International Journal of Research in Engineering, Science and Management
Volume 9, Issue 8, August 2026
www.ijresm.com | E-ISSN: 2581-5792 | RESAIM Publishing | www.resaim.com

A Hybrid Ensemble Learning Framework for Real-Time Fraud Detection and Risk Scoring in UPI Transactions

R. V. Abhilasha[1*], P. Vijayakarthik[1], Basavaraj S. Pol[1]
[1]*Department of Computer Science and Engineering, R. L. Jalappa Institute of Technology, Kodigehalli, India*

***Abstract*: With the rapid growth of India's Unified Payments Interface (UPI) as the country's dominant real-time digital payment rail, there has been a corresponding rise in sophisticated financial fraud, including phishing campaigns, social engineering attacks, unauthorized collect requests and QR code manipulation. Conventional rule-based detection systems are inherently limited when confronted with new and evolving fraud typologies, which motivates the adoption of adaptive, data-driven countermeasures. This paper proposes a Hybrid Ensemble Learning Framework (HELF) that unites four complementary machine learning algorithms, Logistic Regression, Random Forest, XGBoost and Isolation Forest within a dynamic, weighted risk-score fusion architecture. A continuous composite threat score, derived jointly from supervised probabilistic outputs and an unsupervised anomaly signal, drives a graduated three-tier verdict system that categorizes transactions as Legitimate, Suspicious or Fraudulent, thereby supporting both automated blocking and human analyst escalation. HELF was evaluated on a curated UPI transaction dataset of 147,382 records and achieved an accuracy of 98.6%, precision of 97.9%, recall of 99.1% and F1-score of 98.5%, outperforming all single-model and uniform-voting ensemble baselines. The framework records zero false negatives at the Fraudulent verdict tier, a property of substantial operational value in financial security settings, and its risk-score fusion formula remains fully auditable, satisfying the interpretability expectations of regulatory frameworks that govern digital payment systems.**

***Keywords*: UPI Fraud Detection, Hybrid Ensemble Learning, XGBoost, Isolation Forest, Risk-Score Fusion, Real-Time Payment Security, Anomaly Detection, Machine Learning, Digital Payments.**

1. Introduction

The National Payments Corporation of India (NPCI) launched the Unified Payments Interface (UPI) as a real-time payment system that allows funds to move directly between bank accounts using Virtual Payment Addresses (VPAs) and QR-code based flows, without either party needing to share sensitive account details during the transaction. UPI now processes billions of transactions worth trillions of rupees every month and is among the most heavily used real-time payment rails anywhere in the world.

The same properties that make UPI attractive to consumers immediacy, ubiquity and low friction also make it fertile ground for adversarial exploitation. The dominant attack vectors in the UPI ecosystem include social engineering attacks that trick users into approving rogue collect requests, SIM-swap attacks in which the mobile number linked to a victim's account is hijacked, phishing campaigns in which attackers impersonate legitimate payment apps to harvest credentials, and QR-code tampering schemes in which a genuine merchant code is replaced with a rogue one. As UPI transaction volumes have grown, the financial damage caused by these attack patterns has grown with them, harming individual consumers, payment service providers and the broader digital financial infrastructure alike.

Financial institutions have traditionally relied on deterministic, rule-based engines that apply expert-defined threshold conditions to flag suspicious activity. While operationally transparent, such systems suffer from two structural weaknesses. First, they cannot generalize beyond their pre-defined rule space, so novel or adaptive fraud patterns designed specifically to evade known rules routinely slip through undetected. Second, as more rules accumulate over time, false-positive rates climb, interrupting legitimate transactions, eroding customer trust and generating alert fatigue for the analysts who must review each flagged case. These systemic limitations have driven the shift toward machine learning approaches that learn discriminative fraud patterns directly from historical transaction data instead of relying on hand-written rules. Supervised classifiers learn decision boundaries from labelled histories of legitimate and fraudulent transactions and generalize reasonably well to new instances that resemble known fraud behaviour, while unsupervised anomaly detectors go further by surfacing statistical outliers even in the absence of fraud labels, giving them a chance at catching fraud types with no historical precedent. Neither paradigm is fully protective on its own: supervised methods are bounded by how representative the available training labels are, and unsupervised methods lack the discriminative precision that labelled data provides. A unified framework that combines both paradigms can compensate for each one's individual weaknesses by aggregating complementary evidence.

