AI-Driven Real-Time Fraud Detection in High-Volume UPI Transactions: Evaluating Instant Intervention Engines

Dr. Zaker Ul Oman[1], G Swathy[2], and Kanishka Soudai[3]
[1]*Associate Professor, Avinash College of Commerce, Himayathnagar, Hyderabad*
[2,3]*Student, Avinash College of Commerce, Himayathnagar, Hyderabad*

***Abstract*— The rapid growth of Unified Payments Interface (UPI) transactions in India has significantly accelerated digital financial inclusion However, this expansion has also increased vulnerability to sophisticated fraud risk. Traditional rule-based fraud detection systems are increasingly ineffective, as they struggle to handle large transaction volumes, evolving fraud patterns, and often generate high false positives with delayed detection.**

This study proposes an advanced AI-driven intervention engine designed for real-time fraud detection in high-volume UPI environments. The system is designed to address critical challenges such as scalability and low-latency processing enabling efficient handling of billions of transactions. It adopts a multi-algorithm ensemble approach by integrating models such as Random Forest, XGBoost, and Neural Networks to enhance detection accuracy and effectively differentiate between legitimate and fraudulent activities.

A key contribution of this research is the incorporation of Explainable AI (XAI) techniques, including SHAP and LIME, which improve transparency and interpretability of automated decisions. These techniques help build trust among financial institutions and users by providing clear explanations for flagged transactions.Furthermore, the system utilizes behavioral and contextual analysis by examining transaction patterns, geographical locations, and device identifiers. This enables the detection of complex anomalies that traditional systems may overlook.

The proposed framework is aligned with regulatory standards, particularly the guidelines of the Reserve Bank of India (RBI) on digital security and data privacy, ensuring compliance and real-world applicability. Experimental results on large-scale datasets demonstrate that the model significantly reduces fraud-related losses and enhances the overall resilience of digital payment systems.

Overall, this research presents a scalable, intelligent, and secure solution aimed at strengthening trust and integrity within India's UPI ecosystem.

***Index Terms*— AI Fraud Detection, UPI Security, Real-Time Analytics, Explainable AI, Behavioral Analysis, Ensemble Learning, Financial Cybersecurity**

I. INTRODUCTION

The Unified Payments Interface (UPI) has transformed India's financial ecosystem since its launch in 2016 by enabling instant, high-volume, and low-cost money transactions through Virtual Payment Addresses (VPAs) and QR codes. However, this rapid digital adoption has led to significant rise in sophisticated cybercrimes, including phishing, social engineering, identity theft, and malicious payment requests. Recent reports suggest that over 70% of digital fraud cases in India now involve UPI transactions, resulting in substantial financial losses and weakening user trust in digital infrastructures

Traditional fraud detection mechanism primarily relies on static, rule-based systems that are inherently reactive and ineffective against rapidly evolving fraud tactics like deepfake scams or synthetic identity fraud. In a high-volume environment where millions of transactions occur in milliseconds, manual verification becomes impractical. This highlights the need for the development of AI-led intervention engines capable of real-time monitoring and autonomous decision-making.

This research proposes an advanced fraud intelligence framework that leverages Machine Learning (ML) and Explainable AI (XAI) to analyze complex transaction attributes—such as geolocation, device characteristics, and behavioral signals—in real-time. By transitioning from a reactive to a proactive defense posture, the system aims to detect and block fraudulent activities within real time,

ensuring compliance with RBI guidelines for digital security while maintaining a seamless user experience. Ultimately, the study seeks to strengthen the security, reliability, and trustworthiness of the UPI ecosystem, thereby enhancing user confidence in an increasingly digital financial landscape.

