attention. When a legitimate transaction is blocked, users and regulators have a reasonable expectation of understanding why. Black-box AI models, while powerful, cannot natively provide this explanation. The framework incorporates SHAP (SHapley Additive exPlanations) values computed at inference time for every high-risk flagged transaction. These SHAP values quantify each feature's contribution to the fraud score, enabling the generation of human-readable explanations such as "this transaction was flagged because the receiving account was new and the transaction amount was five times higher than your recent average."

Algorithmic bias is another critical concern in the Indian context. AI models trained primarily on urban transaction data may exhibit lower recall on rural user transactions, which have different behavioral baselines. The training methodology explicitly addressed this by ensuring geographic stratification in the training data and evaluating

model performance separately for Tier-1, Tier-2, and Tier-3 city user segments. Results showed no statistically significant disparity in F1-Score across these segments, suggesting the model generalizes well across India's diverse user base.

Challenges and Future Directions

Despite promising results, several challenges must be acknowledged. First, data availability remains the most significant practical barrier. Training high-quality fraud detection models requires large volumes of labeled fraud data, which individual banks and payment providers are reluctant to share due to competitive and regulatory concerns. Federated learning, where models are trained across distributed data sources without sharing raw data, represents the most viable path forward and deserves focused research investment in the Indian banking context.

Second, the adversarial nature of fraud means that any published or deployed detection system risks being studied and circumvented by sophisticated fraudsters. This necessitates continuous model updating, active monitoring of detection rates, and investment in threat intelligence to anticipate emerging attack vectors. The rise of AI-generated voice and deepfake-based social engineering attacks, already reported by Indian law enforcement agencies in 2024, represents a new frontier that current behavioral models are not fully equipped to handle.

Third, India's digital payment infrastructure spans an enormous diversity of devices and network conditions. A significant proportion of UPI transactions occur on entry-level Android smartphones with limited processing capability and intermittent internet connectivity. Any fraud detection system that introduces perceptible latency or requires frequent re-authentication will face adoption resistance, particularly in rural markets. Lightweight model architectures and edge computing approaches that push some detection intelligence onto the device itself are worth exploring.

Future research directions identified by this study include the development of a standardized, privacypreserving benchmark dataset for UPI fraud detection that can be shared across the research community; deeper integration of Graph Neural Networks for fraud ring detection at the NPCI network level; exploration of reinforcement learning approaches where the fraud detection system actively learns from the outcomes of its interventions; and the design of multilingual fraud alert communication systems that can reach users in their preferred language across India's twenty-two scheduled languages.

Conclusion

This paper has presented a comprehensive AI-based framework for real-time secure monitoring and fraud detection in UPI payment transactions, grounded in the specific technical, social, and regulatory realities of India's digital payment ecosystem. The proposed ensemble approach combining Random Forest, XGBoost, and LSTM models demonstrated significantly superior performance compared to conventional rule-based systems, achieving an F1-Score of 0.905 and AUC-ROC of 0.97 while maintaining real-time inference latency under fifty milliseconds.

Beyond the technical results, the paper has argued that effective fraud detection in the Indian context must simultaneously address data privacy compliance under the 2023 DPDPA, algorithmic fairness across India's diverse user demographics, regulatory transparency through explainable AI, and the practical constraints of a heterogeneous device and network environment. These are not peripheral concerns but central design requirements for any system that aspires to operate at UPI's scale.

As India continues its journey toward a less-cash economy and as UPI expands internationally through linkages with payment systems in Singapore, UAE, and other countries, the importance of robust, intelligent fraud detection will only increase. The framework presented in this paper provides a research foundation and a practical blueprint for this critical challenge.

References

1. Agarwal R, Singh P, Tiwari M. Deep learning-based fraud detection in UPI transactions using LSTM networks. Journal of Financial Technology and Innovation,2022:4(2):45–61.
2. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: A comparative study. Decision Support Systems,2011:50(3):602–613.
3. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. IEEE Symposium on Computational Intelligence and Data Mining, 2015, 159–166.
4. Government of India. Digital Personal Data Protection Act, 2023. Ministry of Electronics and Information Technology, New Delhi, 2023.
5. Indian Cyber Crime Coordination Centre (I4C). Annual Report on Cyber Financial Fraud in India 2023–24. Ministry of Home Affairs, Government of India, 2024.
6. Jain A Kaur R. Behavioral analysis of UPI users for fraud identification in Indian digital payment systems. International Journal of Cybersecurity and Digital Forensics,2020:9(3):112–126.
7. Mehta S Bose A. Federated learning for collaborative fraud detection across UPI member banks: A privacy-preserving approach. Proceedings of the International Conference on AI in Financial Services, Mumbai, 2024, 78–93.
8. National Payments Corporation of India (NPCI). UPI Product Statistics and Annual Report 2023–24. NPCI, Mumbai, 2024. Retrieved from www.npci.org.in.
9. Rao KV Venkatesh N. Graph neural networks for fraud ring detection in UPI transaction networks. IEEE Transactions on Network and Service Management,2023:20(1),334–349.
10. Reserve Bank of India. Master Direction on Digital Payment Security Controls (Updated). RBI, Mumbai, 2023.
11. Reserve Bank of India. Report on Currency and Finance 2023–24: Digital Payments and Financial Inclusion. RBI, Mumbai, 2024.
12. Sharma D Mishra V. Hybrid rule-neural model for real-time UPI payment fraud detection. Proceedings of the National Conference on Emerging Technologies in Banking and Finance, New Delhi, 2021, 201–215.
13. Zhang Y, Chen H, Liu W. Explainable AI in financial fraud detection: A survey of SHAP-based approaches. ACM Computing Surveys,2023:55(9),1–34.