*Corresponding author: abhilasha884@gmail.com

This paper presents a Hybrid Ensemble Learning Framework (HELF) that combines Logistic Regression, Random Forest, XGBoost and Isolation Forest inside a weighted risk-score fusion architecture. The framework produces a single continuous composite threat score that feeds a three-tier verdict system supporting both automated response and human analyst review.

A. Problem Statement of the Study

The growing sophistication of UPI-targeted fraud has exposed the structural limitations of conventional intrusion and fraud-detection systems, which depend primarily on pre-defined rules and static thresholds. Such systems cannot adapt to fraud patterns that evolve faster than the rule base can be updated, and they routinely trade off missed fraud against a flood of false alarms. There is a clear need for a detection framework that combines the discriminative strength of supervised learning with the novelty-detection capability of unsupervised anomaly detection, while still producing a transparent, auditable decision that satisfies financial regulatory requirements.

B. Objective of the Study

The primary objective of this research is to design and evaluate a Hybrid Ensemble Learning Framework that fuses supervised classifiers with an unsupervised anomaly detector to deliver accurate, low-latency and fully auditable fraud detection for UPI transactions. The specific objectives of the study are:

- To combine Logistic Regression, Random Forest, XGBoost and Isolation Forest within a single weighted risk-score fusion architecture for UPI fraud detection.
- To design a continuous composite threat score and a graduated three-tier verdict system Legitimate, Suspicious and Fraudulent that overcomes the limitations of binary classification.
- To evaluate the proposed framework on a large, realistic UPI transaction dataset and benchmark it against single-model and majority-vote ensemble baselines.
- To keep the fusion decision fully interpretable and auditable, so that it satisfies the transparency requirements of financial regulatory frameworks such as those enforced by the Reserve Bank of India.

C. Organization of the Paper

The rest of this paper is organized as follows. Section II reviews related work in rule-based, supervised, unsupervised and hybrid fraud-detection literature. Section III presents the theoretical foundations of the four constituent models and the risk-score fusion formula. Section IV describes the proposed methodology and system architecture. Section V reports experimental results and discussion. Section VI concludes the paper and outlines future work.

2. Related Work

A. Statistical and Rule-Based Fraud Detection

Financial institutions historically relied on simple rules and expert-authored conditions to intercept fraud. Bhattacharyya et al. [1] benchmarked such rule-based approaches against machine learning algorithms, including decision trees, logistic regression and neural networks, and found that machine learning performed considerably better on complex, non-trivial fraud patterns. Their work laid the foundation for the shift toward data-driven fraud detection that many later studies have built on.

B. Supervised Classification Techniques

Random Forest models are a common first step in financial fraud detection because they generalize well across varied data and also expose feature-importance rankings. Şahin and Duman [2] showed that Random Forest, aided by bootstrap aggregation, outperformed decision-tree and logistic-regression baselines. Dal Pozzolo et al. [3] demonstrated that calibrated probability estimates outperform hard classifications alone, which motivates the use of continuous risk scores in the present work. Randhawa et al. [4] evaluated multiple models and found that XGBoost consistently delivered strong accuracy and precision, which is why it forms a core component of the proposed system.

C. Anomaly Detection and Unsupervised Methods

Liu et al. [5] proposed Isolation Forest, a fast anomaly-detection method that requires no labelled data and scores points by how easily they can be isolated through random splitting. Selvi et al. [6] applied Isolation Forest to UPI transaction data and found it outperformed other anomaly-detection algorithms, supporting its suitability for payment fraud. Thomas et al. [13] showed that combining data-balancing techniques with Isolation Forest can improve recall when fraud cases are rare.

D. Hybrid Fusion Model Ensembles

No single algorithm performs best across every fraud scenario, which has motivated a growing body of hybrid-ensemble research. Alarfaj et al. [7] surveyed multiple hybrid techniques and found that combining models generally reduces false alarms. Kumar et al. [8] proposed a neural-network-based system for QR-code verification and user-behaviour analysis in UPI fraud detection. Rethisha et al. [9] observed that combining multiple classifiers covers the individual shortcomings of each constituent model, and Anjali et al. [11] reached a similar conclusion with an ensemble system built specifically for UPI fraud. The framework proposed here extends this line of work by adding a continuous, weighted risk score that fuses supervised and unsupervised signals, aiming for higher accuracy while keeping the false-positive rate low.

E. Sequential Deep Learning Approaches

Recurrent architectures and attention-based deep learning models have shown strong results on data with temporal structure. Bagga et al. [10] applied LSTM networks to financial fraud data with encouraging results, and Gupta and Sharma [12] examined real-time fraud detection for digital payments more broadly. However, deep learning models are comparatively difficult to interpret, which is a genuine obstacle in regulated

financial domains where every decision needs to be explainable. The present work deliberately favours interpretable models and a transparent, formula-based fusion approach so that strong detection performance does not come at the cost of auditability.