II. OBJECTIVES OF THE STUDY

- Review of AI model and scalability challenges
- To evaluate the role of explainable Ai techniques and it's evolutionary developments in fraud detection
- To behavioral and contextual intelligence in UPI fraud detection
- To Synthesize Legal and Regulatory Guidelines for AI fraud detection

III. RESEARCH METHODOLOGY

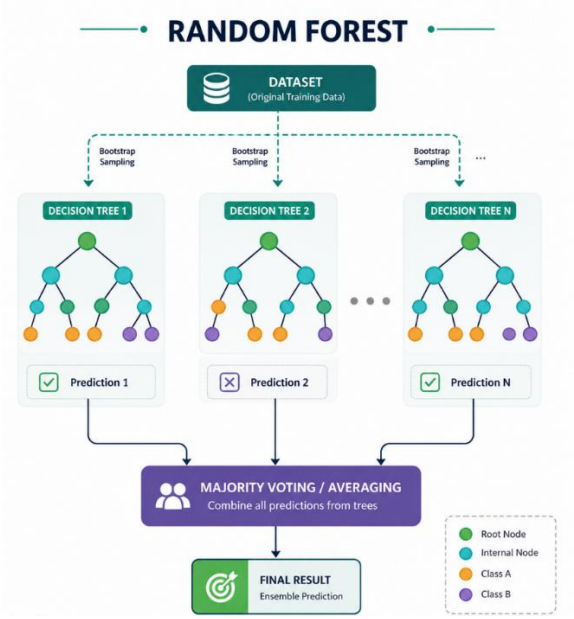
The research methodology employed for this literature survey is a systematic secondary data analysis, focusing on the integration of Artificial Intelligence within the UPI ecosystem. The data collection process involved a structured review of peer-reviewed journals, industry white papers from financial authorities, and technical reports from AI research labs published between 2023 and 2026. This approach ensured a comprehensive understanding of evolving trends in model scalability, the legal implications of the DPDP Act, and the technical shift toward Explainable AI (XAI).

The analysis followed a qualitative synthesis and comparative evaluation of existing literature to identify gaps in traditional rule-based systems. By cross-referencing multi-source data—including latency benchmarks and fraud detection accuracy rates—the study filtered information for technical relevance and regulatory alignment. This descriptive research design allowed for the extraction of key findings regarding hybrid architectures and behavioral biometrics, ensuring that the final conclusions are grounded in the most current and validated industry standards.

III. REVIEW OF AI MODELS IN FRAUD DETECTION

3.1 Random Forest

Random Forest is currently the "gold standard" for tabular data in UPI fraud detection. It is an Ensemble Learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode (for classification) of the individual trees.



1. Core Architecture: “Wisdom of the Crowd”

Random Forest is an ensemble learning technique that combines multiple decision trees to build a robust predictive model. It operates using Bootstrap Aggregating (Bagging), where each tree is trained on different subsets of UPI transaction data generated through sampling with replacement. Additionally, feature randomness ensures that each tree considers only a subset of features (such as transaction amount, time, location, and device ID) at each split, reducing correlation among trees. The final classification is determined through majority voting, where the most frequent outcome (fraud or legitimate) across all trees becomes is selected

2. Advantages in UPI Fraud Detection

Random Forest is highly effective for UPI fraud detection due toits ability to handle imbalanced datasets, where fraudulent transactions are significantly fewer than legitimate ones . The model also captures non-linear patterns, such as sequences of small transactions followed by a large fraudulent one, which simpler models may miss. Furthermore, it offers strong resistance to overfitting, as averaging multiple trees reduces noise and improves generalization to new fraud patterns.

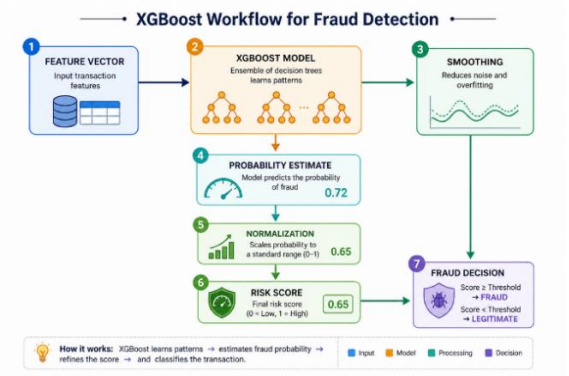
3. Limitations

Despite its effectiveness, Random Forest has some drawbacks. It can be memory-intensive, especially

when using hundreds of trees, making deployment challenging in resource-constrained environments. The model also has a black-box nature, making it difficult to clearly explain decisions to users compared to rule-based systems. Additionally, latency concerns arise in real-time UPI systems, where rapid processing is essential, requiring optimization for faster decision-making.

