

AI-Based Detection of Phishing Attacks in Digital Payment Systems Using Explainable Machine Learning

Prabhat Bisht

Affiliation: Joint Director (IT), The National Informatics Centre (NIC), Chandigarh, Haryana, India
Correspondence: prabhatbisht@gmail.com

Abstract

India's rapid adoption of Unified Payments Interface (UPI), mobile banking, digital wallets, and other digital payment channels has expanded both financial inclusion and the attack surface available to cybercriminals. AI-enabled phishing, voice cloning, facial manipulation, and identity impersonation increasingly challenge conventional rule- and signature-based security mechanisms. This paper proposes an integrated, explainable machine-learning framework for detecting phishing and deepfake-related fraud in digital payment environments. The framework combines **transaction-behavior analysis**, phishing indicators, multimodal deepfake evidence, and an Explainable AI (XAI) layer based on SHAP and LIME. The data-processing pipeline includes multi-source collection, schema harmonization, validation, duplicate removal, leakage-aware preprocessing, feature engineering, stratified splitting, model training, hyperparameter optimization, and cross-dataset evaluation. XGBoost is positioned as a primary tabular fraud-detection model, while CNN/LSTM/transformer-based components are proposed for visual, temporal, audio, and phishing-content analysis. Because the supplied manuscript does not report completed experimental measurements, numerical performance figures are treated as targets rather than observed results. The proposed framework is intended to improve detection effectiveness, interpretability and operational trust while providing a scalable basis for future validation in Indian digital-payment ecosystems.

Keywords: Deepfake Detection; Phishing Detection; Explainable Artificial Intelligence; Machine Learning; XGBoost; UPI Fraud; Digital Payments; SHAP; LIME; Financial Cybersecurity

1. Introduction

India has undergone rapid digitization of retail financial services through UPI, mobile banking, QR-code payments, digital wallets, and related electronic payment channels. This transformation has improved transaction convenience and financial inclusion, but it has also created a large and heterogeneous digital environment in which fraudulent behavior can emerge across devices, accounts, networks, merchants, and user interactions.

Phishing remains a major social-engineering threat because attackers attempt to induce users to disclose credentials, one-time passwords, payment information, or other sensitive data. At

the same time, advances in generative AI have made synthetic or manipulated audio, video, and facial content increasingly convincing. These capabilities can support identity impersonation and may weaken authentication processes when biometric evidence is considered without sufficient behavioral and contextual verification.

Conventional security mechanisms based primarily on fixed rules, signatures, or manually defined thresholds can struggle with adaptive attacks. Machine-learning methods can learn complex relationships among transaction, temporal, device, geographic, and behavioral variables. However, high predictive performance alone is insufficient for financial-security applications: analysts, institutions,

users, and regulators also need to understand why a transaction or identity-verification event was classified as suspicious.

This study therefore proposes an integrated and explainable framework that combines fraud and phishing detection with deepfake-oriented multimodal analysis. The central design principle is that model predictions should be accompanied by interpretable evidence, enabling a risk engine to support automated screening while retaining a pathway for human review.

2. Problem Statement

Existing digital-payment fraud systems face five related challenges: (i) limited ability to identify AI-generated or manipulated identities; (ii) rapidly changing phishing tactics; (iii) false positives that can inconvenience legitimate users; (iv) limited transparency of complex machine-learning decisions; and (v) insufficient integration between identity/biometric verification and transaction-level behavioral monitoring.

An integrated system is required that can combine heterogeneous evidence, adapt to evolving attack patterns, reduce unnecessary alerts, and explain the factors contributing to a decision. The problem is especially important in payment environments where a security decision may affect a legitimate transaction and therefore requires both technical reliability and operational accountability.

