Proceedings of Engineering and Technology Innovation, vol. 33, 2026, pp. 93-105

A Hybrid Autoencoder-XGBoost Framework for High-Performance UPI Fraud Detection

Naga Bhavani Chakka, Shaiku Shahida Saheb*

VIT-AP School of Business, VIT-AP University, Amaravati, Andhra Pradesh, India

Received 21 December 2025; received in revised form 11 March 2026; accepted 17 March 2026

DOI: https://doi.org/10.46604/peti.2026.16002

Abstract

This paper proposes a scalable hybrid autoencoder–XGBoost system for detecting unified payment interface fraud. The approach begins by training an autoencoder on legitimate transactions to learn normal behavior patterns, where reconstruction errors are derived and utilized as anomaly scores. These scores serve as engineered features in an XGBoost classifier for final fraud classification. The system is tested on a synthetic dataset of 2.68 million transactions. The findings show near-perfect performance, with accuracy, precision, recall, and F1-scores close to 1.0 and a receiver operating characteristic curve-area under curve (ROC–AUC) of 0.99999995. However, these results are influenced by deterministic fraud patterns in the simulated dataset, leading to near-separable classes with domain-driven balance features. Therefore, the performance should be interpreted as proof-of-concept under controlled synthetic conditions rather than absolute evidence of real-world effectiveness. The model demonstrates the potential of anomaly-aware feature enrichment for handling severely imbalanced data. Future work will focus on validation with real-world UPI data and adaptive learning upgrades.

Keywords: decision-support systems, digital payments security, financial technology (FinTech), machine learning scalability, real-time transaction monitoring.

1. Introduction

The rapid growth of the digital payment market is primarily attributed to technological innovation, rising internet penetration, and the global shift toward cashless economies [1]. The unified payments interface (UPI) was introduced in India by the National Payments Corporation of India (NPCI) in April 2016, marking a major development in digital payment innovation. Guided by the Reserve Bank of India (RBI, 2019), India's digital payments environment is undergone a significant transformation over the last decade, driven by government efforts, widespread smartphone adoption, and increased internet penetration.

The study examines data from FY 2017-18 to FY 2024-25 and finds a 44.5% compound annual growth rate (CAGR) in transaction volume and 12.1% in transaction value, showing a shift toward smaller, more frequent payments. The UPI is the key growth driver, with its share of total digital transactions increasing from 4.44% to 79.87% at a 129.5% CAGR. Despite this rapid expansion, challenges in cybersecurity and digital literacy remain, along with opportunities in global UPI expansion and offline or voice-based payments [2]. Furthermore, UPI is designed to enable seamless peer-to-peer (P2P) and peer-to-merchant (P2M) transactions, reducing reliance on cash while ensuring accessibility, transparency, and security. This platform fundamentally reshapes the Indian digital finance ecosystem, integrating convenience and interoperability across banks and payment applications [3].

* Corresponding author. E-mail address: shahid.sk@vitap.ac.in

Despite UPI's disruptive impact and rapid adoption, its expansion exposes the payment ecosystem to an increasing range of vulnerabilities. As transaction volumes grow, fraudsters become more active, using tactics such as phishing, account takeover, and the introduction of malicious software to defraud unwary customers. Many incidents exploit deficiencies in digital literacy, with users unintentionally disclosing critical credentials or falling victim to deceptive social engineering attempts [4]. While UPI's adaptability benefits users, it also creates extra security risks: large-scale assaults can target system vulnerabilities at numerous stages, ranging from device-level malware to flaws in third-party integrations.

As UPI continues to expand its reach, both globally and into offline or voice-enabled payments, stakeholders must enhance real-time monitoring, artificial intelligence (AI)-driven detection systems, and continual user education to maintain trust in India's digital payment ecosystem, particularly as fraud attempts become more sophisticated and frequent [5].

However, the rise in adoption has also coincides with an alarming increase in fraud cases. UPI-related fraud incidents in India grew by 85%, from 7.25 lakhs in FY 2022-23 to 13.42 lakhs in FY 2023-24, with financial losses escalating to ₹1,087 crore. Reported cases include unauthorized fund transfers and large-scale social engineering scams such as quick-response (QR) code tampering, fake job postings, and phishing operations. Globally, cyber fraud has also escalated, ranging from the 2016 Bangladesh Bank heist to the recent cryptocurrency “pig butchering” scam, causing $12.4 billion in losses in 2024 [3]. Efforts to mitigate such risks include measures by NPCI and the Securities and Exchange Board of India (SEBI) to mandate verified UPI handles, as well as Google India's digital safety initiatives, which employ AI to prevent fraudulent activities and reduce losses by up to ₹20,000 crore by 2025. Despite these advancements, fraud detection has not kept pace with the rapid adoption of UPI [6].

The emergence of machine learning (ML) and deep learning (DL) technology in UPI fraud detection stems from the increasing sophistication and immediacy of digital payment fraud. Recent studies have shown that a robust fraud detection system that uses an M-DBN with multiple layers of verification achieved very high levels of success in detecting and pinpointing fraudulent activities associated with QR codes and UPI Id's in real time [7]. Other studies developed ML-based secure UPI systems by utilizing XGBoost in the analysis of labeled UPI transactions, incorporating biometric data, compliance history, and user behavior. The conclusion from these studies is that gradient boosting techniques can be utilized most effectively when analysing structured UPI data [8].

Studies comparing the use of various classifiers, including logistic regression (LR), random forest (RF), decision tree (DT), and Naïve Bayes, suggest that Ensemble model approachesspecifically RF Classifier combined with over-sampling techniques such as synthetic minority over-sampling technique (SMOTE) and Tomek links—produced a decrease in the number of false positives when compared to the simpler models [9]. In addition, DL models such as long short-term memory (LSTM) networks have demonstrated the ability to learn and capture the sequential nature of transactions and to prove that temporal relationships play a significant role in UPI fraud detection [10]. Furthermore, convolutional neural networks (CNNs) and hybrid Markov components have shown that a combination of spatial, temporal, and device-based characteristics increases detection rates while also improving the scalability of the approach. Collectively, these studies confirm the effectiveness of advanced ML and DL models for UPI fraud detection while underscoring the need for robust, real-time, and adaptive solutions [11-12].