3.2 XGBOOST :

XGBoost (eXtreme Gradient Boosting) is a high-performance ensemble learning algorithm designed for structured data. Unlike Random Forest, it builds trees sequentially, where each new tree corrects the errors of previous ones. It incorporates advanced features such as L1/L2 regularization, second-order optimization (Hessian-based), missing value handling, and feature subsampling, making it robust and efficient. It is widely recognized for its superior predictive accuracy and scalability in both academic and industry applications.



1. Core Research Mechanics

XGBoost uses Newton Boosting, leveraging both gradients and Hessians for faster and more precise learning. It integrates built-in regularization (L1 & L2) to control model complexity and reduce overfitting. Additionally, its sparsity-aware learning automatically handles missing and sparse data by assigning optimal default directions, enhancing model reliability in real-world datasets.

2. Performance in UPI Fraud Detection

XGBoost is widely regarded as a state-of-the-art model for fraud detection due to its strong performance on imbalanced datasets. It effectively handles class imbalance through parameters such as *scale_pos_weight*, which improves the detection of rare fraudulent transactions. The model typically achieves high accuracy along with enhanced recall and AUC-ROC scores compared to traditional machine learning approaches.

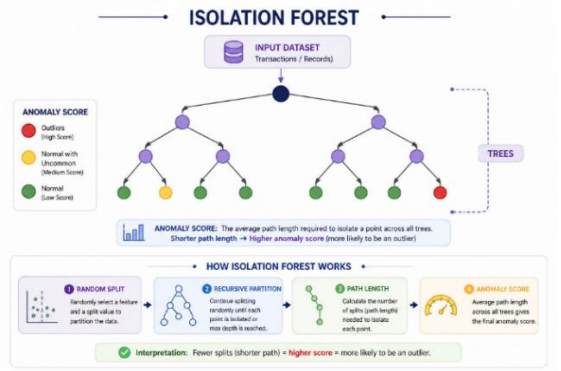
Furthermore, XGBoost's computational efficiency and support for parallel processing enable faster training and prediction, making it suitable for real-time fraud detection. This aligns well with the low-latency requirements of UPI systems, where rapid and accurate decision-making is critical.

Key Hyperparameters

Important tuning parameters include eta (learning rate) for controlled learning, gamma for pruning and split regulation, and max_depth to limit tree complexity and avoid overfitting. These parameters are critical for optimizing model performance in fraud detection scenarios.

3.3 ISOLATION FOREST

Isolation Forest is an efficient anomaly detection algorithm widely used in fraud detection. Instead of learning normal behavior, it identifies anomalies by isolating rare and distinct data points through random partitioning. Since fraudulent transactions are few and different, they are detected quickly with minimal computational effort.



1. Core Concepts

The algorithm is based on three key ideas:

- Isolation: Anomalies are easier to separate because they differ significantly from normal data.
- Partitioning: Data is split using randomly selected features and thresholds to isolate outliers.
- Anomaly Score: Points requiring fewer splits have higher anomaly scores, indicating potential fraud.

2. Working Mechanism

Isolation Forest operates through:

- Random Partitioning: Selecting random features and split values.
- Recursive Splitting: Continuing splits until data points are isolated or depth limits are reached.

- Path Length Calculation: Shorter paths indicate anomalies; longer paths indicate normal behavior.
- Score Aggregation: Final anomaly scores are averaged across multiple trees for accuracy.

3. Application in UPI Fraud Detection

Isolation Forest is highly effective in UPI systems due to:

- High Accuracy (up to 98%) in detecting fraud in imbalanced datasets.
- Unsupervised Learning: No need for labeled fraud data, making it ideal for detecting new (zero-day) scams.
- Scalability: Linear computational complexity enables real-time processing of high transaction volumes.

4. Types of Anomalies in UPI

- Point Anomalies: Single unusual transactions (e.g., sudden high-value transfer at odd hours).
- Contextual Anomalies: Transactions abnormal in context (e.g., impossible location changes).
- Collective Anomalies: Suspicious patterns of multiple transactions (e.g., mule account activity).

IV. ROLE OF EXPLAINABLE AI TECHNIQUES AND ITS EVOLUTIONARY DEVELOPMENTS IN FRAUD DETECTION

In the rapidly shifting landscape of 2026, Explainable AI (XAI) has evolved from a "nice-to-have" transparency feature into the backbone of modern fraud detection systems. As fraud tactics become more automated and sophisticated, the ability to "open the black box" of AI is what allows financial institutions to maintain trust and regulatory compliance.

