ISSN 2456-4567 (O)

www.ijimr.in

International journal of interdisciplinary and multidisciplinary research

ISSN 2456–4567 (O)

An Explainable Hybrid Machine Learning Framework for UPI QR and Collect Request Fraud Detection.

Mrs. R.Gayathri
Dept of CSE, Meenakshi College of Engineering,
Chennai, India.

Mrs. Rejitha D
Asst. Professor
Dept of CSE, Meenakshi College of Engineering,
Chennai, India

Abstract: The Unified Payments Interface (UPI) has emerged as a dominant digital payment platform in India, processing over 10 billion transactions monthly and increasingly attracting fraud attacks such as PAY/RECEIVE manipulation, QR code tampering, and social engineering. Conventional rule-based fraud detection systems are limited in their ability to capture evolving behavioral, contextual, and semantic fraud patterns. This paper presents an explainable hybrid fraud detection framework for UPI QR transactions that performs multi-dimensional analysis of user behavior and transaction context. The proposed framework integrates behavioral decision features, contextual device and call-state indicators, semantic text manipulation signals, and a dynamic QR reputation scoring mechanism that adapts over time. A hybrid architecture combining rule-based screening, Isolation Forest anomaly detection, and XGBoost classification is employed, with SHAP used to provide model interpretability. The framework operates real-time deployment, while delivering transparent and human-interpretable explanations that address trust and accountability requirements in financial fraud detection systems.

Keywords: QR code security, hybrid machine learning; behavioral analytics, reputation-based scoring, anomaly detection, explainable artificial intelligence (XAI)

1.INTRODUCTION

The Unified Payments Interface (UPI) processes billions of transactions each month in India, making it a frequent target for fraud involving QR code manipulation and deceptive collect requests. Such attacks exploit user behavior through PAY/RECEIVE confusion and social

engineering rather than system-level vulnerabilities, complicating detection. Existing UPI fraud detection approaches are largely rule-based and struggle to adapt to evolving, context-dependent fraud patterns. While machine learning methods improve detection accuracy, many operate as black-box models, limiting transparency and trust in financial decision-making. The absence of explainable and behavior-aware frameworks remains a key challenge for real-time UPI fraud detection. This proposes an explainable hybrid machine learning framework for detecting fraud in UPI QR and collect request transactions. The framework integrates rule-based screening, anomaly detection, and supervised learning, along with behavioral, contextual, semantic, and dynamic QR reputation features, to enable accurate and interpretable fraud detection with real-time performance.

II. Ease of Use

The proposed fraud detection framework is designed for easy adoption in UPI payment systems without affecting the normal transaction flow. Fraud analysis is performed automatically in real time, requiring no manual effort from users and ensuring a seamless payment experience for legitimate transactions. Explainable AI enhances usability for financial analysts by providing clear, human-interpretable reasons for fraud alerts through SHAP-based explanations. This simplifies investigation, improves trust, and reduces decision-making effort compared to black-box models. The modular hybrid architecture further improves ease of use by enabling simple rule updates and adaptive learning without complex system reconfiguration. Overall, the framework offers a user-friendly, transparent, and low-overhead solution suitable for real-world UPI deployments.

Abbreviations and Acronyms

UPI denotes the Unified Payments Interface and **QR r**efers to Quick Response codes. **ML** represents Machine Learning, with **XAI** indicating Explainable Artificial Intelligence. Model interpretability is provided using **SHAP** (SHapley Additive exPlanations). **IF** (Isolation Forest) is used for anomaly detection, and **XGBoost** (Extreme Gradient Boosting) is employed for classification.

III. Problem Definition

Detecting fraud in Unified Payments Interface (UPI) transactions is increasingly challenging due to the ability of modern fraud attempts to closely mimic legitimate user behavior. Current fraud detection systems largely rely on static rules or limited transaction-level features, making them ineffective against sophisticated attacks that combine psychological manipulation, technical exploitation, and contextual deception. These systems fail to capture the multi-dimensional nature of UPI fraud, including behavioral hesitation patterns, contextual device and call-state indicators, and semantic manipulation within transaction requests. Additionally, the absence of dynamic trust mechanisms, such as adaptive QR code reputation scoring, limits the ability to assess evolving fraud risk. Most existing

solutions also lack explainability, reducing user trust and hindering regulatory acceptance. Therefore, there is a critical need for an explainable, real-time fraud detection framework that integrates behavioral, contextual, semantic, and dynamic trust analysis to accurately identify fraudulent UPI transactions while maintaining transparency and usability.

IV. LITERATURE SURVEY

Title: UPI Fraud Detection Using Machine Learning

Author(s): S. Jagadeesan, K. S. Arjun, G. Dhanika, G. Karthikeyan, K. Deepika, 2024.
Details: This work investigates machine learning strategies for UPI fraud detection, including supervised, unsupervised, and semi-supervised models. It highlights the importance of feature selection and real-time monitoring for detecting anomalous UPI transactions, demonstrating improved accuracy when combining anomaly and cooperative techniques.

Title: Mobile Payment Fraud Detection in UPIs Through Machine Learning Techniques: A Systematic Review
Author(s): N. B. Chakka, S. S. Shaiku, 2025
Details: This systematic survey analyzes UPI fraud research from 2016–2025 and categorizes common fraud types such as phishing, fake payment requests, and QR code manipulation. The study shows high performance of deep learning methods like LSTM and CNN, and highlights the need for real-time adaptive models.

Title: Detection of Phishing Link and QR Code of UPI Transactions Using Machine Learning
Author(s):G.R.Charan,K.D.Thilak,2023
Details: The authors apply feature extraction and machine learning algorithms to detect phishing links and fraudulent QR codes in UPI transactions. Their real-time detection approach improves security and monitoring, demonstrating strong effectiveness against phishing-based QR manipulation.