Although autoencoder–XGBoost hybrids demonstrated success in general fraud detection contexts, their adaptation to UPI-specific scenarios remains largely unexplored. Key gaps include handling extreme class imbalance, concept drift, UPI-specific fraud typologies (e.g., QR-code tampering and social engineering), and the requirements for model interpretability and scalability under India’s digital banking regulations. Addressing these gaps is essential for developing practical fraud detection solutions tailored to the UPI ecosystem.

This study's main goal is to provide a robust and scalable hybrid fraud detection framework for UPI transactions by combining a supervised XGBoost classifier with an unsupervised autoencoder. The suggested method uses autoencoder-based anomaly detection to extract latent fraud signs and enhance feature representations, which XGBoost then **utilizes** for precise classification. While preserving interpretability for decision support in real-time UPI contexts, this hybrid system seeks to improve detection accuracy, robustness against imbalanced data, and adaptation to evolving fraud trends.

The innovative aspect of this study lies in the structured integration of unsupervised anomaly modeling with supervised gradient boosting within a unified fraud classification pipeline tailored for large-scale UPI transaction data. While previous research has explored standalone DL models or boosting algorithms, this work introduces an anomaly-aware feature enrichment mechanism. In this approach, reconstruction errors derived from legitimate-only training are incorporated directly into a supervised classification framework. This approach enhances discriminative capability under extreme class imbalance without relying on synthetic resampling techniques, while preserving interpretability through domain-driven financial feature engineering. Therefore, the proposed design contributes a scalable and practically deployable methodology for transaction monitoring systems.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 discusses the methodology and framework design; Section 4 presents experimental results and analysis; Section 5 offers discussions and implications; and Section 6 concludes with future research directions.

2. Related Work

India’s digital payment ecosystem has been significantly transformed by UPI, which enables seamless, real-time, API-driven mobile transactions. However, its rapid adoption has expanded exposure to sophisticated fraud, including QR tampering, phishing, impersonation, merchant scams, and malicious applications. These threats exploit user trust and technical vulnerabilities, accelerating the adoption of AI- and ML-based defenses in recent research (2024–2025). Study [13] proposed a hybrid model integrating adversarial autoencoders with GRUs for anomaly detection, reducing false positives. Similarly, Prakash et al. [14] developed an adversarial attention network for sequential data, highlighting the superiority of deep hybrid models in real-time fraud detection

Even though DL techniques have enhanced fraud detection accuracy, critical issues related to data privacy, regulatory adherence, and centralized data exchange continue to challenge the UPI ecosystem. To mitigate these concerns, Singh and Tripathy [15] introduced FedShield, a federated learning framework that enables banks and payment service providers (PSPs) to jointly develop fraud detection models without sharing sensitive customer information. This decentralized methodology aligns with NPCI’s objective of maintaining a secure, interoperable, and scalable UPI infrastructure. Furthermore, Manta et al. [16] emphasized the importance of AI-driven analytics and adaptive learning for real-time transaction monitoring. In a related study, Hamed et al. [17] advocated integrating multi-factor authentication, behavioral biometrics, and AI-based anomaly detection to counter spoofing and impersonation threats.

Beyond accuracy and privacy, interpretability has become essential driven by regulatory requirements imposed by RBI and NPCI, requiring fraud detection models to be transparent and explainable. Talaat et al. [18] incorporated Explainable AI techniques such as SHAP and LIME to clarify AI-driven decisions and enhance accountability. Regarding computational efficiency, Arnaldo et al. [19] showed that LightGBM outperformed SVM in identifying complex, non-linear anomalies in large financial datasets. Building on this, the present study compares RF and XGBoost for UPI fraud detection, finding XGBoost superior under class imbalance. Likewise, Prabu et al. [20] reported the superior performance of XGBoost, despite moderate overall accuracy.

Recent studies further highlight the effectiveness of advanced and hybrid AI approaches over conventional rule-based systems in digital payment fraud detection. Youbi et al. [21] showed that DL models significantly outperformed traditional methods in e-commerce payments by managing class imbalance and enabling real-time analysis. Lawati et al. [22] demonstrated that combining dynamic preprocessing, imbalance handling, CNN-based learning, and drift detection yielded near-perfect accuracy while adapting to evolving fraud patterns. Mozumder et al. [23] enhanced detection performance through hybrid contrastive learning with Siamese networks and attention mechanisms, coupled with explainable AI for transparency.

Emerging paradigms such as quantum-enhanced fraud detection have also shown promise, with QGNNs and QSVMs outperforming classical models in accuracy and latency for real-time systems [24]. Domain-specific innovations include ShuffleNet–SVM architectures for efficient UPI fraud monitoring [25] and artificial neural network-based mobile wallet fraud detection [26]. Other studies introduced optimized CNN–recurrent neural network (RNN) models, which incorporated BiLSTM and gazelle optimization [27]. Furthermore, gradient boost–based whale–hawk optimization with bayesian learning achieved 99.76% accuracy [28]. Finally, the scalable black widow–driven gradient boosting machine addressed adaptive fraud detection [29]. Despite these advancements, existing models still struggle with evolving UPI scam patterns, zero-day frauds, interpretability, and real-time adaptability. These limitations highlighted the need for a hybrid, explainable, and scalable fraud detection framework that integrates unsupervised and supervised learning for robust UPI security.

Notwithstanding the significant advancements in DL, federated learning, and hybrid architectures, existing studies often prioritize predictive accuracy without explicitly integrating unsupervised behavioral modeling into structured ensemble pipelines. Moreover, many approaches depend heavily on complex architectures that may reduce interpretability and increase computational cost. To address these gaps, the present study proposes a streamlined anomaly-enriched boosting framework that balances detection performance, interpretability, and scalability within a unified structure.