4.1 The Critical Role of XAI in Fraud Detection

Explainable AI (XAI) plays a vital role in enhancing the transparency and reliability of fraud detection systems, addressing key limitations of traditional black-box models:

- Regulatory Compliance: XAI provides clear, human-readable explanations for automated decisions, ensuring compliance with regulatory frameworks and supporting the “right to explanation” in digital systems.
- Operational Trust: By generating reason codes (e.g., unusual location or abnormal transaction pattern), XAI helps fraud analysts quickly understand why a transaction is flagged, improving efficiency and decision accuracy.
- Bias Mitigation: XAI identifies the key features influencing model decisions, enabling developers to detect and reduce unintended bias, ensuring fair and ethical outcomes.

1.1 Evolutionary Developments: From Rules to Reasoning

The journey of fraud detection has moved through three distinct stages of evolution:

Stage	Technology	Characteristics	XAI Status
Traditional	Rule-Based Systems	"If-Then" logic (e.g., block if > $5,000).	High (Rules are inherently clear).
Modern	Black-Box ML	Deep Learning, Neural Networks. High accuracy but opaque.	Low (Predictions without explanations).
Current (2026)	XAI-Integrated	SHAP, LIME, and "Self-Explaining" models.	Balanced (High accuracy + transparency).

Key Technical Milestones

1. Post-hoc Explanations: Early XAI used tools like SHAP (Shapley Additive Explanations) and LIME to explain a model's decision *after* it was made.
2. Ante-hoc (Glass-Box) Models: A shift toward models that are interpretable by design, such as Explainable Boosting Machines (EBMs), which offer the performance of complex models with the clarity of a simple decision tree.
3. Semantic Intelligence (GenAI Integration): In 2026, Large Language Models (LLMs) are used to translate technical feature weights into natural language reports for customers and regulators.

1.2 Cutting-Edge Trends in 2026

Recent developments have pushed XAI into more specialized territory:

- Counterfactual Explanations: Instead of just saying why a transaction was flagged, the system tells the user what would have needed to be different for it to be approved (e.g., "If this transaction had been verified via MFA, it would have been authorized").
- Explainable Federated Learning: Financial institutions now train models collaboratively without sharing sensitive customer data. XAI is integrated into these decentralized loops to ensure the global model remains transparent across all participating banks.

1.3 Real-time Root Cause Analysis: Modern platforms now use AI agents to perform "incident autopsies," instantly pulling logs and transaction history to explain a sudden surge in fraud alerts.

In the early 2020s, the "Black Box" problem led to significant friction: legitimate transactions were declined without reason (False Positives), and regulators began imposing heavy fines for "uninterpretable" algorithmic bias.

XAI solves these pain points through:

- Model Debugging: Enabling data scientists to see if a model is "overfitting" on noise (e.g., flagging a transaction just because it happened at 4:02 AM).
- Human-in-the-Loop (HITL) Efficiency: Fraud analysts in 2026 use XAI dashboards to visualize feature importance, reducing investigation time by an average of 40%.
- Adversarial Defense: By understanding the "why," security teams can predict how fraudsters might "probe" a model to find its blind spots.

1.4 Evolutionary Developments (Timeline)

Phase I: The Post-Hoc Era (2017–2021)

During this stage, we used external "wrapper" methods to explain complex models like Random Forests or Gradient Boosting Machines (XGBoost).

- SHAP (Shapley Additive Explanations): Based on cooperative game theory, assigning each feature a "payout" (contribution) to the final prediction.
- LIME (Local Interpretable Model-agnostic Explanations): Perturbing input data to see how the prediction changes, creating a simplified local model.

Phase II: The "Glass-Box" Shift (2022–2024)

Research shifted toward Intrinsic Interpretability. Instead of explaining a complex model, we built models that were complex yet inherently readable.

- EBMs (Explainable Boosting Machines): Developed by Microsoft Research, these provide accuracy rivaling XGBoost while remaining fully transparent via contribution plots.
- Neural Additive Models (NAMs): Combining the power of deep learning with the interpretability of linear models.

Phase III: Cognitive & Generative XAI (2025–2026)

The current frontier involves Multimodal Explanations.

