International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 10 | Oct 2026 www.irjet.net p-ISSN: 2395-0072

A Comparative and Explainable Machine Learning Framework for UPI Payment Fraud Detection

Aditya Natekar[1], Mohit Nitin Pawar[2], Yash Tandon[3], Ms. Swati Kadu[4]

[1234]*Department of MCA, D. Y. Patil University, School of Humanities and Science, Navi Mumbai, India*

ABSTRACT-*UPI transaction processing happens in real time, leaving a very narrow window for fraud risk assessment without impacting legitimate users. This paper proposes a research framework to evaluate explainable machine learning models for real-time UPI fraud detection. The planned study compares decision trees, random forests, gradient-boosted models, and an anomaly detection baseline using features available at authorization time. Evaluation relies on temporal test splits, low fraud prevalence, probability calibration, and realistic daily analyst alert capacities. For flagged alerts, SHAP feature importance scores are mapped to concise reason codes for human review. As a research protocol rather than an empirical trial, this work does not present trained models or performance metrics due to current constraints on data access and institutional approval. Instead, it defines a structured, reproducible study plan that can be executed once data access and ethics clearances are secured.*

Keywords: Unified Payments Interface, payment fraud detection, explainable AI, SHAP, class imbalance, temporal validation, fraud analytics

1. INTRODUCTION

UPI supports account-to-account payments through a fast, widely used digital payment interface. NPCI reported 24,068.38 million UPI transactions for September 2026. This is a measure of transaction scale, not a measure of fraud incidence or loss [8]. At that scale, even a small false-alert rate can create a substantial review workload, while missed fraud can cause direct financial harm.

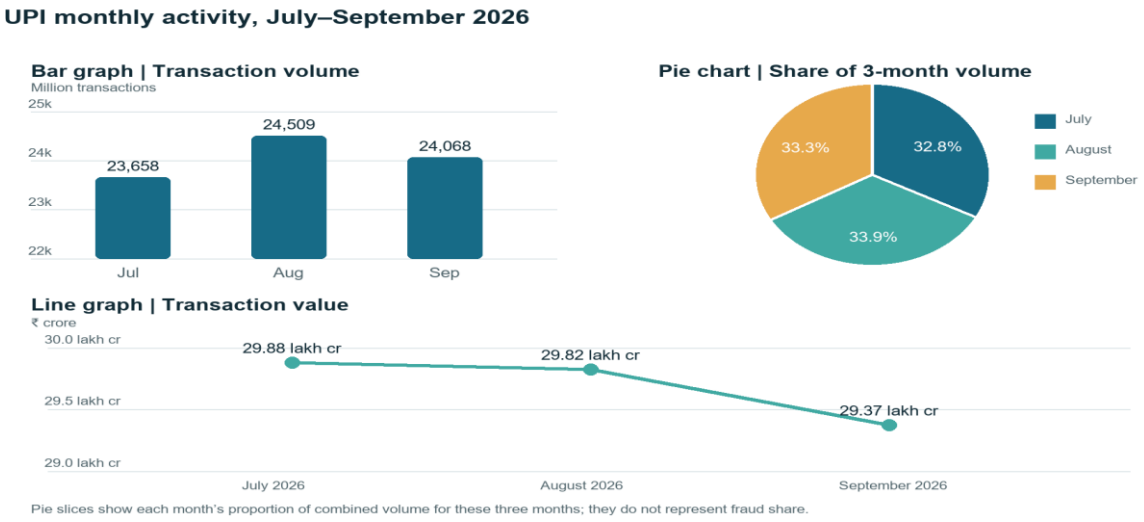


Figure 1. UPI activity shown as monthly volume, period share, and transaction value.

Source: National Payments Corporation of India (NPCI), UPI Product Statistics, monthly statistics. September 2026 is the latest month available on the source page at revision (accessed 22 October 2026). The bar chart compares monthly transaction volume; the pie chart shows each month's share of the combined July–September volume; and the line chart shows transaction value. The pie shares are not fraud proportions. These aggregate activity measures describe transaction-system scale; they do not measure fraud incidence, losses, or model performance.