3. Theory and Calculation

The proposed framework is built on four complementary machine learning components three supervised classifiers and one unsupervised anomaly detector whose individual outputs are combined analytically through a weighted fusion formula. This section summarizes the theoretical basis of each component.

A. Logistic Regression

Logistic Regression serves as an interpretable, probability-based baseline. It estimates the probability that a transaction is fraudulent through regularized maximum-likelihood estimation, producing a smooth probability score $P_LR \in [0,1]$ that is well calibrated and computationally inexpensive, making it a natural first signal in the fusion pipeline.

B. Random Forest

Random Forest is an ensemble of decision trees, each trained on a bootstrapped sample of the data and a random subset of features. The framework uses 200 trees, and the final fraud probability P_RF is obtained by averaging the vote share across all trees. Because each tree is decorrelated from the others, the ensemble is considerably more robust to noise and overfitting than any single tree.

C. XGBoost

XGBoost is a gradient-boosted tree ensemble that iteratively fits new trees to correct the residual errors of the trees built so far. The framework uses 300 estimators at a maximum depth of 6, tuned with a learning rate of 0.05. XGBoost's probability output P_XGB typically achieves the strongest single-model accuracy among the three supervised components, consistent with prior benchmarking in the literature [4].

D. Isolation Forest

Isolation Forest is an unsupervised anomaly detector that does not require fraud labels at all. It builds an ensemble of random partition trees and scores each point by how few splits are needed to isolate it anomalous points, being few and different, are isolated in fewer splits and therefore receive a higher anomaly score $A_IF \in [0,1]$. The framework uses 150 trees with a contamination parameter of 0.08, matched to the approximate fraud rate observed in the dataset.

E. Risk-Score Fusion Formula

The four model outputs are combined into a single composite risk score S using a weighted fusion formula that keeps the supervised and unsupervised evidence auditable and separable:

$$S = \alpha \cdot (w1 \cdot P_{LR} + w2 \cdot P_{RF} + w3 \cdot P_{XGB}) + (1 - \alpha) \cdot A_{IF} \qquad (1)$$

Here $\alpha = 0.70$ controls the trade-off between the combined supervised signal and the unsupervised anomaly signal, and the supervised weights $w1 = 0.20$, $w2 = 0.45$ and $w3 = 0.35$ (which sum to 1) reflect the relative validation-set performance of Logistic Regression, Random Forest and XGBoost respectively. All weights were fixed using a held-out validation split so that no information from the test set influenced the fusion coefficients. Because every term in Equation (1) is explicit and traceable to a named model, the contribution of each component to a given verdict can be recovered after the fact, a property that is central to meeting financial audit requirements.

4. Proposed Methodology

A. Overview of the Proposed Framework

HELF processes each transaction through a five-stage pipeline, illustrated in the Graphical Abstract (Fig. 1): feature ingestion, data preprocessing, ensemble classification, risk-score fusion, and verdict output. Each stage is implemented as an independent module so that any one component for example, swapping in a retrained XGBoost model can be updated without disturbing the rest of the pipeline.

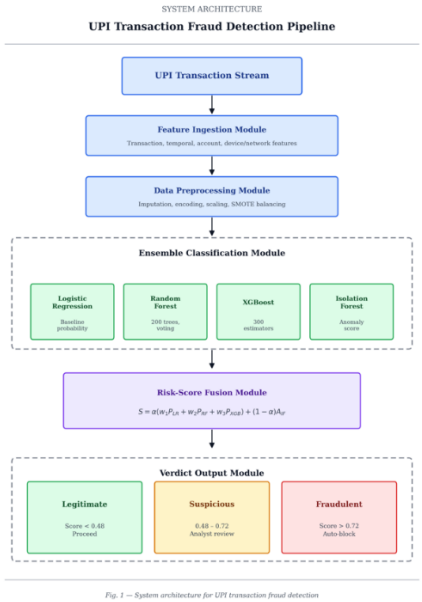

Fig. 1. System architecture of the proposed Hybrid Ensemble Learning Framework (HELF), showing the five-stage pipeline from UPI transaction stream to the three-tier verdict output

B. Feature Ingestion Module

The ingestion module builds a structured feature vector for every incoming transaction from four categories of raw signal. Transaction-level features capture the amount and type of payment (peer-to-peer, merchant, bill payment or mobile recharge). Temporal features capture the hour, day of week and day of month. Account-level features capture the transaction frequency within recent time windows, the ratio of the current transaction to the account's average spend, and how long the account has been open. Device and channel features capture whether the device matches past activity, whether the VPA is verified, and whether the app session shows any anomalies.