3. Research Objectives

- Develop a machine-learning framework for phishing and financial-fraud detection in digital-payment environments.
- Identify indicators associated with AI-generated or manipulated audio, video, and facial-authentication attempts.
- Integrate SHAP and LIME to provide local and global explanations of model decisions.
- Improve detection quality while explicitly monitoring false positives and class imbalance.
- Design a scalable architecture suitable for heterogeneous Indian banking and UPI-related environments.
- Establish a leakage-aware evaluation procedure that supports reproducible comparison across datasets.

4. Literature Review

The rapid adoption of the Unified Payments Interface (UPI) has transformed India's digital payment ecosystem by enabling instant, interoperable, and convenient fund transfers. However, the growing volume, frequency, and velocity of UPI transactions have also increased exposure to fraudulent activities such as phishing, social engineering, unauthorized transactions, fraudulent payment requests, QR-code scams, account takeover, and the misuse of mule accounts. Consequently, recent research has increasingly focused on the use of artificial intelligence (AI) and machine learning (ML) techniques to identify anomalous transaction behaviour and enable timely fraud detection.

Early research in financial fraud detection primarily focused on supervised machine learning algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and Gradient Boosting. These methods generally classify transactions on the basis of historical transaction attributes and behavioural patterns. Existing reviews of ML-based financial fraud detection have identified SVM and Artificial Neural Networks as commonly investigated techniques, with credit-card fraud representing one of the most extensively studied application domains. Although this body of work demonstrates the maturity of conventional ML-based financial fraud detection, UPI-specific fraud detection remains a comparatively recent and evolving research area.

Several studies have subsequently applied conventional ML techniques specifically to UPI transactions. Bodade and Pawade [1] examined a UPI fraud-detection framework incorporating transaction patterns, user behaviour, and device-related information, while considering multiple classification and anomaly-detection approaches. Their work emphasizes the importance of combining transactional and behavioural information rather than relying solely on isolated attributes such as transaction amount or payment status.

More recent studies have investigated ensemble learning, based on the premise that individual classifiers may not adequately capture the heterogeneous characteristics of fraudulent transactions. Kamra et al. [2] evaluated Decision Tree, SVM, Random Forest, and Logistic Regression models using a synthetic and imbalanced UPI transaction dataset. Random Forest achieved an accuracy of 0.92,

while an ensemble combining Random Forest, Logistic Regression, and SVM achieved an accuracy of 0.93 and an F1-score of 0.878. These findings indicate that combining complementary classifiers can enhance predictive performance, particularly where fraudulent behaviour exhibits diverse and nonlinear patterns.

Another significant research direction involves anomaly detection and unsupervised learning. Conventional supervised approaches depend heavily on labelled fraudulent transactions; however, emerging fraud patterns may not be adequately represented in historical training data. Accordingly, hybrid approaches combining supervised and unsupervised learning have received increasing attention for detecting behavioural deviations from established transaction patterns. The UPI Sentinel framework proposed by Sowmiya et al. [3], for example, focuses on behavioural anomaly detection in real-time UPI transactions. The study highlights the limitations of static rule-based systems in identifying evolving fraud behaviours associated with unusual transaction times, unfamiliar locations, newly linked devices, and abnormal transaction amounts.

Deep learning has also emerged as a promising approach for UPI fraud detection. Recent studies have investigated architectures such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Autoencoders. These architectures are capable of modelling complex nonlinear relationships and temporal dependencies that may be difficult to capture using conventional ML algorithms. In particular, CNN-, RNN-, and Autoencoder-based models have been explored for anomaly detection and behavioural modelling in large-scale UPI transaction environments [4].

Hybrid ML and deep-learning architectures are particularly relevant because fraud detection requires both anomaly identification and transaction-level classification. Chakka and Saheb [5] proposed an Autoencoder-XGBoost framework in which the Autoencoder learns representations of normal transaction behaviour and generates reconstruction-error-based anomaly scores. These anomaly scores are subsequently incorporated as additional features into an XGBoost classifier for fraud classification. The framework was evaluated using a synthetic dataset of approximately 2.68 million transactions, illustrating the increasing research emphasis on scalable hybrid architectures capable of processing large transaction volumes.