- LLM-Augmented Logic: Systems now use generative agents to synthesize SHAP values into natural language reports (e.g., *"This transaction was flagged because the user's velocity increased by 300% while utilizing a known VPN exit node."*).
- Counterfactual Reasoning: Models now provide "What-If" scenarios to help customers rectify blocked accounts instantly.

1.5 Comparative Analysis of XAI Techniques

Technique	Type	Best For	Trade-off
SHAP	Post-hoc	Global/Local importance	Computationally expensive
LIME	Post-hoc	Individual transaction audit	Can be unstable (noisy)
EBM	Intrinsic	Regulatory-heavy environments	Slightly lower max-accuracy
Integrated Gradients	Gradient-based	Deep Learning / Neural Nets	Hard for non-technical users to read

V. COMMON UPI SCAMS

A. The "Collect Request" Trap (Refund/Prize Scam)

- The Scenario: A scammer sends you a "Collect Money" request on an app like PhonePe or Google Pay, claiming it’s for a "GST refund,"

"lottery prize," or "reversing a wrong transaction."

- The Hook: You receive a notification from trusted name such as "Request from Netflix" or "Income Tax Dept."
- The Outcome: If you click "Approve" and enter your UPI PIN, money is debited from your account.
- Crucial Rule: You never need to enter your PIN to *receive* money.

B. The QR Code "Payment" Scam

- The Scenario: commonly seen on platforms like OLX. A buyer (the scammer) claims they want to pay you. They send a QR code via WhatsApp and say, "Scan this to receive the payment."
- The Outcome: Scanning a QR code is an authorization to pay. Once you scan and enter your PIN, the money is gone.

C. The Fake Payment Screenshot/Soundbox Scam

- The Scenario: A customer at a shop "pays" via UPI and shows the merchant a screen that looks identical to a successful transaction page.
- The Hook: They use "spoof apps" that generate fake confirmation screens with the merchant's name and the correct amount. Some even play a recorded "Payment Received" voice alert to appear genuine
- The Outcome: The merchant lets the customer leave with goods, only to realize later that no money hit the bank account.

VI. HOW UPI SCAMS ARE DETECTED BY AI & XAI

Modern banking systems in 2026 use a combination of Behavioral Biometrics and Explainable AI (XAI) to detect and prevent this in real time

A. Behavioral Anomaly Detection

AI models (like XGBoost or Random Forest) monitor thousands of data points per second.

- Velocity Checks: If a user who typically makes 2 transactions per day suddenly approves multiple "Collect Requests" in 2 minutes, the system flags it as suspicious
- Device Fingerprinting: If a UPI ID is suddenly accessed on a new device with a different IP address and immediately tries a high-value transfer, the system triggers a mandatory 24-hour cooling period.

B. XAI (Explainable AI) for Risk Scoring

When a transaction is blocked, XAI (using techniques like SHAP) provides the "Reason Code" to the bank's fraud desk:

- Detection Example: A $10,000 transfer is flagged. The XAI dashboard shows:
  - *Factor 1:* Recipient UPI ID is less than 2 hours old (High Risk).
  - *Factor 2:* Sender's GPS location doesn't match their usual city.
  - *Factor 3:* The request was a "Collect" initiated by the recipient.
- Result: The system automatically downgrades the transaction or asks for a secondary Face-ID check.

C. Graph Networks (Identifying "Money Mules")

Banks use Graph AI to track the flow of funds across multiple accounts. If money from a QR scam is transferred to one account and then quickly split into several smaller transactions across multiple accounts, the system identifies this pattern as a potential "mule network."

By analyzing these connections and transaction paths in real time, the system can flag suspicious networks and freeze all linked accounts simultaneously to prevent further fraud.

VII. HOW TO SPOT FAKE PAYMENTS

1. Fake Payment Screenshots & UTRs: Fraudsters use "Counterfeit Apps" (like fake versions of G Pay or Phone Pe) that generate a realistic "Success" screen instantly.

The UTR/Transaction ID Snare: Scammers often show a 12-digit number (UTR) on the screen. Do not trust it. Fake apps generate random 12-digit reference that follow the correct pattern but aren't registered in the banking system.