NPCI reports 24,068.38 million UPI transactions worth ₹2,937,396.67 crore in September 2026. Volume was lower than August but higher than July across this three-month window. This context supports the operational motivation for the proposed study; it is not evidence that any fraud-detection model works.

A useful fraud model must do more than separate two labels in a static file. It should identify suspicious behavior using information available before authorisation, remain useful when patterns change, and provide a reason that a trained reviewer can assess. In practice, fraud labels may arrive after a transaction has been investigated, and the data available to an institution may include fields that should not be used by a model because they arise after the decision.

Published UPI fraud studies show that researchers have tested familiar classifiers and proposed operational systems [1–4]. Existing studies vary significantly in dataset composition, feature sets, validation splits, and evaluation metrics, making direct benchmarks unreliable. To address this gap, this study examines how candidate models perform when evaluated specifically under real-time, capacity-constrained UPI fraud review workflows.

2. REVIEW OF SELECTED UPI FRAUD STUDIES

We analyzed four UPI fraud studies (summarized in Table 1) based on publisher records and full-text papers. The table contrasts their methodology, metrics, and findings. Direct comparisons across these studies are limited because of differences in sampling methods, class imbalance ratios, train-test splits, and target label definitions.

Table 1: Key UPI fraud studies and their alignment with the proposed evaluation framework.

Study	What the source reports	Relevance here	Caution for comparison
Sirisha et al. (2025) [1]	Compares machine-learning classifiers, including decision tree, random forest, gradient boosting, and XGBoost, for UPI fraud detection.	Supports comparing multiple tree-based baselines.	The paper describes a proposed/implemented system, but does not provide a shared, independently reproducible temporal benchmark with the other studies.
Padmanaban and Reshma (2025) [2]	Discusses random forest, K-nearest neighbours, and decision tree classification using transaction-related fields.	Shows a simple model set and transaction-level framing.	The available paper uses a limited feature description; its setup is not equivalent to the richer behavioral feature plan in this protocol.
Amulya et al. (2025) [3]	Uses a Kaggle-sourced dataset, an 80/20 train-test split, and decision tree, random forest, gradient boosting, and AdaBoost; reports high accuracy figures.	Illustrates a common classifier comparison and the use of a public proxy dataset.	A random split and accuracy alone do not establish future-period performance or operational alert quality. The reported figures are the authors' results, not results from this proposed study.
Nazmoddin et al. (2024) [4]	Presents a UPI fraud detection approach using transaction history, user behavior, and anomaly detection.	Motivates including behavioral and anomaly signals as candidates.	The paper's described approach does not provide a directly comparable, time-separated evaluation for this protocol.

The gap we take from this review is methodological rather than a claim that earlier systems cannot detect fraud. The selected papers do not establish which candidate is most useful on later transactions under shared data definitions, realistic alert workloads, and structured explanation workflows. This evaluation will be executed once an appropriate dataset is accessible.

3. RESEARCH AIM AND OBJECTIVES

Aim: To establish and validate an explainable machine learning framework tailored for real-time UPI fraud screening.

3.1 Objectives

- Construct real-time transaction, behavioral, beneficiary-novelty, device, and network features using strictly point-of-authorization data.
- Evaluate baseline interpretable models against ensemble methods using time-aware train-test splits.
- Measure performance under extreme class imbalance using precision-recall curves, probability calibration, and fixed daily alert budgets.
- Generate local SHAP explanations for review and track feature importance stability across temporal windows.
- Detail necessary data governance, privacy safeguards, and operational requirements prior to prospective deployment.

3.2 Research questions

1. RQ1: How do comparative models perform on future out-of-time UPI transactions under a constrained daily alert capacity?
2. RQ2: Do behavioral and beneficiary-novelty features add value over basic transaction context under the same time split?
3. RQ3: Are the resulting risk scores sufficiently calibrated for a tiered review policy?
4. RQ4: Are transaction-level reason codes understandable and reasonably stable across future test windows?

4. PROPOSED STUDY DESIGN

4.1 Data needed

The primary dataset should come from a participating bank, payment service provider, or another institution with authority to share de-identified transaction records for research. Each record would need an authorisation timestamp, a transaction outcome, and a fraud label with a documented definition and confirmation delay. The dataset should cover enough consecutive time to create training, validation, and later test periods, and it should include a record of missingness and label maturity.