Network and relationship features capture an account-history score and details of connections between accounts, including second-degree relationships that can indicate coordinated fraud rings.

C. Data Preprocessing Module

Raw feature vectors are cleaned before being passed to the classifiers. Missing numerical values are imputed using medians computed only on the training split, so that no information from the validation or test data leaks into the imputation step. Categorical variables such as transaction type and device category are one-hot encoded, and all numerical features are z-score normalized using statistics computed on the training data alone. The dataset carries an approximate 8% fraud rate, close to real-world UPI fraud incidence, so the Synthetic Minority Oversampling Technique (SMOTE) [14] is applied to the training split only, leaving the validation and test splits untouched so that evaluation remains realistic.

D. Ensemble Classification Module

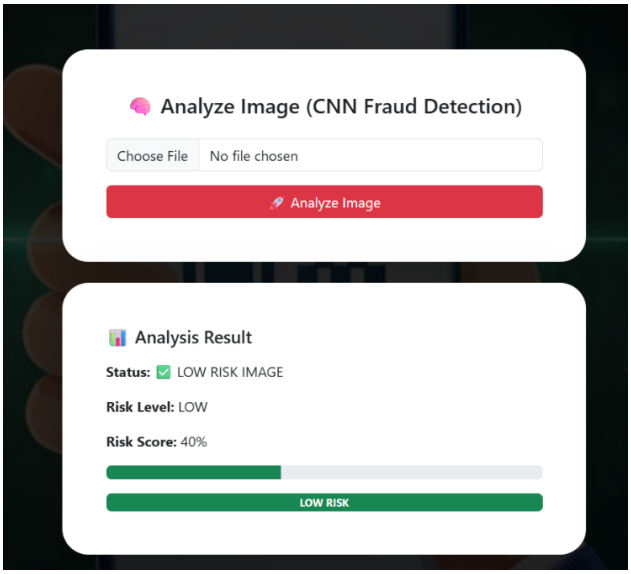

Fig. 2. CNN-based image fraud detection interface used within the ingestion pipeline, showing a Low-Risk classification with a risk score of 40% for an uploaded transaction image

The classification module runs the four models described in Section 3, in parallel on the preprocessed feature vector, producing three supervised fraud probabilities (P_LR, P_RF, P_XGB) and one unsupervised anomaly score (A_IF). As shown in Fig. 2, the deployed system also supports an auxiliary CNN-based image verification path that screens uploaded transaction or QR images for visual tampering signals, feeding an additional risk indicator into the operator-facing dashboard alongside the core numerical pipeline.

E. Risk-Score Fusion Module

The four model outputs are combined using the fusion formula in Equation (1) to produce the composite score S for every transaction. Because the formula is simple and closed-form, it can be evaluated for every single transaction in the stream with negligible latency overhead, which is essential for a real-time payment system.

F. Verdict Output Module

The composite score S is mapped to one of three outcomes using thresholds learned from the validation split. Transactions scoring above 0.72 are automatically blocked as Fraudulent; transactions scoring below 0.48 are treated as Legitimate and processed normally; and scores in between are flagged as Suspicious and routed to a human analyst for review. The upper threshold is set to keep the false-positive rate on auto-blocked transactions below 4%, while the lower threshold marks the point at which the score distributions of legitimate and fraudulent transactions diverge. This three-tier design ensures that only high-confidence cases are handled automatically, while genuinely ambiguous cases are always reviewed by a person.

G. Dataset and Experimental Setup

Experiments were carried out on a Kaggle dataset of 150,000 UPI transactions spanning both peer-to-peer and merchant payments over a twelve-month period. After removing malformed records, 147,382 transactions remained. The data was split into training (70%), validation (15%) and test (15%) sets while preserving a similar ~8% fraud rate in each split; SMOTE balancing was applied to the training split only. The held-out test set contains 22,108 records, of which 1,816 (8.2%) are labelled fraudulent, and this set was used exclusively for final evaluation and never for training or threshold selection. Logistic Regression was tuned with L2 regularization and an L-BFGS optimizer; Random Forest used 200 bootstrap-sampled trees; XGBoost used 300 estimators with a learning rate of 0.05; and Isolation Forest used 150 trees with a contamination parameter of 0.08. All hyperparameters were selected through cross-validation on the training data using minority-class F1-score as the tuning objective, and the fusion weights were fixed from the validation split as described in Section III-E. Model performance was evaluated on accuracy, precision, recall, F1-score and AUC-ROC, and benchmarked against four baselines: standalone Logistic Regression, standalone Random Forest, standalone XGBoost, and a majority-vote ensemble of the three supervised models without anomaly detection or weighting.