Similarly, recent work has explored combinations of CNN, Autoencoder, Local Outlier Factor (LOF), and K-Means clustering for UPI fraud detection [6]. Such hybrid approaches integrate supervised classification with unsupervised anomaly detection and are particularly relevant when labelled fraudulent transactions are scarce or when novel fraud strategies continuously emerge.

A further area of interest is real-time and adaptive fraud detection. Conventional rule-based systems rely primarily on predefined thresholds and known fraud signatures. Although such mechanisms can detect established fraud patterns, they may generate excessive false positives and often exhibit limited capability in identifying previously unseen or rapidly evolving fraudulent behaviour. Consequently, adaptive ML-based systems have been proposed to learn changing transaction patterns and generate real-time or near-real-time alerts when suspicious behaviour is detected [7].

The literature also demonstrates increasing interest in Explainable Artificial Intelligence (XAI) for financial fraud detection. High-performing models such as XGBoost, Random Forest, neural networks, and hybrid deep-learning architectures can function as relatively opaque decision-making systems. This presents an important challenge in financial applications, where fraud analysts, payment institutions, and other stakeholders may require clear explanations of the factors contributing to a fraud prediction. Recent UPI-related research has therefore considered techniques such as SHAP (Shapley Additive Explanations) and feature-importance analysis to quantify the contribution of individual transaction attributes to model predictions [8]. Incorporating explainability can improve transparency, facilitate model validation, and support the practical adoption of AI-based fraud-detection systems.

Another emerging research direction involves identifying fraudulent accounts and mule-account networks, rather than analysing individual transactions in isolation. Recent studies have combined techniques such as Gradient Boosted Decision Trees, Graph Neural Networks (GNNs), and LSTM networks to analyse transaction behaviour, fund-flow relationships, and temporal patterns for identifying mule accounts and potentially fraudulent entities [9]. This reflects a broader transition from conventional transaction-level fraud detection toward entity-level,

relationship-aware, and graph-based financial intelligence, in which interactions among accounts and transactions are analysed alongside individual transaction characteristics.

Despite considerable progress in UPI fraud detection, several important research gaps remain.

First, the limited availability of authentic UPI fraud datasets remains a major methodological challenge. UPI transaction data are highly sensitive and are generally not publicly available because of privacy, security, and financial confidentiality considerations. Consequently, many existing studies rely on synthetic or simulated datasets. Recent work, for example, has evaluated fraud-detection approaches using synthetic and imbalanced UPI transaction data [10]. Although such datasets are useful for algorithm development and methodological experimentation, they may not fully represent the complexity, diversity, and behavioural dynamics of real-world UPI transactions. This raises concerns regarding the generalizability and external validity of reported model performance.

Second, excessive reliance on classification accuracy can produce misleading conclusions in highly imbalanced fraud-detection settings. Legitimate transactions substantially outnumber fraudulent transactions, meaning that a model may achieve high overall accuracy while failing to identify a meaningful proportion of actual fraud cases. Therefore, performance evaluation should extend beyond accuracy to include precision, recall, F1-score, ROC-AUC, and particularly Precision–Recall AUC (PR-AUC). In operational fraud detection, minimizing false negatives is especially important, while an acceptable false-positive rate must also be maintained to avoid unnecessary disruption of legitimate transactions.

Third, comprehensive feature integration remains insufficiently explored. Existing studies frequently focus on transaction-level attributes such as transaction amount, time, location, device information, transaction frequency, and selected behavioural indicators. Comparatively fewer approaches integrate transactional, behavioural, device-related, temporal, and network-level features within a unified modelling framework. Such multidimensional representations may provide a more comprehensive characterization of fraudulent behaviour and improve the detection of complex, coordinated, and evolving fraud patterns.