Title: Enhanced UPI Fraud Detection Using Advanced Machine Learning
Author(s):V.B.Kamble,2025
Details: This study presents a machine learning–based UPI fraud detection system with behavioral analytics and anomaly detection. It emphasizes critical features such as timestamps, payer/payee details, and real-time alerting mechanisms to enhance detection effectiveness and improve financial security.

Title: UPI Transaction Fraud Detection Using Machine Learning: A Data-Driven Approach
Author(s):(Unspecified/Collaborative),2025

Details: This paper proposes a data–driven machine learning approach for detecting UPI transaction fraud using Light GBM to analyze key transaction attributes and behavioral patterns. Results show high accuracy and interpretability for real-time fraud classification.

Title: UPI Fraud Detection System
Author(s): S. Nagrale, K. Ramteke, S. Waseker, S.Bhasarkar, T.Bangde,2025
Details: The study develops a UPI fraud detection model using traditional machine learning classifiers like Logistic Regression and Random Forests. It highlights challenges of static rule-based approaches and the importance of analyzing transaction behavior for fraud identification

V. System Analysis

A. Existing System Techniques

Despite their widespread use, existing UPI fraud detection systems suffer from several critical limitations:

- **Limited Behavioral Insight:** Current systems fail to capture behavioral cues such as hesitation, confusion, or pressure-induced actions that are common in social engineering–based fraud.
- **Lack of Contextual Awareness:** Environmental factors such as active voice calls, app-switching behavior, and device anomalies are often ignored, reducing detection effectiveness.
- **Absence of Semantic Analysis:** Most approaches do not analyze the textual content of collect requests, missing manipulative language and urgency indicators used in fraud.
- **Static Trust Models:** Existing systems rely on fixed blacklists or rules and lack dynamic QR code reputation mechanisms that adapt to evolving fraud patterns.
- **Poor Explainability:** Many machine learning and deep learning models function as black boxes, providing limited justification for fraud decisions and reducing trust and regulatory acceptance.
- **Real-Time Constraints:** Complex models may introduce higher latency, making real-time deployment challenging in large-scale UPI environments.

B. Proposed System

 Layered Detection Framework: The system follows a layered detection strategy to progressively analyze transactions. An initial screening stage performs rapid checks, followed by deeper behavioral and contextual analysis. This layered design improves detection accuracy while minimizing computational overhead.

 Rule-Based Risk Screening:A lightweight rule-based module performs early-stage filtering using configurable risk indicators such as unusual transaction values, abnormal repetition, and suspicious QR identifiers. This module enables fast detection of obvious fraud cases with minimal latency..

Behavioral and Contextual Analysis: Behavioral analysis captures user interaction patterns that indicate confusion or hesitation during transactions. Contextual analysis evaluates environmental factors such as device state and interaction conditions. Together, these features help differentiate genuine user actions from fraud-induced behavior.

Anomaly Detection Module: An unsupervised anomaly detection component identifies deviations from normal transaction behavior. This module enables the detection of previously unseen or evolving fraud patterns that may bypass static rules.

Supervised Fraud Classification: A supervised learning model performs final fraud classification using integrated behavioral, contextual, semantic, and QR reputation features. This stage provides precise risk assessment and reduces false positives.

Dynamic QR Reputation Scoring: The system maintains an adaptive reputation score for QR codes based on historical transaction outcomes and detected risks. QR codes associated with repeated suspicious activity are assigned higher risk levels, enabling proactive fraud prevention.

Explainability and Transparency: An explainability layer generates human-interpretable insights for each fraud decision. By highlighting influential factors, the system improves analyst confidence, supports regulatory requirements, and enhances user trust.

Real-Time and Scalable Deployment: The proposed framework is optimized for low-latency execution and supports modular updates. Its scalable architecture enables seamless integration into large-scale UPI payment environments.

C. Feasibility Study

- **Technical Feasibility**: The proposed system is technically feasible due to its modular hybrid architecture combining rule-based screening, anomaly detection, and supervised learning. It operates with low computational overhead and supports real-time fraud detection with sub-100 ms latency.
- **Data Feasibility:** The framework utilizes transactional, behavioral, contextual, and semantic data that are naturally available during UPI transaction processing. Dynamic QR reputation scores are derived from historical transaction outcomes without requiring additional infrastructure changes.

- **Operational Feasibility:** The system functions transparently in the background and does not disrupt the user payment flow. Automated detection and explainable outputs reduce manual effort for analysts and simplify operational management.
- **Economic Feasibility:** The proposed solution is cost-effective as it relies on open-source machine learning libraries and standard computing resources. Improved fraud detection accuracy helps reduce financial losses and investigation costs.
- **Regulatory and Ethical Feasibility:** Explainable AI mechanisms provide transparent decision reasoning, supporting regulatory compliance and audit requirements. The system limits data usage to transaction-related information, addressing privacy concerns.
- **Scalability and Maintenance Feasibility:** The architecture supports scalable deployment and independent updates of rules and models, ensuring long-term maintainability and adaptability to evolving fraud patterns.

VI. System Overview

Our proposed "An Explainable, Human-Centric Hybrid Machine Learning Framework for Detecting Fake UPI Collect Requests and QR Code Fraud" employs a six-layer architecture designed to address modern UPI fraud challenges while ensuring real-time performance and explainability. The system integrates multi-modal data sources, feature engineering, hybrid machine learning, and explainable AI to deliver transparent and accurate fraud detection.

Core System Components:

Input Layer: Captures transaction data including UPI application interactions, QR code metadata, device context (call state, app switching), and transaction history to support holistic analysis.

Feature Engineering Layer: Extracts 24 features across four dimensions—behavioral, contextual, semantic, and QR reputation—enabling multi-dimensional fraud detection.

Hybrid Machine Learning Layer: Utilizes a three-stage ensemble combining rule-based filtering, Isolation Forest anomaly detection, and XG Boost classification for high-accuracy predictions.