A public synthetic or proxy dataset could be used to demonstrate the analysis pipeline, but its results would only describe that proxy. They would not show that a model is effective on production UPI payments. Before access, the researcher would need written institutional permission, ethics review where applicable, a data minimisation plan, and an agreement on retention and reporting.

Table 2. Candidate feature groups for a future dataset; all features must be available at decision time

Feature group	Candidate fields	Purpose and boundary
Transaction context	Amount, channel, merchant category where available, hour and day	Describe the payment; timestamps must be normalised and recorded consistently.
Prior customer behavior	Rolling transaction count/value, deviation from prior amount patterns, previous successful payments	Compare with the account's earlier activity only; do not use future events.
Beneficiary relationship	First-payment indicator, time since first observed beneficiary, prior sender-beneficiary history	Represent novelty from historical transfers; do not expose raw UPI IDs.
Device/session context	Device-change indicator, app/session consistency, failed authentication count if lawfully available	Potentially identify account takeover signals; define availability and missingness.
Network context	Prior beneficiary fan-in or shared-device counts computed from earlier records	Explore links to mule-account activity; avoid joins that use later investigations.

4.2 Leakage and privacy controls

Features must be constructed strict to the point in time when the transaction occurs. To prevent target leakage, post-authorization signals such as chargebacks, subsequent blocks, future activity, or retrospective blacklists must be excluded from real-time features. Any resampling or class weighting strategies must strictly fit on training data. Personally identifiable details must be anonymized or excluded, accompanied by audit logging, retention limits, and clear data dictionaries.

5. PROPOSED MODELLING AND EVALUATION METHODOLOGY

5.1 Candidate Models

Baseline models include shallow decision trees and logistic regression for high interpretability, alongside ensemble models like Random Forest and gradient-boosted frameworks (XGBoost, LightGBM, or CatBoost). Isolation Forest can be incorporated to capture unsupervised anomaly scores, acting as an auxiliary signal rather than a direct supervised baseline. The model selection will be pruned based on computational limits and training feasibility.

5.2 Time-aware evaluation

Records would be ordered by authorisation time. Earlier periods would be used for training, the next contiguous period for model selection and threshold setting, and the latest period would remain untouched until the protocol is fixed. If the dataset is long enough, rolling-origin windows would show how performance changes over time. The same split boundaries, feature definitions, and label maturity rules would apply to each candidate model.

Because fraud is expected to be the minority class, overall accuracy would not be the primary selection measure. Candidate measures and their planned interpretation are listed in Table 3. The fixed alert budget should be chosen with the data-owning institution, based on review capacity, rather than selected after viewing final test results.

Table 3. Evaluation measures planned for the empirical phase

Measure	Why it is included	How it would be reported
Precision-recall curve and PR-AUC	Summarises ranking quality when positive cases are rare.	Per held-out period, with fraud prevalence and sample counts.
Recall at fixed alert budget	Reflects how much confirmed fraud appears within a manageable review queue.	Fraud captured among the top N alerts per day or a pre-agreed alert rate.
Precision and false-positive workload	Shows the number of legitimate payments or cases sent for review.	Counts and rates per period; distinguish review alerts from automatic blocks.
Calibration and Brier score	Checks whether risk scores correspond to observed frequencies.	Calibration plot and score on untouched periods.
Temporal performance change	Shows whether later windows differ from earlier evaluation periods.	Same metrics per window; explain changes without claiming causation.
Explanation review	Checks whether reason codes are understandable and grounded in available fields.	Predefined reviewer rubric, reviewer count, and agreement; no claim until performed.

5.3 Explanations and decision support

Upon training the supervised models, global SHAP feature importance will identify overall key predictors, while local SHAP values will provide transaction-level explanations for analyst alert validation [5]. Concise reason codes might say that the transaction amount is unusually high relative to the account's prior history or that the beneficiary is newly observed. These are model associations, not causal explanations. A reviewer study would be needed before claiming that the explanations improve decision speed or consistency.

The proposed workflow would use a risk score to prioritise review or step-up verification. Automatic blocking should not be recommended from an unvalidated research model. Any action policy would require institution-defined thresholds, customer recourse, monitoring for unequal error burdens, and a human escalation path for uncertain cases.