Fourth, an effective fraud-detection framework

must balance predictive performance with interpretability, computational efficiency, scalability, and real-time deployability. Highly complex deep-learning architectures may offer strong predictive capability but can introduce challenges related to computational cost, latency, interpretability, and operational implementation. Conversely, conventional ML models may provide greater interpretability and computational efficiency but may be less effective in representing complex temporal or nonlinear fraud patterns. Therefore, hybrid and explainable ensemble architectures offer a promising direction for balancing predictive performance, transparency, scalability, and operational feasibility.

Overall, the literature demonstrates a clear progression from conventional rule-based and supervised ML approaches toward ensemble learning, anomaly detection, deep learning, hybrid architectures, explainable AI, and graph-based fraud intelligence. Nevertheless, significant challenges persist, including the scarcity of authentic UPI datasets, severe class imbalance, fragmented feature representation, inadequate modelling of evolving fraud behaviour, limited explainability, and difficulties associated with real-time deployment. These limitations provide a strong rationale for developing a robust, scalable, explainable, and adaptive UPI fraud-detection framework that integrates multidimensional transaction and behavioural information while maintaining practical applicability in real-time digital payment environments.

5. Proposed Framework and Methodology

The framework is organized into four analytical layers. The first layer analyses transaction and behavioral signals. The second evaluates phishing-related indicators such as URL, text, and contextual characteristics. The third analyses multimodal identity evidence, including visual and audio characteristics where suitable datasets are available.

The fourth layer combines model outputs into an interpretable risk assessment and supports manual review for high-impact decisions.

5.1. Data Set Construction

The experimental study was conducted on a consolidated hybrid dataset containing 5,00,000 transaction records. To improve data

diversity and support robust evaluation of the proposed fraud-detection framework, the dataset was constructed by integrating information from multiple sources, including synthetically generated transaction data using the Faker library and fraud-related transaction patterns obtained from established datasets such as PaySim. The data from different sources were carefully processed and transformed into a common schema to ensure consistency across transaction attributes, feature representations, and target labels.

The data preparation process included data validation, missing-value treatment, duplicate removal, consistency checking, feature transformation, and schema harmonisation.

5.2. Data Validation, Cleaning, and Schema Harmonization

A systematic data validation process was performed to ensure the quality, consistency, and reliability of the consolidated dataset. The validation process examined the total number of records and attributes, data types, unique transaction identifiers, transaction amounts, timestamps, IP address representations, account-balance consistency, missing values, outliers, and the validity of target labels. Exact duplicate records were identified using a combination of key attributes, including Transaction_ID, Account_ID, Timestamp, Amount and Merchant.

However, transactions were not considered duplicates solely because they shared the same transaction amount or merchant, as identical values may legitimately occur across different transactions. Missing-value treatment was performed according to the nature and data type of the corresponding attribute. Where appropriate, median imputation was applied to numerical features, while an explicit “Unknown” category was used to represent missing values in categorical attributes. Invalid timestamps and transaction identifiers were subjected to validity and consistency checks rather than being blindly imputed. Missing or questionable target labels were not imputed without investigation, as inappropriate label imputation could introduce bias into the classification of fraudulent and legitimate transactions.

To minimise the risk of data leakage, all data-dependent preprocessing parameters, including imputation values, scaling parameters, and encoding rules, were estimated exclusively from the training partition. The learned parameters

were subsequently applied unchanged to the validation and test partitions. This procedure ensured a fair and reliable assessment of the predictive performance of the developed models. Schema harmonisation was performed to integrate data obtained from different sources into a consistent analytical structure. The process included the adoption of common naming conventions, data-type conversion, categorical encoding, numerical transformation where required, temporal feature extraction, and target-label normalisation. The provenance of each data record was also preserved to maintain source traceability and support the reproducibility of the experimental framework. Where the source datasets provided the required financial attributes, financial consistency checks were additionally performed.

In particular, logical and numerical relationships among Balance-Before, Transaction Amount, and Balance-After variables were examined to identify potentially inconsistent or erroneous records. These validation and harmonisation procedures were designed to remove or flag unreliable observations and establish a consistent, high-quality dataset suitable for subsequent feature engineering, model training, evaluation, and explainability analysis.