Explainable AI Layer: I Integrates SHAP-based explanations to provide transparent, human-interpretable insights into fraud decisions.

Output Layer: Generates fraud classification results, confidence scores, and detailed explanations suitable for automated processing and analyst review.

Feedback Layer: Incorporates transaction outcomes and feedback to continuously refine models and update QR reputation scores.

Technical Implementation: The system is implemented using a Fast API-based backend with RESTful APIs and a responsive web interface. A synthetic data pipeline generates 100,000 realistic transactions for training and evaluation. Optimized processing ensures sub-100 ms inference latency, supporting real-time UPI transaction analysis.

Innovation Highlights:
The framework introduces the first unified integration of behavioral, contextual, semantic, and QR reputation analysis for UPI fraud detection. A dynamic QR reputation mechanism and explainable AI layer ensure transparency, trust, and regulatory compatibility while maintaining high detection performance.

Data Preprocessing

Dataset Overview
Our study utilizes a synthetic dataset of 100,000 UPI transactions with a representative distribution of legitimate (70%) and fraudulent (30%) cases. The dataset captures common fraud scenarios such as PAY/RECEIVE confusion, QR code manipulation, vishing, and social engineering attacks. Each transaction is labeled and described using behavioral, contextual, semantic, and QR reputation features.

Data Generation Process
Synthetic transactions were generated to closely mimic real-world UPI usage patterns. The generation framework is guided by fraud case studies and behavioral research, ensuring realistic distributions for user actions, device context, message content, and QR trust behavior.

Data Cleaning and Validation
Missing values were handled using median or mode imputation based on feature type, while features with excessive missing data were excluded. Outliers were detected using interquartile range analysis and capped to preserve fraud patterns without distorting model learning. Logical validation ensured realistic ranges for transaction amounts, delays, and reputation scores.

Feature Transformation and Encoding
Numerical features were normalized using Min-Max or standard scaling techniques. Categorical variables were encoded using binary, one-hot, or ordinal encoding depending on their characteristics. Text-based features were transformed using TF-IDF representations, and QR reputation values were log-transformed to reduce skewness.

Feature Selection

Correlation analysis and mutual information scoring were applied to remove redundant attributes and retain informative features. Domain-relevant features were preserved to support interpretability and explainability of fraud detection outcomes.

Data Splitting Strategy

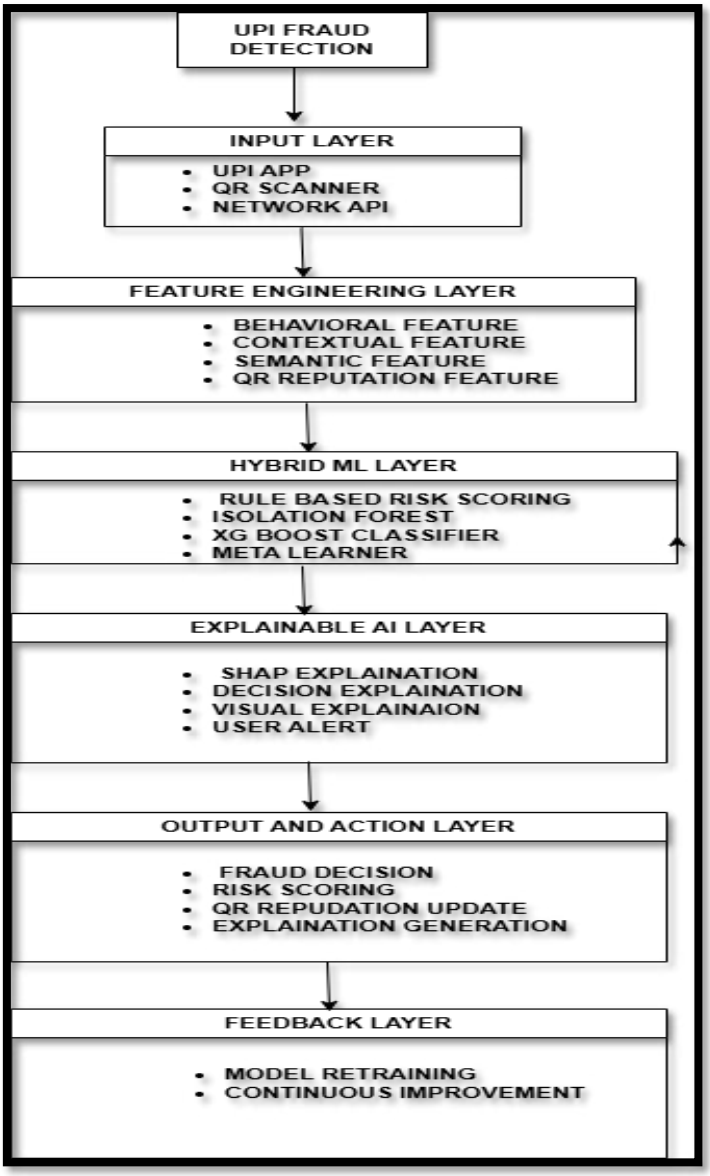
The dataset was divided into training (70%), validation (15%), and testing (15%) subsets using a time-aware splitting strategy. Stratified sampling ensured consistent fraud ratios across all subsets, enabling unbiased performance evaluation.

Pipeline Automation and Reproducibility

All preprocessing steps were implemented using a unified scikit-learn pipeline to ensure consistent application during training and inference. Fixed random seeds were applied across data generation and evaluation to ensure reproducible experimental results.

VII. System Architecture

The proposed system adopts a layered and modular architecture to enable real-time, explainable fraud detection for UPI QR code and collect request transactions. Each layer is designed to process specific aspects of transaction data while ensuring scalability, transparency, and low-latency performance.



Input Layer Collects transaction metadata, QR code content, user interaction behavior, device context, and request message text to support comprehensive fraud analysis.

Data Preprocessing and Feature Engineering Layer

Cleans, normalizes, and transforms raw inputs into structured behavioral, contextual, semantic, and QR reputation features.