3. Methodology

Fig. 1 depicts the overall methodology for UPI fraud detection in this study using a hybrid framework. The process begins with data collection from a publicly accessible financial transactions dataset, followed by data preprocessing and feature engineering to ensure the data is clean, consistent, and relevant. The dataset is inspected for missing and invalid values prior to feature engineering. Since the PS_20174392719 datasets does not contain missing values in the selected attributes, no imputation is required; however, in real-world deployments, standard strategies such as median imputation would be applied within the same preprocessing pipeline. During preprocessing, transaction categories containing no fraudulent instances are excluded to prevent trivial class separability and artificial performance inflation.

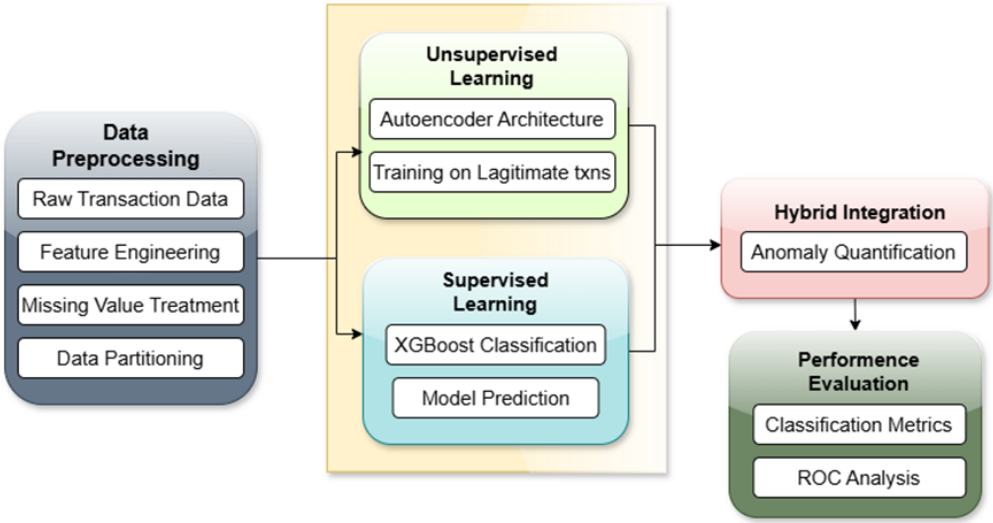

Fig. 1 Proposed AI-driven fraud detection framework for UPI transactions

Exploratory analysis confirmed that fraudulent events occur exclusively within the TRANSFER category, while certain transaction types are structurally fraud-free. Including such categories would allow the model to discriminate based on transaction type rather than behavioral irregularities. Therefore, the dataset is restricted to fraud-prone transaction flows to preserve a meaningful fraud detection setting. The dataset is then partitioned into training and testing sets to enable systematic model development and evaluation. The core of the proposed methodology is a two-stage hybrid model. First, an unsupervised autoencoder is trained exclusively on legitimate transactions to learn a low-dimensional representation of normal behavior.

The autoencoder is implemented as a fully connected symmetric neural network with an input dimension of 14 engineered features. The encoder consisted of three hidden layers with 16, 8, and 4 neurons respectively, where the 4-neuron layer represents the latent embedding space. The decoder mirrored this structure with layers of 8 and 16 neurons, followed by an output layer of 14 neurons. Rectified linear unit (ReLU) activation functions are applied in all hidden layers, while a sigmoid activation function is used in the output layer. The model is trained using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) as the reconstruction loss function. Training is conducted for 10 epochs with a batch size of 256 and a validation split of 10% to monitor convergence.

The reconstruction error for each transaction is computed using the mean absolute error (MAE) between the original and reconstructed feature vectors, and this value serves as an anomaly score. In the proposed framework, this anomaly score is not thresholded to produce a binary decision. Instead, it is used as a continuous, real-valued feature that is concatenated with the engineered financial features and provided to the XGBoost classifier. Consequently, no explicit anomaly score threshold is tuned in this study, and the classifier automatically learns how to weight this feature in combination with the other inputs. Changes in the autoencoder architecture would alter the distribution of the reconstruction error feature and may influence downstream classification performance; a systematic sensitivity analysis of this effect is left for future work**.**

Second, this anomaly score is concatenated with the original engineered features to form an enriched feature vector, which is then used to train a supervised XGBoost classifier for the final fraud classification task. The XGBoost classifier was configured with 100 DT (n_estimators = 100) and a maximum tree depth of 5 (max_depth = 5). The learning rate was set to 0.1 to balance convergence speed and predictive stability. The subsampling ratio (subsample) and column sampling ratio (colsample_bytree) were not explicitly modified and therefore retained their default value of 1.0. Regularization parameters were kept at default settings, with L1 regularization (reg_alpha = 0) and L2 regularization (reg_lambda = 1). A fixed random state of 42 was used to ensure reproducibility. Lastly, the hybrid model is evaluated through various evaluation metrics (accuracy, precision, recall, F1-score, and receiver operating characteristic curve-area under curve (ROC–AUC)). This comprehensive methodology ensures the development of a robust and behaviorally discriminative fraud detection framework.

3.1. Dataset description

The study uses a publicly available simulated financial transactions dataset (PS_20174392719_1491204439457_log.csv) containing 6,362,620 records with 11 attributes. Each entry includes parameters such as transaction type, amount, account balances, and binary fraud indicators (isFraud and isFlaggedFraud). To ensure analytical rigor, the *isFlaggedFraud* variable was excluded due to its extreme sparsity.

Exploratory data analysis indicates that fraudulent transactions in the dataset are observed only within the TRANSFER category, whereas CASH_IN, CASH_OUT, and DEBIT contain no fraud instances. Accordingly, the analysis focuses on TRANSFER transactions. This restriction prevents the model from trivially classifying fraud based on transaction type—a superficial pattern that would exacerbate class imbalance and inflate performance metrics.

Instead, it ensures the model identifies meaningful behavioral patterns. After filtering, the dataset contained 2,684,404 records. Furthermore, the account identifiers ‘nameOrig’ and ‘nameDest’ were removed to prevent identity leakage and improve the model generalization.