5.2.1. Creation of Unified Dataset Structure

To maintain data provenance, a Source_Dataset attribute should be retained for every transaction. A standardized representation can include:

Table 1: Unified Dataset Structure

Unified Attribute	Description
Source_Dataset	Original dataset/source
Transaction_ID	Unique transaction identifier
Account_ID	Account/user identifier
Transaction_Type	Type of financial transaction
Amount	Transaction amount
Merchant	Merchant/payee information
Location	Geographic information
Device_ID	Device identifier/fingerprint

Unified Attribute	Description
IP_Address	IP address, where available
Timestamp	Transaction time
Balance_Before	A/C balance before transaction
Balance_After	Account balance after transaction
Transaction_Status	Transaction outcome/status
Fraud_Flag	Standardized fraud indicator
Phishing_Flag	Standardized phishing indicator
Risk_Score	Risk score, where available
Original_Label	Label provided

5.2.2. Data Type Conversion

Since the dataset combines information from multiple sources, all attributes should be converted to a common schema.

Table 2: Data Type Conversion

Attribute	Standard Type
Transaction_ID	String
Account_ID	String
Transaction_Type	Categorical
Amount	Float
Merchant	Categorical
Location	Categorical
Device_ID	String
IP_Address	String
Timestamp	Datetime
Balance_Before	Float
Balance_After	Float
Transaction_Status	Categorical
Fraud_Flag	Binary Integer
Phishing_Flag	Binary Integer
Risk_Score	Float
Original_Label	Categorical

5.2.3. Data Set Preparation

Table 3: Data set formulation from various sources

Dataset Source / Category	Data Characteristics and Variables	Target Label	Research Purpose
PaySim – Synthetic Financial Dataset	Synthetic mobile-money transactions generated using the PaySim simulator. Variables include transaction type, amount, origin and destination accounts, sender/receiver balances, transaction time, isFraud and isFlaggedFraud.	Is_Fraud	Financial fraud detection, anomaly detection, class-imbalance analysis and evaluation of ML/DL models.
Faker-Generated Financial Transaction Dataset	A controlled synthetic dataset of 1,000 transactions generated using the Faker library. Variables include Transaction ID, Account ID, User Name, amount, transaction type, merchant, location, timestamp, balance after transaction and phishing indicator.	Is_Phishing	Phishing-related transaction detection, supervised classification and controlled model experimentation.
Credit Card Transaction Data	Transaction-level records containing transaction/card identifiers, merchant category, transaction amount, timestamp, location, device information and fraud indicators.	Fraud/Legitimate	Credit-card fraud detection, anomaly detection and transaction-risk prediction.
UPI Transaction Data	Digital-payment transactions containing UPI identifiers, sender/receiver banks, transaction amount, device fingerprint, IP address, merchant type, transaction status and fraud indicators.	Fraud/Legitimate	UPI fraud detection, account-takeover analysis, velocity fraud and behavioral risk assessment.
Bank Transfer Data	Account-to-account transfer records containing account/beneficiary information, IFSC, amount, transfer	Suspicious/Legitimate/Risk	AML detection, mule-account identification, suspicious transaction

Dataset Source / Category	Data Characteristics and Variables	Target Label /	Research Purpose
	type, transaction frequency and risk-related attributes.		monitoring and transaction-network analysis.
Mobile Wallet Data	Wallet-based transactions contain wallet ID, device ID, geographical information, recharge type, wallet balance, merchant category and transaction details.	Fraud/ Legitimate	Wallet fraud detection, behavioural profiling, device-based analysis and geospatial anomaly detection.

5.2.4. Research Methodology

The proposed methodology follows a unified, leakage-aware pipeline:

Multi-source data collection → Data validation and cleaning → Schema harmonization → Exploratory analysis → Feature engineering → Feature selection → Stratified data splitting → Class-imbalance analysis → Model training → Hyperparameter optimization → Independent testing → Cross-dataset evaluation → Explainability analysis → Risk-based decision support.