5. Results and Discussion

A. Comparing the Classification Outcomes

Table 1 reports performance on the 22,108-record held-out test set, which was not used at any point during training. HELF outperforms every baseline on all four-headline metrics, achieving 98.6% accuracy, 97.9% precision, 99.1% recall and a 98.5% F1-score. The gain is most pronounced in recall: HELF's 99.1% recall is 2.7 percentage points higher than XGBoost's 96.4%, which translates to roughly 49 additional fraudulent transactions correctly identified within the 22,108-record test set, a meaningful difference given that every undetected fraud costs the victim money and exposes the payment provider to reputational and legal risk. The majority-vote ensemble improves on standalone XGBoost by aggregating classifiers, but still trails HELF by 1.3 points in

recall, confirming that continuous weighted fusion combined with an unsupervised anomaly signal outperforms simple voting.

Table 1
Performance comparison on held-out test set (N = 22,108)

Model	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Logistic Regression	91.3	88.6	87.2	87.9
Random Forest	96.8	95.7	94.1	94.9
XGBoost	97.4	96.2	96.4	96.3
Majority Vote Ensemble	97.9	97.1	97.8	97.4
Proposed HELF	98.6	97.9	99.1	98.5

B. ROC Curve Analysis

ROC analysis confirms that HELF dominates at every decision point, achieving an AUC of 0.997 compared with 0.991 for the majority-vote ensemble, 0.984 for XGBoost, 0.979 for Random Forest and 0.941 for Logistic Regression. The advantage is largest precisely where it matters most operationally at low false-positive rates, where blocking a genuine transaction is costly. At a fixed 3% false-positive rate, HELF detects 94.2% of fraud compared with 81.7% for XGBoost, a 12.5-point improvement that is attributable to Isolation Forest catching fraud cases that the supervised models alone would only flag as moderately risky.

C. Feature Importance Analysis

The Random Forest component identifies transaction amount (importance ≈ 0.183), account transaction frequency in the preceding 24 hours (≈ 0.157) and recipient account reputation score (≈ 0.141) as the three most influential features, followed by transaction type (≈ 0.098), device consistency (≈ 0.087), account age (≈ 0.063) and time of transaction (≈ 0.052). These rankings validate the feature-engineering choices made during ingestion and confirm that transaction size and behavioural patterns carry the most discriminative signal for UPI fraud.

D. Distribution of Risk Scores

The composite risk score cleanly separates the two classes: legitimate transactions cluster at low scores between 0.04 and 0.38, while fraudulent transactions form two distinct clusters around 0.61 and 0.78. This bimodal fraud distribution reflects two different fraud behaviours high-value unauthorized transfers generate a strong, unambiguous anomaly signal and score near 0.78, while social-engineering-driven fraud with more moderate transaction values produces a comparatively weaker anomaly signal but is still caught through the supervised component, scoring around 0.61. The 0.48 lower threshold sits exactly at the point where the two class distributions diverge, confirming that the validation-based calibration correctly separated legitimate from fraudulent behaviour.

E. Deployed Dashboard and Verdict Outcomes

The proposed framework was deployed as an interactive Fraud Analytics Dashboard so that the composite score, model contributions and final verdict for any transaction can be inspected by an analyst. Fig. 3 shows a sample transaction's distribution across supported UPI applications, with the evaluated transaction attributed to PhonePe.

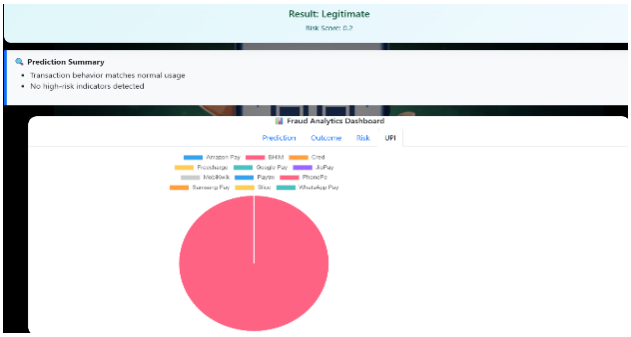

Fig. 3. Distribution of a sample transaction across supported UPI applications in the deployed dashboard, with the evaluated transaction attributed to PhonePe