3.2. Feature engineering

To improve the model's predictive power, several domain-specific features are engineered. Temporal features are created by calculating the transaction hour from the step variable (hour = step % 24) and by determining whether the transaction occurred during nighttime hours (is night = 1 if the hour is less than 6; otherwise, 0). A proportional feature, amount ratio, is computed as the transaction amount divided by the sender's original balance to normalize transaction sizes. Additionally, balance change features are generated, specifically sender balance change and receiver balance change. Two binary indicators, orig balance zero and dest balance zero, are added to identify cases where the initial balance is zero. Finally, one-hot encoding is used for the transaction type. The final dataset consists of 14 numerical features for ML.

3.3. Data partitioning

The processed dataset is partitioned into training and test sets using a stratified 75:25 split to preserve the original class distribution. The training set contains 2,013,303 records, while the testing set contains 671,101 records. The split is performed at the transaction level rather than at the account (sender/receiver) level. Although account identifiers ('nameOrig' and 'nameDest') are removed to prevent direct identity leakage, transactions associated with the same sender or receiver may still appear in both training and testing subsets. This design evaluates generalization across transaction behavior patterns rather than unseen account identities; however, future work will incorporate strict group-based (account-level) partitioning to assess out-of-entity robustness.

To prevent data leakage, the split is performed before model training and evaluation. All preprocessing and feature engineering steps are derived exclusively from the training data and subsequently applied to the test set without recalculation. The autoencoder is trained only on legitimate transactions within the training subset. Reconstruction errors for the test set were computed using the trained model without retraining or parameter adjustment. The test data remains completely unseen during model training and hyperparameter tuning.

Currently, this experimental setup uses a static train–test split and does not explicitly model temporal distribution shifts. Consequently, concept drift is not evaluated in this study, and the framework is assessed under a stationary data assumption due to the static nature of the simulated dataset. The impact of evolving fraud patterns and non-stationary data distributions will be investigated in future work using time-ordered splits, sliding-window retraining strategies, and drift-aware evaluation protocols

3.4. Hyperparameter optimization

The XGBoost hyperparameters (n_estimators = 100, max_depth = 5, learning_rate = 0.1) are selected based on standard configurations for gradient boosting models and are not optimized using the test dataset. No hyperparameter tuning is performed via cross-validation on the test split. This approach ensures that the reported results are free from test data bias and reflect the baseline performance of the proposed hybrid framework. While these default settings provide stable and reliable performance, they may not represent the optimal configuration for the dataset. Future research will incorporate systematic optimization techniques, such as grid search or bayesian optimization to further enhance model performance and generalizability.

3.5. Evaluation metrics

The hybrid model's performance is assessed using accuracy, precision, recall, F1-score, and area under the ROC–AUC. Accuracy provides an overall measure of correctly classified transactions, while precision evaluates the proportion of correctly identified fraud cases among all predicted frauds. Recall measures the model's ability to detect actual fraudulent transactions, which is critical for minimizing financial losses. The F1-score balances precision and recall, offering a comprehensive

performance metric, especially under class imbalance conditions. Additionally, a confusion matrix was used to analyze classification errors in detail, and the ROC curve is plotted to illustrate the trade-off between true positive and false positive rates across different thresholds.

4. Results

The performance of the proposed hybrid model is exceptionally high, as summarized in Table 1. The discriminative strength of the engineered features is first evident in the feature correlation heatmap shown in Fig. 2, highlighting strong correlations with key variables: amount ratio, sender balance change, and receiver balance change. This confirms that features designed to capture irregular fund transfers and sudden balance shifts play a crucial role in achieving high detection accuracy.

The robustness of the model is further validated by the ROC presented in Fig. 3, which exhibits an AUC close to 1.0. This indicates the model's excellent ability to distinguish between fraudulent and legitimate transactions, achieving a high true positive rate while maintaining a very low false positive rate. Finally, the confusion matrix illustrated in Fig. 4 demonstrates near-perfect classification performance on a test dataset comprising 671,101 transactions. The model correctly classifies 670,053 legitimate transactions with only one false positive and successfully detects 1,042 out of 1,047 fraudulent cases, resulting in just five false negatives.

Table 1 Performance metrics of the hybrid model

Metric	Fraudulent class	Non-fraudulent class	Overall/Score
Accuracy	1.00	1.00	1.0000
Precision	1.00	1.00	-
Recall	1.00	1.00	-
F1-score	1.00	1.00	-
ROC–AUC	-	-	0.99999995
Correctly classified	1,047	670,054	671,101