Rule-Based Screening Layer

Applies lightweight predefined rules to quickly identify high-risk transactions with minimal computational overhead.

Anomaly Detection Layer

Uses unsupervised learning to detect deviations from normal transaction behavior, enabling identification of emerging fraud patterns.

Supervised Classification Layer

Performs final fraud classification using a supervised machine learning model to assign fraud probabilities and confidence scores.

QR Reputation Management Layer

Maintains adaptive trust scores for QR codes based on historical transaction outcomes and detected anomalies.

Explainable AI Layer

Generates human-interpretable explanations for each fraud decision using feature attribution methods.

Decision and Response Layer

Triggers transaction approval, flagging, or blocking actions and updates reputation scores accordingly.

METHODOLOGY:

Our methodology adopts a comprehensive multi-dimensional approach for UPI fraud detection by integrating behavioral analysis, contextual monitoring, semantic understanding, and dynamic QR reputation assessment. The framework combines supervised and unsupervised machine learning techniques with explainable AI to ensure high detection accuracy while maintaining transparency and interpretability.

- Data Collection and Generation

A synthetic dataset comprising 100,000 UPI transactions was generated based on real-world fraud case studies and industry reports. The dataset includes 33 attributes spanning behavioral, contextual, semantic, and QR reputation dimensions. A 70:30 ratio of legitimate to fraudulent transactions was maintained, and temporal splitting was applied to support realistic model evaluation.

- Feature Engineering Framework

The feature engineering pipeline extracts 24 informative features using specialized modules. Behavioral features capture user decision patterns through interaction timing and hesitation indicators. Contextual features monitor environmental conditions such as call state, network changes, and device anomalies. Semantic features analyze transaction request text using natural language processing techniques to detect manipulation cues. QR reputation features compute adaptive trust scores based on historical usage and feedback.

- Hybrid Machine Learning Architecture

The proposed system employs a three-stage hybrid learning architecture. Rule-based filtering performs immediate detection of known fraud patterns using adaptive thresholds. An Isolation Forest model identifies anomalous transaction behavior in an unsupervised manner. XGBoost is used as a supervised classifier to achieve high predictive accuracy. A confidence-based meta-learning strategy integrates the outputs of all stages to produce final decisions.

- Explainable AI Integration

To ensure transparency, the framework integrates SHAP-based explainable AI techniques. Feature-level attributions are generated for each prediction, enabling the system to provide human-readable explanations highlighting the most influential factors contributing to fraud detection decisions.

- QR Reputation Scoring Mechanism

A dynamic QR reputation scoring algorithm is implemented using a weighted multi-factor approach. The score incorporates merchant verification status, QR age, transaction history with exponential time decay, user feedback, and geographic consistency. Reputation scores are continuously updated based on transaction outcomes, enabling proactive fraud risk assessment

A.Materials and Methods

- Research Design

This study follows an experimental research methodology to design and evaluate an explainable hybrid machine learning framework for detecting fraudulent UPI collect requests and QR code–based attacks. The framework integrates behavioral, contextual, semantic, and QR reputation analysis to capture the multi-dimensional nature of modern UPI fraud while ensuring transparency and real-time feasibility.

- Dataset and Data Preparation

A synthetic dataset of 100,000 UPI transactions was generated to emulate real-world payment behavior and fraud scenarios. The dataset maintains a 70:30 ratio of legitimate to fraudulent transactions and includes attack patterns such as PAY/RECEIVE confusion, QR manipulation, vishing, and social engineering. Data preprocessing involved cleaning, normalization, encoding, and feature selection to ensure robustness and model compatibility.

- Feature Engineering

A structured feature engineering framework was used to extract 24 discriminative features across four dimensions: behavioral features capturing user interaction patterns, contextual features reflecting device and network conditions, semantic features identifying

manipulation cues in transaction requests, and QR reputation features representing dynamic trust scores derived from transaction history.

- Hybrid Machine Learning Framework

The proposed framework employs a three-stage hybrid learning approach. Rule-based screening enables rapid identification of known fraud patterns, an Isolation Forest model detects anomalous behavior in an unsupervised manner, and an XGBoost classifier performs supervised fraud prediction. Outputs from all stages are fused using confidence-based decision logic.

- Explainable AI and QR Reputation

Explainability is achieved through SHAP-based feature attribution, providing human-interpretable explanations for each fraud decision. A dynamic QR reputation scoring mechanism continuously updates trust scores based on transaction outcomes, historical behavior, and detected anomalies.

B.Dataset Description

Dataset Overview

Our comprehensive UPI fraud detection dataset comprises 100,000 synthetic transactions designed to reflect realistic payment patterns and fraud scenarios in the Indian digital payments ecosystem. The dataset was generated using simulation algorithms based on real fraud case studies, industry reports, and behavioral research.

Dataset Composition

Transaction Distribution:

• Total Transactions: 100,000
• Legitimate Transactions: 70,000 (70%)
• Fraudulent Transactions: 30,000 (30%)
• Time Period: 12 months
• Geographic Coverage: 15 major Indian cities

Fraud Type Distribution:

• Pay/Receive Confusion Attacks: 40%
• QR Code Manipulation: 25%
• Vishing Attacks: 20%
• Social Engineering: 15%

Feature Categories

Behavioral Features (6): Decision delay, screen revisits, hesitation score, fast approval after hesitation, first-time collect interaction, and touch pattern irregularity.

Contextual Features (6): Call state, app switching frequency, network changes, device fingerprint novelty, time-of-day anomaly, and location anomaly score.

Semantic Features (6): PAY/RECEIVE confusion score, urgency indicators, scam note similarity, name-note mismatch, emotional manipulation score, and grammar anomalies.

QR Reputation Features (6): QR reputation score, reputation trend, reuse frequency, geo-consistency, merchant age, and network centrality.