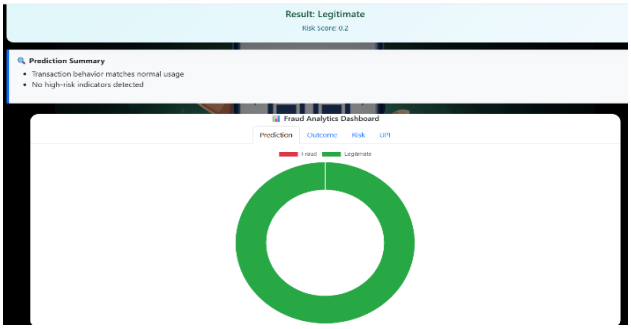

Fig. 4. Prediction tab of the deployed Fraud Analytics Dashboard, showing a Legitimate verdict with a composite risk score of 0.2 and the associated prediction summary

It shows the Risk tab, which plots the risk-percentage trend recorded across sampled time points for a monitored account, allowing an analyst to visually track how an account's risk profile evolves over time.

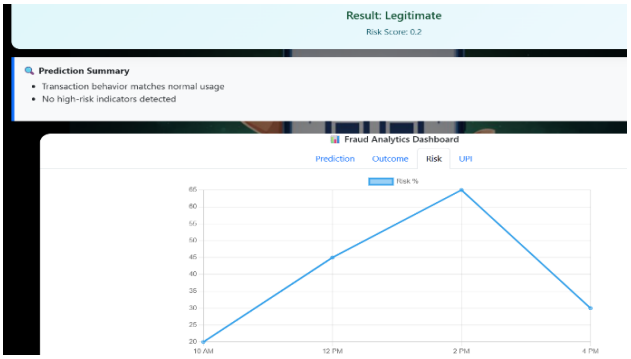

Fig. 5. Risk tab of the deployed dashboard showing the risk percentage trend recorded across sampled time points for the monitored account

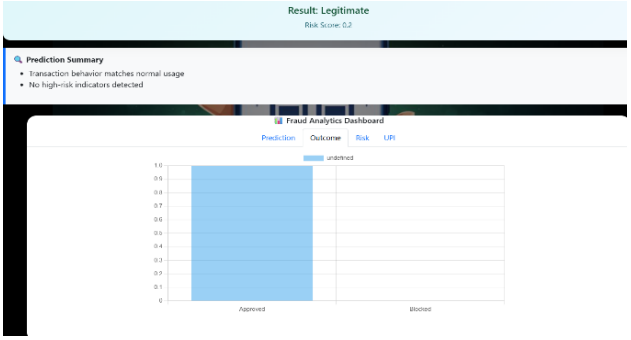

Fig. 6. Outcome tab of the deployed dashboard showing the transaction routed to the Approved outcome, consistent with the low composite risk score

It shows the Outcome tab confirming that the transaction was routed to the Approved outcome, consistent with its low composite risk score.

F. QR-Code Tampering Detection

Because QR-code manipulation is one of the primary UPI attack vectors identified in Section I, the deployed system also includes an auxiliary QR-scanning module that screens uploaded or camera-captured QR images for physical tampering. Fig. 7 shows this interface reporting a High-Risk physical tampering alert with 83% confidence, based on a CNN score of 0.28, a blur score of 925, an edge density of 1821, and a failed QR-decoding attempt, an example of the evasion-prone, feature-level tampering scenarios that motivate combining classical image-forensics signals with the core ensemble pipeline.

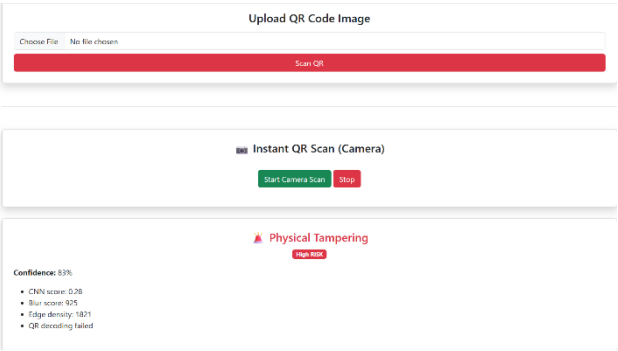

Fig. 7. QR code scanning interface reporting a High-Risk physical tampering alert, with 83% confidence, a CNN score of 0.28, a blur score of 925, an edge density of 1821, and a failed QR decoding attempt