6. HYPOTHESES AND REPORTING PLAN

The following are hypotheses for a future empirical phase, not findings: H1, adding prior behavioral and beneficiary-novelty features will improve recall at a fixed alert capacity compared with basic transaction context; H2, later-period performance will differ from a random-split estimate; H3, probability calibration will make risk thresholds more interpretable; and H4, trained reviewers will rate concise reason codes as useful for assessing alerts. Each hypothesis would be accepted, rejected, or left inconclusive only after the corresponding analysis is completed.

A completed report should include dataset provenance and approval, time coverage, label definition and delay, class prevalence by split, excluded fields, feature availability, model versions and settings, threshold-selection procedure, confidence intervals where appropriate, confusion counts at the chosen operating point, and limitations. It should not report only a headline accuracy score.

7. LIMITATIONS AND ETHICAL CONSIDERATIONS

This manuscript does not include the original data, code, a fitted model, evaluation outputs, or a reviewer study. It therefore cannot establish detection performance, cost savings, generalisation to UPI institutions, or the usefulness of SHAP explanations. The feature list is a candidate design; some fields may not be collected, accessible, lawful to use, or available early enough for an authorisation-time decision.

Fraud screening may delay legitimate payments or create unequal friction. Data access and label quality can also shape the result: confirmed fraud labels may miss unreported cases and may arrive later than ordinary transaction outcomes. These risks require institutional review, data minimisation, access control, subgroup error analysis where legally and ethically appropriate, and an appeals or escalation route. A public proxy can test code, but cannot substitute for representative UPI data.

8. CONCLUSION

I have framed this work as a proposed study because no original UPI dataset or experimental result was available for this revision. The protocol focuses on a future-time evaluation, practical alert capacity, calibrated risk, and explanations that a reviewer can inspect. Its contribution at this stage is a clear plan for testing those questions, not a claim that one model is already more accurate or ready for deployment. The next research step is to secure approved data and labels, preregister or otherwise freeze the evaluation design, run the analysis, and report the resulting evidence and limitations.

REFERENCES

[1] P. Sirisha, S. Jaheda, G. A. Tahaseen, K. D. Madhuri, and P. M. Arshad, “Unified Payments Interface Fraud Detection Using Machine Learning,” in Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT ’25), pp. 420–426, 2025. https://doi.org/10.5220/0013930800004919

[2] K. Padmanaban and S. Reshma, “UPI Fraud Detection Using Machine Learning,” International Journal of Research and Analytical Reviews, vol. 12, no. 2, 2025, pp. 822–827. DOI: 10.56975/ijrar.v12i2.313098.

[3] K. Amulya, M. Bhumika, T. S. Murari, S. Komaravalli, and V. V. Gopal, “UPI Fraud Detection Using Machine Learning,” International Journal of Scientific Development and Research, vol. 10, no. 4, Apr. 2025, pp. b418–b424. https://ijsdr.org/papers/IJSDR2504164.pdf

[4] M. D. Nazmoddin, M. Swetha, G. Yashwanthi, and Y. Divyasree, “UPI Fraud Detection Using Machine Learning,” Journal of Computational Analysis and Applications, vol. 33, no. 6, 2024, pp. 1704–1707. https://doi.org/10.48047/jocaaa.2024.33.06.33

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.

[6] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," in 2015 IEEE Symposium Series on Computational Intelligence, pp. 159–166, 2015. https://doi.org/10.1109/SSCI.2015.33

[7] The author's original proposal also listed F. Provost and T. Fawcett, Data Science for Business, 1st ed., O'Reilly Media, 2013. It is retained as general background rather than as evidence about UPI-specific model performance.

[8] National Payments Corporation of India (NPCI), "Unified Payments Interface (UPI) Product Statistics," monthly statistics, September 2026. Accessed 22 October 2026. https://www.npci.org.in/product/upi/product-statistics

[9] National Payments Corporation of India (NPCI), "Unified Payments Interface (UPI) Product Statistics," monthly statistics for 2026-27. https://www.npci.org.in/product/upi/product-statistics. Accessed 22 October 2026.