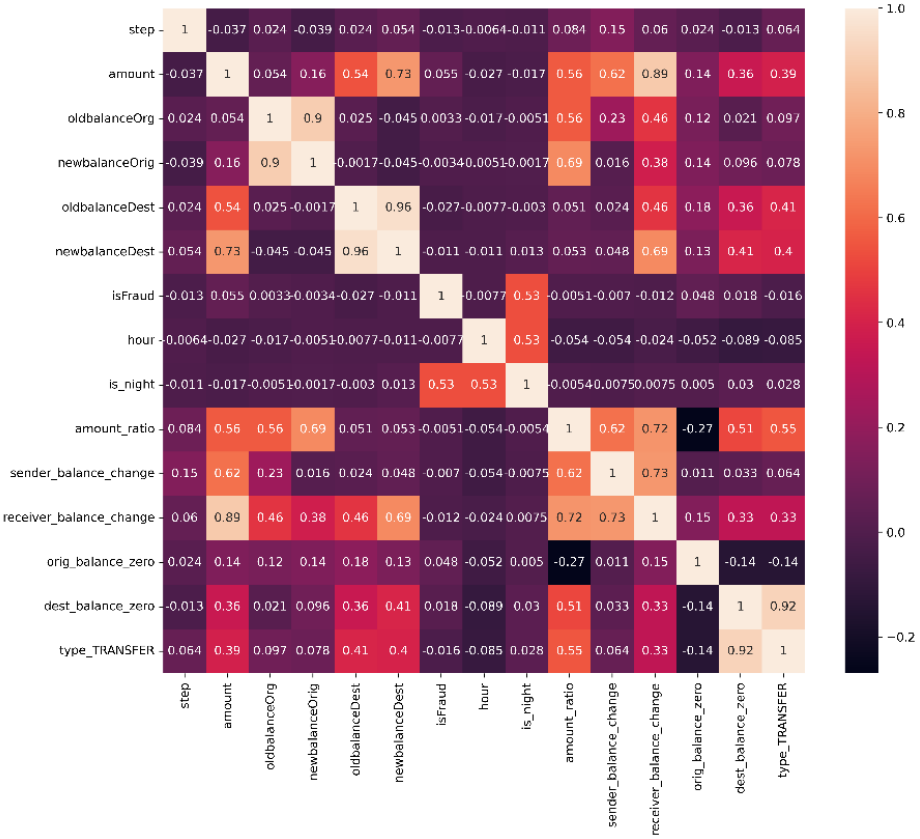

Fig. 2 Feature correlation and fraud association analysis

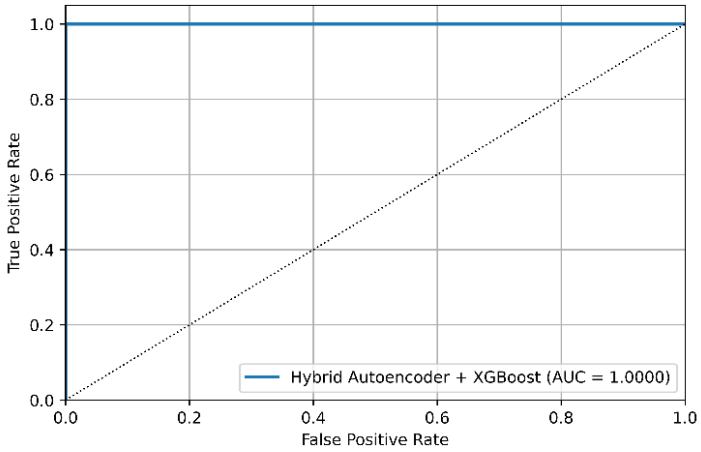

Fig. 3 ROC for hybrid autoencoder+XGBoost

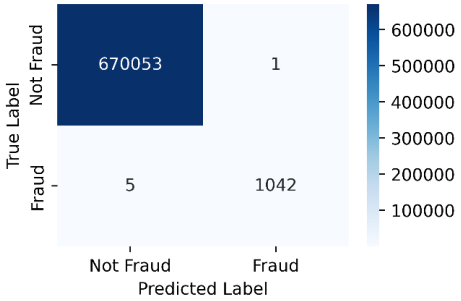

Fig. 4 Confusion matrix for autoencoder+XGBoost

To quantify the contribution of the autoencoder-derived reconstruction_error feature, an additional baseline experiment is conducted using an XGBoost classifier trained solely on the engineered transactional features, excluding the anomaly score. The baseline XGBoost model achieves an overall accuracy of 1.0000 and a ROC–AUC of approximately 0.999999, with classification performance nearly identical to that of the hybrid model. When the reconstruction_error feature is incorporated, the hybrid autoencoder–XGBoost framework achieved an accuracy of 1.0000 and a ROC–AUC of 0.99999995, as reported in Table 1. The negligible performance difference indicates that deterministic balance-related features in the simulated dataset already provide near-perfect class separability. Therefore, under these controlled conditions, the anomaly-aware enrichment contributes only marginal incremental discriminative improvement. This finding aligns with the structural simplicity of the simulated fraud patterns.

5. Discussion

The near-perfect performance of the hybrid framework must be understood in light of its primary limitation: the use of a simulated dataset. The outstanding results are directly influenced by deterministic fraud signatures embedded in the data. A closer analysis reveals that fraudulent 'TRANSFER' transactions consistently follow a straightforward, unchanging rule: the originator's account is drained to zero (oldBalanceOrig – amount = 0), while the recipient's balance stays the same (newBalanceDest = oldBalanceDest). The crafted features, such as sender balance change and destination balance zero, are intentionally created to detect these patterns, making the classes nearly perfectly separable for the XGBoost classifier. Therefore, the observed performance should not be interpreted as conclusive evidence of real-world robustness. Instead, it represents a consequence of deterministic patterns in the synthetic dataset, which can inflate performance metrics and mask the true complexity of real-world fraud scenarios.

Importantly, the near-perfect performance does not arise from data leakage or improper experimental design. The dataset was partitioned using a stratified 75:25 train–test split prior to model training, with all preprocessing and feature engineering performed exclusively on the training data and subsequently applied to the test set without recalculation. The autoencoder was

trained only on legitimate transactions within the training subset, and the test data remained completely unseen during model fitting. Furthermore, the XGBoost hyperparameters (n_estimators = 100, max_depth = 5, learning_rate = 0.1) were fixed and not tuned using the test set. Thus, the high performance reflects deterministic structural separability in the simulated dataset rather than experimental leakage.

While this outcome strongly supports the proposed feature engineering approach, it also shows that simple, uniform fraud patterns are not representative of real-world situations, where fraudulent tactics are varied, subtle, and constantly evolving. Moreover, the model's impressive performance serves primarily as a proof of concept for the hybrid architecture rather than a direct measure of its effectiveness in a live setting. Generalizability can only be confirmed by testing the framework on authentic, anonymized transaction data from financial institutions. Additionally, the framework is not experimentally benchmarked for inference latency, throughput capacity, or computational efficiency. Consequently, any reference to real-time readiness should be interpreted as architectural feasibility rather than empirically validated deployment performance. Rigorous latency benchmarking and throughput measurement under realistic transaction loads are necessary before confirming operational suitability.

Although concept drift is a critical challenge in real-world UPI fraud detection systems, the present study does not explicitly model or evaluate temporal distribution shifts because the employed dataset is static and simulated. As a result, the proposed methodology is assessed under a stationary data assumption, and its robustness under evolving fraud patterns remains untested. This represents an important limitation of the current work. Future research will incorporate drift-aware evaluation strategies such as chronological train–test splits, rolling-window validation, incremental retraining, and drift detection or synthetic drift injection experiments to assess model stability under non-stationary fraud distributions.