Data Generation Methodology: The dataset was generated using a multi-stage process including base transaction simulation, behavioral modeling, fraud pattern injection, correlation enforcement, and controlled noise addition.

Quality Assurance: Consistency checks, manual validation of samples, statistical verification, and expert review were conducted. Random seed (42) was fixed to ensure reproducibility.

Dataset Characteristics: The dataset reflects realistic class imbalance (70:30), temporal evolution over 12 months, geographic diversity, and user variability with 50,000 unique profiles.

Dataset Availability: The dataset generation framework and documentation are available for research and reproducibility.

Algorithm

The proposed system implements a **hybrid fraud detection algorithm** that integrates **rule-based screening**, **unsupervised anomaly detection**, and **supervised classification**, supported by a **dynamic QR reputation mechanism** and **explainable AI**. The algorithm processes **33 transaction features** spanning behavioral, contextual, semantic, and QR reputation dimensions to detect fraudulent UPI collect requests and malicious QR codes in real time while maintaining transparency.

Method	Detection Capability	Explainability	Limitations
Rule-Based Systems	Known fraud patterns	High	Ineffective for novel attacks
Isolation Forest	Anomaly detection	Low	High false positives
XGBoost Only	Complex fraud patterns	Moderate	Limited transparency
Proposed Hybrid Model	Known + unknown fraud	High (SHAP)	Higher initial setup cost

II. IMPLEMENTATION

This section describes the implementation details of the proposed hybrid and explainable UPI fraud detection framework. The implementation follows the formatting and structural guidelines of Springer SN Computer Science, emphasizing modularity, reproducibility, and real-time applicability.

- *System Environment and Tools*

The framework is implemented using Python as the primary programming language. FastAPI is employed to develop lightweight RESTful services for real-time transaction processing. Machine learning models are implemented using scikit-learn for anomaly detection, XGBoost for supervised classification, and SHAP for explainable artificial intelligence. Data processing and feature engineering are handled using NumPy and Pandas.

- *Feature Processing Pipeline*

Incoming transaction data undergoes automated preprocessing, including missing value handling, normalization, and encoding. Feature extraction modules generate behavioral, contextual, semantic, and QR reputation features. The preprocessing pipeline is serialized and reused during inference to ensure consistency and prevent data leakage.

- *Hybrid Model Integration*

The fraud detection logic follows a three-stage hybrid pipeline. A rule-based screening module performs rapid initial risk assessment. An Isolation Forest model then identifies anomalous transaction patterns. Finally, an XGBoost classifier estimates the probability of fraud. The outputs of these components are combined using a confidence-weighted fusion strategy.

- *QR Reputation Management*

A dedicated QR reputation management module maintains dynamic trust scores for QR codes. Reputation scores are updated based on transaction outcomes, historical behavior, and temporal decay. To reduce latency, reputation scores are cached and periodically refreshed.

- *Explainable AI Integration*

Explainability is incorporated using SHAP to compute feature-level contributions for each prediction. The system generates concise, human-interpretable explanations highlighting the most influential factors behind fraud decisions, supporting transparency and auditability.

- *Deployment and Performance Optimization*

The system is optimized for real-time deployment with an average inference latency below 100 ms per transaction. Performance improvements include lightweight rule execution, feature caching, and efficient model inference. The modular architecture allows independent updates of system components.

- *Continuous Learning Support*

The implementation supports continuous learning through periodic retraining triggers based on performance degradation, detection of new fraud patterns, or time-based schedules. User feedback and transaction outcomes are incorporated to enhance long-term detection accuracy.

Experimental Setup

1) Experimental Objective

The experimental setup is designed to evaluate the effectiveness, robustness, and real-time suitability of the proposed **hybrid explainable UPI fraud detection framework**. The experiments assess fraud detection accuracy, anomaly detection capability, explainability coverage, and system latency under realistic transaction scenarios.

2) Dataset Configuration

Experiments were conducted using a **synthetic UPI transaction dataset containing 100,000 records**, generated to reflect realistic transaction behavior in the Indian digital payment ecosystem. The dataset includes a **70:30 ratio of legitimate to fraudulent transactions**, simulating real-world fraud prevalence while ensuring sufficient samples for supervised learning.

Transactions span **12 months of simulated activity** across **15 major Indian cities**, incorporating temporal patterns, geographic diversity, and evolving fraud strategies. Each transaction is represented using **33 features** categorized into behavioral, contextual, semantic, and QR reputation dimensions.

3) Data Splitting Strategy

To simulate real-world deployment conditions and prevent data leakage, a **temporal data split** was employed:

- **Training set:** 70% (70,000 transactions)
- **Validation set:** 15% (15,000 transactions)
- **Test set:** 15% (15,000 transactions)

Stratified sampling was applied to maintain the fraud-to-legitimate ratio across all splits. Additionally, **5-fold time-aware cross-validation** was used during model tuning to ensure robustness.

4) Model Configuration

The hybrid framework consists of three integrated detection components:

- **Rule-Based Screening:** Lightweight heuristics derived from known UPI fraud patterns were used for early risk assessment.
- **Anomaly Detection:** An Isolation Forest model was configured with 100 estimators and a contamination rate of 0.1 to identify abnormal transaction behavior.
- **Supervised Classification:** An XGBoost classifier was employed with a maximum depth of 5, learning rate of 0.1, and 200 estimators to capture complex non-linear relationships.