5.3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand the underlying characteristics, distribution, and quality of the consolidated dataset. The analysis quantified the proportions of fraudulent, phishing-related, and legitimate transactions and summarised the distributions of both numerical and categorical features. Statistical analysis was conducted to identify skewness, extreme values, and potential outliers that could influence model performance.

Temporal characteristics of the transactions were also examined to identify potential patterns related to transaction time, frequency, and periodic variations. Furthermore, relationships among the input features were analysed using appropriate statistical measures and correlation analysis. Multicollinearity among numerical predictors was assessed to identify highly correlated or redundant features that could adversely affect model interpretation and predictive performance.

5.4. Feature Engineering and Feature Selection

Feature engineering was performed to derive temporal, financial and behavioural features from transaction data. Temporal features included transaction hour, day of week, weekend and night-time indicators, while financial features included transaction-to-balance ratios and balance differences. Behavioural features captured transaction frequencies across accounts, merchants, devices, IP addresses, and locations to identify deviations from normal transaction patterns.

Direct identifiers such as Transaction_ID and high-cardinality attributes were excluded or transformed into meaningful behavioural representations. Raw timestamps were converted into relevant temporal features, and target-related variables were excluded to prevent data leakage. Feature selection was performed using correlation analysis, mutual information, Recursive Feature Elimination (RFE), and model-based feature importance. All feature-selection and data-dependent transformation procedures were fitted only on the training data and subsequently applied unchanged to validation and test datasets.

5.5. Data Splitting and Class-Imbalance Handling

The proposed experimental split is 70% training, 15% validation and 15% testing, using stratification to preserve class distributions.

5.6. Machine-Learning Models

XGBoost was selected as the primary model for tabular fraud detection because of its ability to capture nonlinear relationships and feature interactions within heterogeneous transaction data. An initial configuration of 200 estimators, maximum depth of 6, and learning rate of 0.1, along with appropriate positive-class weighting, was considered for model development. These parameters were treated as initial settings rather than fixed optimal values and were subsequently optimized using the validation-based hyperparameter tuning procedure.

For multimodal deepfake analysis, the manuscript proposes CNNs for image/frame analysis, LSTMs for sequential patterns, and transformer-based models for phishing-text analysis. In the refined framework these components are treated as modular models whose outputs can be calibrated and passed to the common decision layer rather than assuming that one architecture is optimal for every modality.

5.7. Explainable AI Layer

The Explainable Artificial Intelligence (XAI) layer was incorporated to provide both global and local interpretations of model predictions. SHAP was used to quantify the contribution of individual features to each prediction, while aggregated SHAP values across the test dataset were used to identify the dominant factors associated with fraud risk. LIME was considered for generating local surrogate explanations for selected transactions.

The generated explanations were assessed in terms of stability, consistency, and practical usefulness rather than being interpreted as evidence of causal relationships. For instance, a fraud alert may indicate that unusual transaction timing, a newly observed device, increased transaction frequency, or geographic deviation contributed substantially to the model's estimated risk. Similarly, for deepfake detection, relevant visual or audio characteristics may be presented as supporting model evidence. All explanations are therefore interpreted as model-attribution evidence indicating the factors influencing a prediction, rather than as a definitive causal conclusion.

5.8. Integrated Decision & Workflow

The proposed framework begins with the acquisition of transactional data and, where applicable, identity-verification information. Relevant features are extracted and provided to the corresponding fraud, phishing, and deepfake detection models. The calibrated outputs of these models are then integrated to generate a unified risk score. Based on predefined risk thresholds and institutional security policies, low-risk transactions may proceed under standard controls, medium-risk transactions may require additional verification, and high-risk transactions may be subjected to manual review or enhanced security checks.