With respect to emerging threats such as zero-day fraud strategies and long-horizon social engineering schemes (e.g., pig butchering scams), the proposed framework is conceptually supported by its anomaly-aware design. Since the autoencoder is trained exclusively on legitimate transactions, transaction patterns that deviate from learned normal behavior—even if they correspond to previously unseen or novel fraud strategies are expected to yield elevated reconstruction errors. These anomalies can therefore be highlighted as suspicious by the downstream classifier. This provides a degree of resilience to previously unseen attack patterns that are not explicitly represented in the training data. However, the present study does not include dedicated zero-day or pig-butchering-style attack simulations, and systematic empirical evaluation against such emerging threats remains an important direction for future work.

Additionally, claims regarding real-time readiness remain preliminary. The study does not experimentally evaluate inference latency, computational load, memory consumption, or energy efficiency under operational throughput constraints. Recent work by Bakirci (2025) [30] proposes a systematic methodology for assessing real-time AI deployment by jointly analyzing inference latency, computational burden, and energy efficiency in resource-constrained environments. Such trade-off analysis is highly relevant for large-scale UPI fraud detection systems, where millisecond-level response times and high transaction throughput are essential. Incorporating similar performance evaluation methodologies in future validation of the proposed framework would provide a more comprehensive assessment of its practical deployability and operational robustness.

Fraudulent cases frequently exhibit specific balance patterns, such as the sender’s post-transaction balance being zero and the receiver’s balance remaining unchanged. These patterns are effectively captured by the engineered features, especially the amount ratio, sender balance change, receiver balance change, and destination balance zero. The integration of the autoencoder’s reconstruction error further enhanced the model’s discriminatory power by providing an anomaly-focused signal.

To examine whether the model relies primarily on deterministic balance-change heuristics, a gain-based feature importance analysis is conducted using the trained XGBoost classifier. The results indicated that sender balance change, amount ratio, receiver balance change, and dest balance zero are the dominant predictors, confirming that balance-related

features drive much of the class separability in the simulated dataset. Importantly, the autoencoder-derived reconstruction error also demonstrated measurable importance, indicating that anomaly-aware feature enrichment contributed additional discriminative information beyond explicit handcrafted rules. This analysis clarifies that while structural heuristics strongly influence separability under simulated conditions, the hybrid framework integrates both engineered financial signals and learned anomaly representations in its decision-making process.

From a regulatory and governance perspective, recent work, such as "Embedding Accountability in the AI Lifecycle for Critical Finance Applications" emphasizes the need for auditable AI frameworks in high-stakes financial systems, including fraud detection. In particular, mechanisms such as decision traceability logs and compliance rules engines are highlighted as key for ensuring transparency, auditability, and regulatory alignment. Incorporating these accountability-oriented design principles strengthens the framework's suitability for deployment in regulated digital banking environments and aligns it with emerging financial AI governance requirements.

Despite the promising experimental outcomes, the proposed framework cannot be considered production-ready at this stage. The absence of validation on real-world UPI transaction data, the lack of temporal drift evaluation, and the absence of comprehensive operational benchmarking (e.g., latency, throughput, and energy profiling) limit its immediate deployability within live financial infrastructures. The current findings should therefore be interpreted as research-stage validation of methodological feasibility rather than confirmation of readiness for integration into production-grade UPI monitoring systems.

Deployment of AI-based fraud detection systems in operational UPI environments involves multiple practical and regulatory constraints. These include strict data privacy and compliance requirements, the need for continuous monitoring to address concept drift and potential adversarial adaptation by fraudsters. Moreover, scalability under high transaction throughput, and the risk of false positives affecting legitimate users. Additionally, explainability, auditability, and regulatory transparency are essential for financial dispute resolution and supervisory oversight. Without robust monitoring, automated retraining pipelines, and stress testing under real transaction loads, premature deployment may introduce financial, operational, and reputational risks.

Table 2 Contextual overview of selected UPI fraud detection studies

Study/ year	Dataset Name	Model	Accuracy	Findings
R. U. et al. (2023) [7]	UPI transactions dataset. (96,545)	M-DBN	98.4%	robust in detecting UPI QR-code & UPI-ID fraud in real time
Rani et al. (2024) [8]	Labeled UPI transaction dataset (timestamps, amount, user information, etc.)	XGBoost	98.2%	include data, compliance, biometrics, and AI.
Rethisha et al. (2025) [9]	Real UPI transaction data	LR, RF, DT, NB	LR: 97% RF: 98.3% DT: 89% NB: 84%	RF achieved the highest accuracy with strong recall & F1; effective in minimizing false positives
Raju et al. (2024) [10]	Online transactional dataset	LSTM	99.74%	combining reinforcement learning, extending datasets, and conducting continuous model updates to better fraud detection.
Naikl et al. (2024) [11]	Custom / Synthetic UPI transaction dataset	CNN	97.68%	fraud detection with ML, scalability, time, and location analysis.
Bharath et al. (2024) [12]	Financial transaction dataset	HMLM	98.61%	real-time deployment, handling device fraud types, adapting to evolving threats, and integrating with the national payment system.
Proposed Hybrid autoencoder + XGBoost (This Work)	Simulated transaction dataset (2.68M)	(Hybrid Model) AE+ XGBoost	= 1.00	hybrid anomaly-aware framework evaluated on a large-scale simulated dataset; results reflect structural separability under controlled conditions.

Table 2 presents a contextual overview of selected recent UPI fraud detection studies to situate the proposed framework within the broader research landscape. It is important to note that these studies differ substantially in terms of dataset characteristics, sample sizes, fraud definitions, preprocessing strategies, and evaluation protocols. Hence, the reported performance metrics are not directly comparable and should not be interpreted as strict benchmark comparisons. Instead, the table highlights the diversity of modeling approaches employed in the domain and illustrates the range of performance outcomes reported under heterogeneous experimental conditions. The proposed hybrid autoencoder + XGBoost model is therefore positioned as a methodological contribution rather than a competitively benchmarked alternative across standardized datasets.