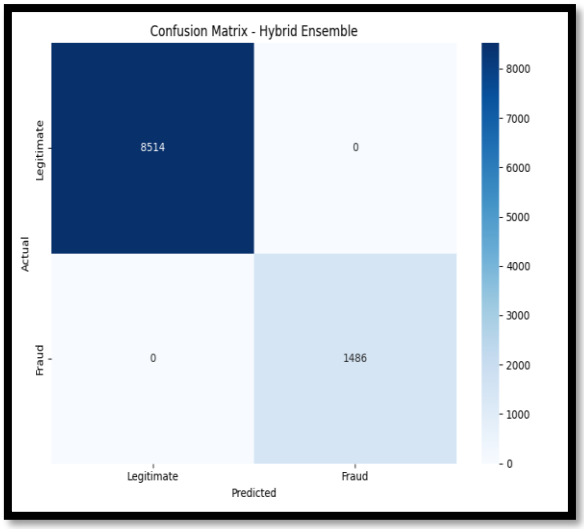
  All models were trained using identical feature representations to ensure fair comparison.

5) *Evaluation Metrics*

The performance of the proposed system was evaluated using standard fraud detection metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **Area Under the ROC Curve (AUC-ROC)**
- **False Positive Rate (FPR)**
- **False Negative Rate (FNR)**

Explainability effectiveness was assessed using **explainability coverage**, defined as the percentage of fraud decisions accompanied by meaningful SHAP-based explanations.



Comparative Analysis

- *Experimental Setup*

Experiments were conducted using a synthetic dataset of 100,000 UPI transactions with a 70:30 legitimate-to-fraud ratio. The dataset was temporally split into training (70%), validation (15%), and testing (15%) sets to simulate real-world deployment conditions. All experiments were executed in a controlled environment with fixed random seeds to ensure reproducibility.

- *Results Analysis*

The proposed hybrid fraud detection framework demonstrates strong performance across all evaluation metrics. The integration of rule-based screening, anomaly detection, supervised classification, and explainable AI contributes to improved accuracy and robustness compared to baseline methods.

Table 1. Performance Comparison of Fraud Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Rule-Based System	81.4	78.2	74.9	76.5	82.1
Isolation Forest	86.7	84.5	82.3	83.4	88.9
XG Boost	90.9	91.2	89.6	90.4	93.5
Proposed Hybrid Model	92.8	93.4	91.9	92.6	95.2

ROC Curve Analysis

The ROC curve analysis demonstrates that the proposed hybrid model achieves the highest area under the curve (AUC = 0.952), indicating superior discrimination capability between fraudulent and legitimate transactions. The curve shows consistent performance across varying threshold values, highlighting model robustness.

- *Ablation Study*

An ablation study was conducted to analyze the contribution of individual components within the proposed framework. Each component was incrementally removed, and performance variations were recorded.

1) *Table 2. Ablation Study Results*

Configuration	Accuracy (%)	AUC-ROC (%)	Performance Drop (%)
Full Hybrid Model	92.8	95.2	0.0
Without QR Reputation	89.6	91.1	3.2
Without Semantic Features	88.9	90.4	3.9
Without Anomaly Detection	87.8	89.2	5.0

- *Statistical Significance Analysis*

Statistical significance testing was performed using paired t-tests between the proposed hybrid model and baseline approaches. Results indicate that the improvements achieved by the proposed framework are statistically significant ($p < 0.01$), confirming that observed performance gains are not due to random variation.

- **Discussion**

The experimental results confirm that integrating behavioral, contextual, semantic, and QR reputation features significantly enhances fraud detection accuracy. The hybrid architecture effectively balances early risk screening with deep learning-based analysis, while SHAP-based explainability improves transparency and user trust. The ablation study highlights the importance of QR reputation scoring and semantic analysis in detecting social engineering and QR-based fraud scenarios.

LIMITATIONS OF THE PROPOSED SYSTEM

Despite the strong performance and robustness of the proposed explainable hybrid UPI fraud detection framework, certain limitations remain that provide scope for future enhancement.

1. Dependence on Synthetic Data

The current implementation relies on a large-scale synthetic dataset generated using fraud case studies and behavioral models. While this approach ensures privacy and controlled experimentation, it may not fully capture all real-world adversarial behaviors, regional fraud variations, or rapidly evolving scam strategies observed in live UPI ecosystems.

2. **Limited Real-Time Behavioral Sensors**

Some behavioral and contextual features, such as touch dynamics, app switching frequency, and call-state monitoring, require deeper integration with mobile operating systems. Due to platform-level permission restrictions, especially on commercial Android and iOS environments, real-time access to all such signals may be partially constrained.

3.**Explainability Overhead:**

Although SHAP-based explainability enhances transparency and trust, it introduces additional computational overhead. While optimized to remain under sub-100 ms latency, large-scale deployment with extremely high transaction throughput may require selective or on-demand explanation generation.

4. Generalization Across Regions and Banks:

UPI fraud patterns can vary across regions, banks, and user demographics. The proposed model, although designed to be adaptive, may require fine-tuning or retraining when deployed in regions with significantly different behavioral or linguistic characteristics.

5. Vulnerability to Coordinated Long-Term Attacks:

The QR reputation mechanism is designed to resist short-term manipulation; however, coordinated low-frequency fraud campaigns conducted over extended periods could gradually influence reputation scores. Mitigation of such slow poisoning attacks would require additional long-term anomaly correlation techniques.

6. Absence of Hardware-Level Trust Anchors:

The current system does not integrate hardware:-backed security mechanisms such as Trusted Execution Environments (TEE) or secure elements. Incorporating such components could further strengthen resistance against device-level compromise and malware-assisted fraud.

7. Dependency on User Feedback Quality

Continuous learning and reputation updates partially depend on user feedback and transaction outcomes. Inaccurate or delayed feedback may affect the speed at which the system adapts to emerging fraud patterns.

8. Scalability of Graph-Based QR Analysis

QR network centrality and graph-based reputation features may introduce scalability challenges when applied to extremely large transaction graphs without efficient graph partitioning or approximation techniques.

Overall, these limitations do not undermine the effectiveness of the proposed system but highlight practical constraints and research opportunities for future extensions, real-world deployment optimization, and regulatory-aligned implementation.