- Graphical inconsistencies: Look for blurry logos, incorrect fonts, or misaligned text. Authentic apps have crisp, high-resolution UI elements.
- The Time Gap: Check the timestamp on the screenshot. Often, scammers use old screenshots or the clock on the fake app doesn't match the actual current time.
- How the system identifies the scam by: The only proof of payment is a notification on your device or a credit entry in your bank statement. Never let a customer leave until you see the money in your own transaction history.

2. UPI IDs and Fake Payment apps: Scammers install modified APKs (Android packages) that look identical to G Pay, Paytm, or Phone Pe but have a "Manual Entry" mode where they can type any amount and hit "Pay" to show a fake success animation.

- Associated Name Trick: When you scan a QR code, the name displayed Must be the legal business name. If a name like "Software Update" or "Refund Verify" appears, it's a scam ID.
- The @Handle: Professional and official IDs usually end in @sbi, @icici, @okaxis, or @upi. Be aware of long, strange handles like paytm-refund-department-77@xyz.

3. False Security Alerts & SMS Notifications: This is "Smishing" (SMS Phishing). Scammers send a message that looks like it's from your bank to make you believe money has arrived or your account is at risk.

- Sender ID plishing: They use sender IDs like AD-BNKINF or BZ-PAYTM to mimic bank headers.
- "Balance Added" SMS: You might receive an SMS saying: *"HDFC Bank: Rs 5000.00 credited to A/c XXXXX1234 by UPI Ref 678901..."* * The Catch: Check your actual balance. Often, these SMS messages are sent from a regular mobile number or contain a link like bit.ly/check-balance. Real bank SMS will never contain a bit.ly link.

Soundbox Scam: Scammers sometimes play a recorded sound of a "Payment Received" voice alert from their own pocket or a hidden Bluetooth speaker to trick a busy shopkeeper.

4. How to identify the "Fake" Elements:

Element	Detection Method	Official Verification
Fake Screenshot	Look for font mismatches or weird alignments.	Only trust your own "Transaction History" in your app.
Fake UTR	Scammers show a 12-digit number.	Verify it against your bank statement; fake ones won't exist there.
Fake SMS	Check the Sender Header.	Real banks use headers like HX-HDFCBNK, not +91 mobile numbers.
Fake App	Check if it was downloaded via APK.	Official apps are only on the Google Play Store or Apple App Store.
Fake Sound	Scammer plays a sound from their phone.	Listen for the sound from your own official Soundbox or phone.

VIII. OFFICIAL ONLINE FINANCIAL FRAUD REPORTING PORTALS IN INDIA

If you have seen a fake app/screenshot, use these links immediately:

- National Cyber Crime Reporting Portal:
- *What to do:* Click on "Report Financial Fraud." This is the primary portal for filing complaints about fake apps and digital theft.
- NPCI UPI Dispute Redressal:
- *What to do:* Scroll to the "Complaint" section. You can report specific transaction issues, fake UPI IDs, and technical glitches here.
- RBI Sachet Portal: sachet.rbi.org.in
- *What to do:* Use this to report "unregistered entities" or fake financial apps that are pretending to be banks or official payment providers.
- Emergency Contact Numbers: If money has been stolen, time is critical. Use these numbers within the first "Golden Hour":
  - 1930: The National Cybercrime Helpline (Call this immediately to freeze the scammer's account).
  - 14448: RBI's Financial Education helpline for guidance on fraud.

IX. BEHAVIORAL AND CONTEXTUAL INTELLIGENCE IN UPI FRAUD DETECTION

Behavioural Intelligence: Behavioural intelligence analyzes how users interact with digital systems—such as typing patterns, device usage, and navigation behavior—to detect discrepancies. These unique micro-patterns are difficult to

replicate, making them highly effective for detect fraud and ensuring secure digital transactions.