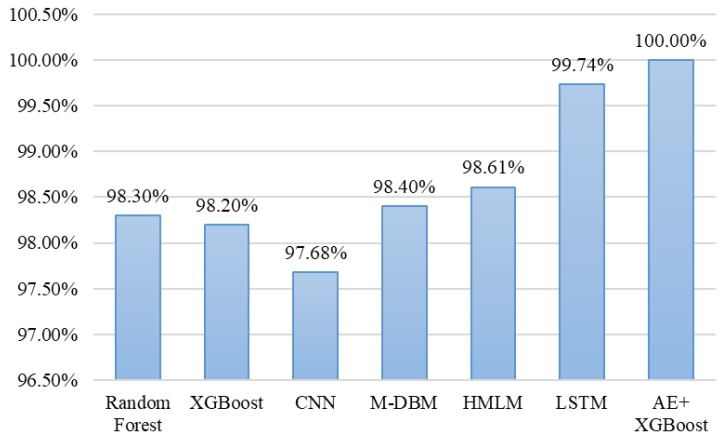

Fig. 5 Accuracy of Classification Models for UPI Fraud Detection

Fig. 5 presents the accuracy of various ML, DL, and hybrid models for UPI fraud detection. The visualization demonstrates a clear improvement trend in performance, where advanced models achieve higher accuracy than traditional ML approaches. However, these results are derived from different datasets and methodologies. Hence, the figure should be interpreted as a contextual comparison rather than a direct benchmark.

6. Conclusion

A hybrid autoencoder–XGBoost framework for detecting fraudulent UPI transactions was developed and evaluated. The approach combines supervised XGBoost classification with unsupervised autoencoder-based anomaly detection to handle highly imbalanced transaction data. Using domain-driven feature engineering and anomaly-aware feature enrichment on a large-scale simulated dataset, the model enhances fraud detection performance. Its key contribution lies in integrating reconstruction errors from legitimate-only training as structured inputs into the boosting model, improving discrimination without relying on synthetic resampling techniques. The main findings of the study are summarized as follows:

(1) The hybrid autoencoder–XGBoost model achieved near-perfect performance (ROC-AUC 0.99999995, accuracy and F1-score of 1.00) on the simulated dataset. This outcome is driven by deterministic fraud patterns and domain-engineered features that create near-separable classes.

(2) Without using resampling methods like SMOTE or Tomek linkages, class imbalance was handled by using autoencoder-generated reconstruction errors as anomaly scores, which improved model stability.

(3) The significance of contextual feature design was highlighted by the discovery that domain-specific characteristics such as sender balance change, receiver balance change, and transaction amount ratio were the most influential predictors in identifying fraudulent conduct.

(4) The framework shows computational feasibility under simulated high-volume conditions; however, real-time deployment claims remain preliminary due to the absence of latency, throughput, and energy benchmarking, requiring further real-world validation and stress testing.

(5) Although results are strong, they reflect a proof-of-concept validation on simulated data rather than real-world complexity. Validation on anonymized, authentic UPI datasets is essential for confirming generalizability.

(6) Access to real-world UPI transaction data is restricted due to privacy regulations and institutional confidentiality constraints; therefore, collaboration with financial institutions for anonymized external validation represents an important direction for future research.

(7) The study used transaction-level splitting, which may allow overlap of account behavior across datasets despite removing identifiers. Future work should adopt group-based partitioning to ensure robustness under unseen account-level conditions and avoid cross-entity overlap.

(8) Concept drift was not evaluated due to the static dataset; future work will incorporate drift-aware validation using time-based data, sliding-window retraining, and drift detection to ensure robustness under evolving fraud patterns.

Conflicts of Interest

The authors declare no conflict of Interest.

References

[1] J. Putrevu and C. Mertzanis, "The Adoption of Digital Payments in Emerging Economies: Challenges and Policy Responses," Digital Policy, Regulation and Governance, vol. 26, no. 5, pp. 476-500, 2024.

[2] S. M. Mathew, S. Shanimon, S. Joseph, and M. Abraham, "Exploring the Exponential Growth of UPI in India: A Study on Digital Payment Transformation (2016–2024)," Library Progress International, vol. 44, no. 3, pp. 10796-10805, 2024.

[3] M. Razi-ur-Rahim, M. R. Rabbani, F. Uddin, and Z. H. Shaikh, "Adoption of UPI Among Indian Users: Using Extended Meta-UTAUT Model," Digital Business, vol. 4, no. 2, article no.100093, 2024.

[4] D. Devassy, D. K. S, G. P. Mannickathan, A. Biju, A. T, and D. R, "FedUPI: Federated Learning Empowered Detection of Fraudulent UPI Transactions," Proceedings of the 5th International Conference on Soft Computing for Security Applications (ICSCSA), IEEE, pp. 1007-1015, 2025.

[5] A. Aljaradat, G. Sarkar, and S. K. Shukla, "Modelling Cybersecurity Impacts on Digital Payment Adoption: A Game Theoretic Approach," Journal of Economic Criminology, vol. 5, article no. 100089, 2024.

[6] N. B. Chakka and S. S. Saheb, "Mobile Payment Fraud Detection in UPIs through Machine Learning Techniques: A Systematic Review," Multidisciplinary Reviews, vol. 9, no. 6, article no. 2026280, 2025.

[7] R. U., M. P. Raj, J. N. Mithra, S. S. Balaji, L. A. Narayanan, and J. M. D. Y., "A Robust UPI Fraud Identification Scheme over Digital Money Transactions Using Learning Powered Classification Principles," Proceedings of International Conference on Electronics and Renewable Systems (ICEARS), IEEE, pp. 1551-1558, 2025.

[8] R. Rani, A. Alam, and A. Javed, "Secure UPI: Machine Learning-Driven Fraud Detection System for UPI Transactions," Proceedings of 2nd International Conference on Disruptive Technologies (ICDT-2024), IEEE, pp.924-928, 2024.

[9] R. Rethisha, S. Cyindia H., and R. Geetha, "Leveraging Machine Learning Techniques of Real Time Detection of UPI Fraud," Proceedings of International Conference on Intelligent Sustainable Systems (ICISS), IEEE, pp. 1506-1510, 2025.

[10] M. N. Raju, Y. C. Reddy, P. N. Babu, V. S. P. Ravipati, and V. Chaitanya, "Detection of Fraudulent Activities in Unified Payments Interface using Machine Learning - LSTM Networks," Proceedings of 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), IEEE, pp.769-774, 2024.