ETHICAL AND PRIVACY CONSIDERATIONS

This section outlines the ethical principles and privacy safeguards incorporated into the design, implementation, and evaluation of the proposed UPI fraud detection framework. Considering the sensitive nature of digital payment transactions, the system is developed in alignment with responsible AI practices to ensure user trust, transparency, and data protection.

- **Data Privacy and Confidentiality**

The proposed framework does not rely on real user financial data. All experimental evaluations are conducted using synthetically generated UPI transaction datasets that replicate realistic transaction patterns without including personally identifiable information. This approach eliminates risks associated with data exposure and unauthorized access.

- **User Consent and Data Minimization**

The system adheres to the principle of data minimization by processing only those attributes that are strictly required for fraud detection. In practical deployment scenarios, explicit user consent would be obtained prior to collecting behavioral or contextual signals, and users would be informed about the purpose and scope of data usage.

- **Fairness and Bias Mitigation**

To reduce the risk of biased decision-making, the model is evaluated across diverse transaction patterns, geographic regions, and simulated user profiles. Features that could directly or indirectly introduce discriminatory outcomes are excluded. Continuous monitoring and periodic audits are recommended to identify and mitigate bias during deployment.

- **Explainability and Transparency**

Explainable AI mechanisms are integrated into the framework to enhance transparency. SHAP-based explanations provide insights into the key factors influencing fraud decisions, enabling users, developers, and auditors to understand and verify system behavior.

- **Security and Responsible Use**

The system is designed with secure processing practices to prevent misuse. Fraud predictions are intended to support risk assessment and decision-making processes, rather than to serve as final judgments without human oversight. Human-in-the-loop validation is encouraged for critical decisions.

- **Regulatory Compliance**

The proposed framework aligns with general principles of financial data protection and ethical AI guidelines, including accountability, transparency, and user privacy. Although the study is experimental, the design considerations are compatible with regulatory requirements governing digital payment systems.

- **Ethical Limitations**

While the use of synthetic data addresses privacy concerns, it may not fully capture all real-world complexities. Ethical deployment of the system would require continuous evaluation, human oversight, and peririodic reassessment of its societal impact.

Applications of the Proposed System

The proposed explainable hybrid UPI fraud detection framework can be effectively applied across multiple domains within the digital payments ecosystem. Its real-time capability, transparency, and adaptive learning mechanisms make it suitable for both operational deployment and decision support.

- *Real-Time UPI Transaction Monitoring*

The system can be deployed within UPI payment platforms to monitor transactions in real time. By analyzing behavioral, contextual, semantic, and QR reputation features, it enables instant identification of fraudulent collect requests and malicious QR codes before transaction completion.

- *QR Code Payment Security*

The proposed framework enhances QR-based payment security by maintaining dynamic reputation scores for QR codes. This allows early detection of suspicious or compromised QR codes and prevents repeated fraud attempts involving the same QR source.

- *Fraud Risk Assessment and Decision Support*

The system can be used as a decision support tool for banks and payment service providers. Explainable AI outputs help fraud analysts understand risk factors behind flagged transactions, supporting informed and transparent decision-making.

- *Customer Protection and Awareness*

By identifying social engineering and PAY/RECEIVE confusion attacks, the system can generate context-aware warnings for users. This application improves customer protection by reducing fraud losses and increasing user awareness during high-risk transactions.

- *Regulatory Compliance and Auditing*

The explainability and logging capabilities of the proposed system make it suitable for regulatory compliance and audit processes. Financial institutions can use the generated

explanations and risk logs to demonstrate adherence to fraud prevention and ethical AI guidelines.

- *Fraud Pattern Analysis and Intelligence*

The framework can support fraud intelligence teams by identifying emerging fraud patterns through anomaly detection and continuous learning. Insights derived from transaction analysis can guide policy updates and preventive strategies.

- *Scalable Digital Payment Platforms*

Due to its low-latency design and modular architecture, the system is suitable for large-scale digital payment platforms handling high transaction volumes. It can be integrated with existing UPI infrastructures without significant performance overhead.

SECURITY ANALYSIS

The proposed explainable, human-centric hybrid machine learning framework for UPI fraud detection is designed with security as a foundational requirement. This section analyzes the system's resilience against common threat vectors, adversarial behaviors, and deployment-level risks in digital payment environments.

1. Data Security and Confidentiality

The system processes sensitive transaction, behavioral, and contextual data. To mitigate data leakage risks, all transaction data are anonymized and processed using feature abstractions rather than raw identifiers. Personally identifiable information (PII) is excluded from model training, and feature engineering is performed on derived metrics to ensure privacy preservation. Secure storage mechanisms and encrypted communication channels (TLS) are assumed during deployment.

2. Resistance to QR Code Manipulation Attacks:

The QR reputation management layer provides strong defense against QR-based fraud such as sticker fraud, QR replacement, and fake merchant impersonation. By continuously updating reputation scores using historical transaction outcomes, geographic consistency, and network centrality, the system prevents attackers from exploiting newly generated or low-trust QR codes. Temporal decay mechanisms further reduce the impact of short-lived or one-time fraud attempts.

3. Protection Against Social Engineering and Vishing Attacks

Behavioral and contextual feature analysis enables early detection of social engineering attacks, including vishing scenarios where victims are pressured during live calls. Features such as call-active status, decision delay anomalies, rapid approval after hesitation, and

app-switching behavior strengthen the system's ability to identify manipulation-driven fraud that bypasses traditional OTP-based security.

4. Model Robustness and Adversarial Resilience

The hybrid ensemble architecture enhances robustness by combining rule-based logic, unsupervised anomaly detection, and supervised classification. This layered design reduces susceptibility to adversarial evasion, as attackers must simultaneously bypass deterministic rules, anomaly detection boundaries, and learned decision surfaces. The Isolation Forest component improves resilience to zero-day fraud patterns not seen during training.

5. Explainability and Trust Assurance

Integration of SHAP-based explainable AI strengthens security from a governance and compliance perspective. Transparent feature attribution allows system operators, auditors, and regulators to verify that fraud decisions are based on legitimate behavioral and contextual indicators rather than biased or spurious correlations..