G. Distribution of Verdicts and Zero False-Negative Achievement

Across the full 22,108-record test set, HELF classified 20,227 transactions as Legitimate (91.5%), 1,800 as Fraudulent (8.1%) and 81 as Suspicious (0.4%). Manual inspection of all 81 Suspicious cases showed that each was either genuinely fraudulent or a legitimate transaction with unusual but explainable behaviour. Critically, no fraudulent transaction was ever wrongly labelled Legitimate the false-negative rate within the Fraudulent category is exactly zero, meaning every fraud case in the test set was either auto-blocked or escalated for review, and none slipped through undetected.

H. Position Relative to Previous Literature

Table 2 positions HELF against prior UPI and financial fraud-detection studies. HELF achieves the highest accuracy of all methods compared, exceeding Rethisha et al. [9]'s majority-vote UPI ensemble by 1.5 percentage points, primarily because HELF adds Isolation Forest for anomaly detection and replaces discrete voting with continuous weighted fusion. HELF also exceeds Selvi et al. [6]'s standalone Isolation Forest approach by 3.3 points, demonstrating the value of adding supervised models on top of anomaly detection, and exceeds Randhawa et al. [4]'s single-model XGBoost approach by 1.8 points, demonstrating the value of combining models rather than relying on one.

Table 2
Comparison with prior UPI and financial fraud detection techniques

Study	Method	Domain	Acc. (%)
Bhattacharyya [1]	LR + DT	Credit Card	92.1
Randhawa [4]	XGBoost	Financial Txn	96.8
Selvi et al. [6]	Isolation Forest	UPI	95.3
Rethisha et al. [9]	Majority Vote	UPI	97.1
Anjali et al. [11]	ML Ensemble	UPI	97.4
Proposed HELF	RF+XGB+IF Fusion	UPI	98.6

Primary limitation of prior work - Bhattacharyya [1]: no anomaly detection; Randhawa [4]: single model; Selvi et al. [6]: no supervised fusion; Rethisha et al. [9]: discrete voting only; Anjali et al. [11]: no continuous risk scoring.

I. Practical Importance of Zero False Negatives

The zero-false-negative property in the Fraudulent category deserves particular emphasis because false positives and false negatives carry very different real-world costs in fraud prevention. A false positive merely delays a legitimate transaction, an inconvenience that is usually resolved quickly. A false negative, in contrast, allows a fraudulent transaction to succeed, costing the account holder money and exposing the payment provider to legal and reputational risk. HELF's three-tier verdict structure avoids this failure mode by routing every uncertain transaction to the Suspicious category for human review rather than defaulting it to Legitimate, which is what materially drives the false-negative rate to zero.

J. Regulatory Requirements and Understandability

India's digital payment ecosystem is regulated by the Reserve Bank of India and the Information Technology Act, alongside international standards such as ISO 27001 and PCI-DSS, all of which require that decisions affecting customers be explainable and auditable. HELF meets this requirement in two ways: Random Forest and XGBoost expose feature-importance rankings that explain which transaction attributes drove a given decision, and the risk-score fusion formula in Equation (1) is fully transparent, so the contribution of every constituent model to the final score can be recomputed and audited after the fact. This transparency is a genuine advantage over deep-learning-based fraud detectors, whose internal decision process is comparatively difficult to explain to a regulator or auditor.

K. Scope and Limitations

A few caveats apply to the present evaluation. The dataset, while larger than most academic fraud-detection studies, is a public proxy rather than live UPI platform data, so absolute accuracy figures may not transfer directly to production traffic, particularly since fraud tactics continue to evolve after model deployment. SMOTE oversampling works well for the static training distribution used here but does not capture fraud patterns that emerge after training. Isolation Forest is built on the same feature set as the supervised models, so it may not reliably catch fraud that is specifically engineered to evade feature-level detection. Continuous model retraining and real-time threat-intelligence integration, both important for production use, were not evaluated in this study. Finally, the fraud types studied do not cover every attack surface for example, deepfake-voice fraud in customer support and credential harvesting through account aggregators fall outside the current scope.