[11] S. K. L. Naikl, A. Kiran, V. P. Kumar, S. Mannam, Y. Kalyani, and M. Silparaj, "Fraud Fighters - How AI and ML are Revolutionizing UPI Security," Proceedings of International Conference on Science Technology Ingineering and Management (ICSTEM), IEEE, pp. 1-7, 2024.

[12] S. Bharath, G. L. Vara Prasad, V. Sujatha, S. Hemajothi, D. Sharada Mani, and N. G. Merlin, "HMLM: An Intelligent Artificial Intelligence Assisted Strategy to Identify UPI Frauds Based on Hybrid Markov Learning Methodology," Proceedings of International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), IEEE, pp. 1-6, 2024.

[13] S. Janbhasha, C. H. N. S. Kumar, V. Sitharamulu, B. N. V. Madhu Babu, H. R. Battu, and K. Venkataramana, "A Hybrid Approach for Fraud Detection in Digital Wallet Transactions Using Adversarial Autoencoders and Gated Recurrent Units," Engineering, Technology and Applied Science Research, vol. 15, no. 4, pp. 25532-25537, 2025.

[14] K. Prakash, M. Franklin, M. Shunmugasundaram, K. Sankar Ganesh, and S. Gangadharan, "Fraud Detection in the Banking Sector Using Gated Green Anaconda Progressive Generative Axial Adversarial Attention Network," Intelligent Decision Technologies, vol. 19, no. 5, pp. 3281-3303, 2025.

[15] N. Singh and S. Tripathy, "FedShield: Federated Learning Based Robust Online Payment Fraud Detection," Journal of Supercomputing, vol. 81, no. 13, article no. 1289, 2025.

[16] O. Manta, V. Vasile, and E. Rusu, "Banking Transformation Through FinTech and the Integration of Artificial Intelligence in Payments," FinTech, vol. 4, no. 2, article no.13, 2025.

[17] A. Hamed, Y. Bansal, M. Mohamad, A. Jimenez-Aranda, and T. Gaber, "AI Security in Contactless Payments and Education: A Review," Journal of Engineering Education Transformations, vol. 39, no. s1, pp. 20-25, 2025.

[18] F. M. Talaat, T. Medhat, and W. M. Shaban, "Precise Fraud Detection and Risk Management with Explainable Artificial Intelligence," Neural Computing and Applications, vol. 37, no. 23, pp. 19199-19229, 2025.

[19] C. G. Arnaldo, R. D. A. Jurado, F. P. Moreno, and M. Z. Suárez, "Enhancing Security in Airline Ticket Transactions: A Comparative Study of SVM and LightGBM," Applied Sciences, vol. 15, no. 17, article no. 9581, 2025.

[20] G. Prabu, A. J. F. Prakash, E. R. Joseph, S. M. Poonkuzhali, R. Ramya, and S. Subha, "Unlocking the Power of UPI: A Machine Learning Journey Through Digital Payment and Data Visualization Models," Proceedings of the International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), pp. 1-7, 2025.

[21] R. El Youbi, F. Messaoudi, M. Loukili, and R. Loukili, "From Click to Checkout: Deep Learning for Real-Time Fraud Detection in E-Payment Systems," Statistics, Optimization & Information Computing, vol. 14, no. 6, pp. 3398-3408, 2025.

[22] H. M. R. Al Lawati, A. Zainal, B. A. S. Al-Rimy, M. A. N. Al-Azawi, M. N. Kassim, S. A. Almalki, et al., "An Integrated Preprocessing and Drift Detection Approach With Adaptive Windowing for Fraud Detection in Payment Systems," IEEE Access, vol. 13, pp. 92036-92056, 2025.

[23] M. S. A. Mozumder, M. B. H. Sakil, M. R. Hasan, M. A. Hasan, K. M. N. R. Fuad, M. F. Mridha, et al., "Hybrid Contrastive Learning With Attention-Based Neural Networks for Robust Fraud Detection in Digital Payment Systems," IEEE Open Journal of the Computer Society, vol. 6, pp. 1053-1064, 2025.

[24] H. N. Himabindu Gurajada and R. Autade, "Quantum Computing for Fraud Detection in Real-Time Payment Systems," Proceedings of International Conference on Computing Technologies & Data Communication (ICCTDC), IEEE, pp. 1-8, 2025.

[25] M. Tamilselvi, R. Begum, K. K, J. Giri, D. Sheela, and M. O. Sabri, "Experimental Evaluation Unified Payment Interface (UPI) Fraud Detection System Using Elevated Deep Learning Methodology" Proceedings of the International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), IEEE, pp. 40-45, 2024.

[26] A. A. Abdirahman, A. O. Hashi, U. M. Dahir, M. A. Abdi, and O. E. R. Rodriguez, "Enhancing Security in Mobile Wallet Payments: Machine Learning-Based Fraud Detection Across Prominent Wallet Platforms," International Journal of Electronics and Communication Engineering, vol. 11, no. 3, pp. 96-105, 2024.

[27] T. Madhavappa and B. Sathyanarayana, "An Efficient Framework Based on Optimized CNN-RNN for Online Transaction Fraud Detection in Financial Transactions," International Journal of System Assurance Engineering and Management, vol. 10, no. 53s, pp. 689-713, 2025.

[28] S. Renukadevi, B. C. Manujakshi, T. M. Shashidhar, and N. Sivakumar, "Fraud Detection in Financial Transactions Using Gradient Boost with Hybrid Optimization," Journal of Machine and Computing, vol. 5, no. 4, pp. 2328-2344, 2025.

[29] C. Dong and S. Xiao, "Enhancing Financial Fraud Detection in Digital Finance Applications Through Machine Learning Algorithms and Real-Time Data Analytics," Journal of Computational Methods in Sciences and Engineering, article no.14727978251352131, 2025.

[30] M. Bakirci, "Performance Evaluation of Low-Power and Lightweight Object Detectors for Real-Time Monitoring in Resource-Constrained Drone Systems," Engineering Applications of Artificial Intelligence, vol. 159, part B, article no. 111775, 2025.