As fraud patterns continuously evolve, continuous monitoring and periodic model

evaluation are incorporated into the framework. Feedback from confirmed fraudulent transactions, legitimate transactions, analyst assessments, and model-performance monitoring can be used for model updating and retraining. Where centralised data sharing is constrained by privacy or regulatory requirements, privacy-preserving techniques such as federated learning may be explored to support collaborative model training without requiring direct centralisation of sensitive transaction data.

6. Experimental Setup

The experiments were conducted using a computing environment equipped with an Intel Core i7-class processor, 16 GB RAM, and an NVIDIA GPU. The software environment comprised Python, Jupyter Notebook, TensorFlow, scikit-learn, OpenCV, SHAP and LIME. To ensure experimental reproducibility, the final implementation should document the operating system, Python and library versions, GPU model, random seeds, preprocessing configurations, hyperparameter-search space, and model training time.

The experimental protocol was designed according to the characteristics of each detection task. Transaction datasets were used for financial fraud detection, phishing datasets for phishing classification, and video/audio deepfake datasets for multimodal manipulation detection. Each task was evaluated independently using appropriate performance metrics before combining the outputs at the decision-support level. This task-wise evaluation ensures that the performance of individual models can be assessed objectively and prevents the integrated risk score from masking weaknesses in any particular detection component.

7. Evaluation Metrics and Reporting Protocol

Accuracy alone is not considered an adequate performance measure for highly imbalanced fraud-detection datasets, as a model may achieve high accuracy while failing to identify a significant proportion of fraudulent transactions. Therefore, the proposed evaluation framework considers Precision, Recall, F1-Score, ROC-AUC, and Precision-Recall AUC (PR-AUC) as the primary performance metrics. Confusion

matrices are also reported at the selected operating threshold to provide a detailed assessment of true-positive, true-negative, false-positive, and false-negative predictions. For financial fraud-detection applications, the classification threshold should be selected by considering the relative operational costs of false positives and false negatives. To improve the reliability of the experimental findings, the final study should report confidence intervals or variability across repeated experimental runs, wherever feasible.

8. Security, Privacy, and Operational Considerations

A production system must minimize the collection and retention of personally identifiable and biometric information. Device identifiers, IP information, account information, audio, and facial data should be handled according to applicable organisational controls and legal requirements. Access control, encryption, audit logging, retention limits, and secure model-serving practices are important complements to the detection models.

Deepfake and phishing detectors can also be attacked through adversarial or distribution-shift techniques. Model monitoring should therefore include drift detection, robustness testing, periodic threshold review, and controlled retraining. XAI outputs should support analysts without exposing sensitive internal security rules to attackers.

9. Limitations

The proposed framework has limitations arising from the use of synthetic and publicly available benchmark datasets, which may not fully represent the complexity of real-world Indian digital-payment transactions. Fraud, phishing, and deepfake components also differ in data characteristics and therefore require independent task-specific validation before integration. In addition, comprehensive experimental validation of the complete multimodal framework remains necessary. Future work will focus on larger and more representative datasets, cross-dataset evaluation, reproducible data-generation procedures, and rigorous validation of the integrated framework.

10. Future Scope

- Federated learning for privacy-preserving collaboration across financial institutions.
- Real-time multimodal deepfake detection combining visual, audio, and behavioural evidence.
- Robustness evaluation against adversarial manipulation and distribution shift.
- Blockchain or tamper-evident mechanisms for selected fraud-evidence and audit workflows.
- Integration with institutional fraud-monitoring and regulatory-compliance processes where technically and legally appropriate.
- Continual learning and drift-aware retraining for rapidly changing phishing and fraud patterns.

11. Conclusion

This paper presents a unified, explainable machine-learning framework for detecting phishing, financial fraud, and deepfake-related identity threats in digital-payment environments. The framework consolidates heterogeneous transaction data, leakage-aware preprocessing, behavioural and temporal feature engineering, XGBoost-based tabular modelling, modular multimodal deepfake and phishing models, and SHAP/LIME explanations within a common risk-decision pipeline.