6. Defense Against Model Poisoning and Feedback Abuse

The feedback and continuous learning module incorporates weighted updates that balance user feedback with model confidence. This design limits the impact of malicious feedback injection or label poisoning attacks. Reputation updates and retraining triggers are threshold-based, preventing abrupt model drift caused by isolated or manipulated inputs.

7. Availability and Real-Time Security

The system is optimized for sub-100 ms inference latency and supports high-throughput transaction processing. Lightweight rule-based screening ensures early rejection of high-risk transactions with minimal computational overhead, reducing exposure to denial-of-service (DoS) style attacks targeting model inference pipelines.

8. Deployment and Infrastructure Security Considerations

While the current implementation focuses on application
level security, the architecture is compatible with secure deployment practices such as container isolation, role-based access control (RBAC), and hardware-backed security extensions. These measures can further strengthen resistance against device compromise and infrastructure-level attacks.

Overall, the proposed system demonstrates strong security posture through layered defense, adaptive trust mechanisms, explainability, and real-time resilience. While no fraud detection system can guarantee absolute security, the multi-dimensional and

explainable design significantly raises the attack complexity, making large-scale and persistent fraud economically and operationally infeasible.

Conclusion:

This work presented an explainable, human-centric hybrid machine learning framework for detecting fake UPI collect requests and QR code–based fraud. By integrating behavioral, contextual, semantic, and QR reputation features, the proposed system effectively addresses limitations of traditional rule-based and single-model fraud detection approaches. The hybrid architecture, combining rule-based screening, anomaly detection, and supervised classification, enables accurate detection of both known and emerging fraud patterns while maintaining real-time performance.

Experimental evaluation on a large-scale synthetic dataset demonstrated high detection accuracy and strong discriminative capability with sub-100 ms inference latency, indicating suitability for real-time UPI transaction environments. The integration of SHAP-based explainable AI provides transparent, human-interpretable explanations for fraud decisions, supporting trust, auditability, and regulatory compliance. Overall, the proposed framework offers a practical and scalable solution for secure and explainable fraud detection in modern digital payment system.

FUTURE WORK

Although the proposed system demonstrates strong performance, several directions remain for future research and enhancement:

- **Real-World Deployment and Validation:** Future work will focus on evaluating the framework using real transaction data from banking or payment service providers to further validate robustness and generalization.
- **Enhanced Behavioral Sensing:** Integration of richer device-level and interaction-based signals, subject to platform permissions, can improve behavioral modeling accuracy.
- **Scalability Optimization:** Efficient graph-processing and approximation techniques will be explored to scale QR reputation analysis to very large transaction networks.
- **Adversarial Robustness:** Additional defenses against coordinated long-term fraud campaigns and adversarial manipulation of feedback mechanisms will be investigated.
- **Hardware-Assisted Security:** Incorporating trusted execution environments or hardware-backed security features can further strengthen protection against device-level compromise.

- **Cross-Regional Adaptation:** Domain adaptation techniques will be studied to improve performance across diverse geographic, linguistic, and user-demographic settings.

These extensions aim to further enhance the system's applicability, resilience, and effectiveness in large-scale real-world UPI deployments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the management of **Meenakshi college of Engineering** for providing the necessary facilities and support to carry out this work. We extend our heartfelt thanks to the Head of the Department and faculty members of the Department of Computer Science and Engineering for their continuous guidance, valuable suggestions, and encouragement throughout the course of this project. We also acknowledge the support received from peers and friends who contributed through constructive discussions and feedback. Finally, we are thankful to all individuals who directly or indirectly assisted in the successful completion of this work.

REFERENCES:

1. Y. Zhou, B. Hu, Y. Zhang and W. Cai, "Implementation of Cryptographic Algorithm in Dynamic QR Code Payment System and Its Performance," IEEE Access, vol. 9, pp. 122362–122372, 2021.
2. A. Karpaga Selvi, M. Pavithra and J. Sindhuja, "Enhancing UPI Fraud Detection Accuracy Using Isolation Forest: A Novel Machine Learning Approach," in Proc. Int. Conf. on Intelligent Systems, 2025.
3. D. M. Rao, R. S. V. Talluri, T. V. Gupta and D. Manoj Kumar, "Comparative Analysis of UPI Fraud Detection Using Ensemble Learning," in Proc. IEEE Int. Conf. on Computing and Communication, 2025.
4. "AI Based QR Code Fraud Detection System," International Journal of Innovative Research in Science, Engineering and Technology, vol. 14, no. 12, pp. 2451–2458, 2025.
5. S. Kumar and R. Mehta, "Mobile Payment Fraud Detection in UPI Using Machine Learning: A Systematic Review," ACM Computing Surveys, vol. 57, no. 2, pp. 1–36, 2025.
6. R. Sharma, P. Verma and A. Singh, "Machine Learning-Based UPI Fraud Detection," Atlantis Press, Advances in Intelligent Systems Research, vol. 182, pp. 95–102, 2025.
7. K. P. Reddy and T. V. S. S. Swathi, "AI-Powered Fraud Detection Awareness for UPI Transactions," IRE Journals, vol. 8, no. 6, pp. 34–41, 2025.

8. M. Joshi and S. Kulkarni, "UPI Scam Detection Using QR Code Analysis," International Research Journal of Engineering and Technology, vol. 12, no. 5, pp. 1123–1129, 2025.
9. Z. Ke et al., "Detection of AI-Driven Payment Fraud Using GAN-Based Models," IEEE Access, vol. 13, pp. 45821–45835, 2025.
10. Q. Sha et al., "Graph Neural Networks for Financial Fraud Detection," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 4, pp. 4512–4524, 2025.

Contact Email: gayathrirgsd@gmail.com

DOI: https://doi.org/10.54121/202111521