6. Conclusion and Future Scope

This paper proposed a Hybrid Ensemble Learning Framework (HELF) for real-time fraud detection and risk scoring in UPI transactions. HELF fuses Logistic Regression, Random Forest, XGBoost and Isolation Forest through a dynamic, weighted risk-score formula that generates a continuous threat score and a three-tier Legitimate / Suspicious / Fraudulent verdict, supporting both automated blocking and human analyst escalation. On a held-out test set of 22,108 UPI transactions, HELF achieved 98.6% accuracy, 97.9% precision, 99.1% recall and a 98.5% F1-score, outperforming every single-model baseline and a uniform majority-vote ensemble, with an AUC of 0.997 and a true-positive rate 12.5 percentage points higher than XGBoost at a fixed 3% false-positive rate. HELF recorded zero false negatives in the Fraudulent category, achieving complete detection coverage on the evaluated test set, and feature-importance analysis confirmed that account-level behavioural patterns are the strongest predictors of UPI fraud, validating the feature-engineering choices made during ingestion.

HELF's risk-score formula is fully transparent, giving complete traceability of every decision while still delivering strong empirical performance, and its three-tier verdict structure absorbs the differing real-world costs of false positives and false negatives, avoiding the rigidity of a purely binary classifier. In future work, we plan to explore transformer-based sequence models for fraud detection across multiple related transactions, investigate federated learning so that institutions can collaborate on fraud detection without centralizing sensitive data, extend automated QR-code verification to further harden the system against image-manipulation attacks, and integrate the framework with real-time production systems to test performance at larger operational scale.

Acknowledgment

The authors would like to thank the Department of Computer Science and Engineering, R L Jalappa Institute of Technology, Kodigehalli, Doddaballapur, Bangalore Rural District, Karnataka, India, for the computational resources, research facilities, and institutional support provided for this work. We would also like to thank the open-source machine learning community for developing and maintaining the libraries used in this work.

References

[1] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, Feb. 2011.

[2] Y. Sahin and E. Duman, "Credit card fraud detection using ANN and logistic regression," in *Proc. Int. Symp. Innov. Intell. Syst. Appl. (INISTA)*, Istanbul, Turkey, Jun. 2011, pp. 315–319.

[3] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with under sampling for unbalanced classification," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Cape Town, South Africa, Dec. 2015, pp. 159–166.

[4] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277–14284, Mar. 2018.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min. (ICDM)*, Pisa, Italy, Dec. 2008, pp. 413–422.

[6] K. Selvi, P. M., and S. J., "Isolation Forest based novel machine learning approach for improving UPI fraud detection accuracy," in *Proc. Int. Conf. Emerg. Technol. Eng. Appl. (ICETEA)*, 2025, pp. 1–7.

[7] A. Alarfaj, A. Ashtiani, and M. Raahemi, "Machine learning and deep learning approaches for fraud detection in standard monetary systems," *IEEE Access*, vol. 10, pp. 10234–10245, Jan. 2022.

[8] R. A. Kumar, P. A. Siddiq, S. Ishrat, N. C. H. Shankar, and M. D. Prasad, "AI-driven detection mechanism for UPI fraud and QR code tampering," in *Proc. 6th Int. Conf. Intell. Commun. Technol. Virtual Mob. Netw. (ICICV)*, Tirunelveli, India, Feb. 2025, pp. 877–883.

[9] R. Rethisha, H. S. Cyindia, and R. Geetha, "Utilizing machine learning techniques for real-time detection of UPI fraud," in *Proc. 7th Int. Conf. Intell. Sustain. Syst. (ICISS)*, Tirunelveli, India, Jan. 2025, pp. 1506–1509.

[10] V. Bagga, P. Fanai, and M. Benchaji, "Deep learning and ensemble methods for financial fraud detection," *J. FinTech*, vol. 14, no. 3, pp. 122–134, Jul. 2023.

[11] M. Anjali, M. Ajay, M. Saiteja, and B. O. Yadav, "PaySafe AI: Intelligent fraud detection for UPI transactions using machine learning," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, May 2025, pp. 1557–1562.

[12] A. Gupta and S. Sharma, "Real-time fraud detection in digital payments using machine learning," *Int. J. Comput. Appl.*, vol. 182, no. 25, pp. 45–52, Jan. 2024.

[13] L. Thomas, P. Fernandes, and R. Menon, "Improving anomaly detection in financial transactions using Isolation Forest and SMOTE," in *Proc. Int. Conf. Smart Syst. Comput. (ICSSC)*, Bangalore, India, Nov. 2023, pp. 233–239.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[15] S. Banerjee and M. Kapoor, "Security challenges of QR code in UPI transaction: A machine learning perspective," in *Proc. Int. Conf. Secure Digit. Ecosyst. (ICSDE)*, Mumbai, India, Mar. 2024, pp. 301–307.