The principal contribution is the integration of detection and interpretability rather than treating model accuracy as the sole objective. The proposed methodology is designed to support transparent investigation, reduce avoidable false positives, and provide a scalable basis for future validation in Indian digital-payment ecosystems. However, the supplied manuscript does not yet provide sufficient experimental evidence to claim the numerical performance targets originally stated. Future work should therefore prioritise reproducible experiments, cross-dataset validation, real-world or institutionally representative data, robustness testing, and rigorous evaluation of explanation quality.

References

- [1] S. S. Bodade and P. P. Pawade, "Review paper on UPI fraud detection using machine learning," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 11, no. 12, pp. 1283–1285, Dec. 2023, doi: 10.22214/ijraset.2023.57551.
- [2] V. Kamra, V. Varshney, and S. Yadav, "A novice approach for UPI fraud detection by using machine learning techniques," in *Proc. 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, Nov. 2025, pp. 1–6, doi: 10.1109/ICCCA66364.2025.11325722.
- [3] N. Sowmiya, G. Maitra Varuna, S. Nesana, and I. Kasthuri, "UPI Sentinel: Anomaly-aware fraud detection using hybrid ML models," in *Proc. 2026 International Conference on Visual Analytics and Data Visualization (ICVADV)*, Tirunelveli, India, Feb. 2026, pp. 1609–1614, doi: 10.1109/ICVADV67766.2026.11470456.
- [4] J. Kumar and N. Rani, "Optimized machine learning and deep learning approaches for effective detection of fraud in Unified Payments Interface (UPI) transactions," *International Journal on Science and Technology (IJSAT)*, vol. 16, no. 4, pp. 1–15, Oct.–Dec. 2025, doi: 10.71097/IJSAT.v16.i4.9525.
- [5] N. B. Chakka and S. S. Saheb, "A hybrid autoencoder-XGBoost framework for high-performance UPI fraud detection," *Proceedings of Engineering and Technology Innovation*, vol. 33, pp. 93–105, 2026, doi: 10.46604/peti.2026.16002.
- [6] Meghana L, Amutha S, and S. Patil, "A machine learning-based UPI security and fraud detection," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 14, no. 8, pp. 1263–1268, Aug. 2026, doi: 10.22214/ijraset.2026.84675.
- [7] N. Dhinakaran, A. B. Hajira Be, and J. Syed Raffi Ahamed, "Smart UPI fraud detect: UPI fraud detection using machine learning," *International Journal of Creative and Open Research in Engineering and Management*, vol. 2, no. 6, pp. 1–8, Jun. 2026, doi: 10.55041/ijcope.v2i6.172.
- [8] N. Sindhura and N. P. Sai, "An intelligent machine learning framework for fraud detection in Unified Payments Interface (UPI) transactions," *International Journal of Engineering Research and Science & Technology*, vol. 22, no. 2(1), pp. 29–39, 2026.
- [9] S. Sahu, P. Chaudhury, S. P. Senapati, S. Pradhan, and S. K. Nahak, "Detection of mule accounts and fraudsters in UPI transactions using AI and machine learning techniques," *Iconic Research and Engineering Journals*, vol. 9, no. 10, pp. 904–913, Apr. 2026, doi: 10.64388/IREV9I10-1716166.
- [10] S. Jagadeesan, K. S. Arjun, G. Dhanika, G. Karthikeyan, and K. Deepika, "UPI fraud detection using machine learning," in *Challenges in Information, Communication and Computing Technology*, 1st ed., V. Sharmila, S. Kannadhasan, A. R. Kannan, P. Sivakumar, and V. Vennila, Eds. Boca Raton, FL, USA: CRC Press, 2024, pp. 755–760, doi: 10.1201/9781003559085-130.

Publisher's Note & Copyright

IRJIST Journals remains neutral regarding jurisdictional claims in published maps and institutional affiliations; the views expressed are solely those of the authors.

© 2026 by the authors. Open access under the CC BY 4.0 license.