Effectiveness of Behavioural Intelligence: Behavioural intelligence is ever-evolving and adaptive, unlike static rule-based systems. It enables:

- Real-time fraud detection through identifying subtle behavioural anomalies
- Reduced false positives, ensuring genuine users are not blocked
- Scalability for high-volume transaction environments
- Proactive security, evolving with new fraud techniques

Need of Advanced Behavioural Detection

Modern fraud call for advanced solutions due to:

- AI-powered attacks like phishing and deepfakes, detected through behavioural anomalies
- Fraud ecosystems, where AI identifies coordinated mule networks
- Real-time processing needs, enabling instant risk assessment in UPI systems

Augmenting Customer Experience

Behavioural intelligence balances security with usability by:

- Reducing false alarms through personalized behaviour analysis
- Authorising continuous authentication using behavioural biometrics
- Deploying adaptive authentication, applying security checks only when needed

Key Benefits (2026)

- Higher accuracy in fraud detection with fewer false declines
- Operational efficiency through automation of detection processes
- Improved consumer trust via seamless and proactive protection

Contextual Intelligence Run-through

Contextual intelligence adds situational awareness by analyzing the environment, relationships, and history behind transactions, ensuring decisions are not made in isolation.

Environmental Framework: It evaluates multiple risk factors such as:

- Positional-temporal patterns (e.g., impossible location changes)
- Device fingerprint to detect unfamiliar devices
- Network reputation to flag risky connections like VPNs or public Wi-Fi

Relational Intelligence and Mule Identification: Using graph-based analysis, contextual intelligence:

- Assesses recipient (VPA) reputation and suspicious linkages
- Detects collective anomalies, such as coordinated fraud or mule account networks

X. LEGAL AND REGULATORY GUIDELINES FOR AI FRAUD DETECTION

RBI's Evolutionary Shift: From Rules to "FREE-AI"

The RBI has transitioned from issuing rigid circulars to creating dynamic, technology-neutral frameworks.

A. The "FREE-AI" Framework (2025-2026)

RBI outlines the FREE-AI Sutras (Safety, Transparency, Accountability, Fairness, Inclusivity, Sustainability, and Explainability).

Mandate: Financial institutions should move away from "Black Box" models. Any AI blocking a transaction must have an "Understandability by Design" architecture.

- Impact: This legally mandates banks to use XAI (like SHAP values) to justify fraud alerts to both the regulator and the customer.

B. Beyond SMS OTP: The 2FA Revolution (April 2026)

As of April 1, 2026, the "Authentication Mechanisms for Digital Payment Transactions Directions" came into effect.

- The Update: RBI recognized that SMS OTPs are vulnerable to SIM swapping.
- New Protocol: All digital payments should use Two-Factor Authentication (2FA) where at

least one factor is dynamic (linked to that specific transaction) and non-SMS based—such as biometrics, device-bound tokens, or secure in-app prompts.

NPCI's Evolution: Real-Time Behavioural Defense
While RBI sets the policy, NPCI (as the utility provider for UPI) implements the technical guardrails.

A. Federated AI & Risk Scoring

NPCI has transformed simple transaction limits to Federated Learning models.

- How it is useful: NPCI allows banks to share "Fraud Signals" without sharing "Raw Data."
- MuleHunter.AI: NPCI now provides partner banks with tools like MuleHunter.AI, which is useful AI to identify "Money Mule" accounts (accounts used to rotate stolen money) in milliseconds.

B. Friction as a Feature (2025-2026)

NPCI has also outlined "Intelligent Friction" to slow down scammers:

- Delaying Credits: For high-risk or first-time P2P (Peer-to-Peer) transactions, there is often a mandated "cooling period" or delayed credit to give victims time to report a scam.
- Stoppage of P2P Collect: To stop the "Collect Request" scam, NPCI phased out the P2P collect feature, making most transactions Payer-Initiated only.

Comparison of Regulatory Evolution

Era	Focus	Key Mechanism
Traditional (Pre-2021)	Liability	Post-fraud reporting (Zero Liability rules).
Modern (2022-2024)	Prevention	Device binding, SMS OTPs, transaction limits.
Advanced (2025-2026)	Prediction & XAI	Behavioral Biometrics, FREE-AI compliance, and Federated Risk Scoring.

C. How RBI and NPCI identify Fraud (2026 Mechanism)

The detection process is not just about matching a PIN. It is a multi-layered "Digital Risk Filter" that operates in milliseconds.

NPCI's Real-Time AI Detection

- MuleHunter.AI: NPCI mandates a specialized AI engine called *MuleHunter.AI* to all partner banks. It identifies "Money Mules" by spotting accounts that suddenly receive multiple small payments and immediately transfer them out.
- Behavioral Baselines: It builds a "DNA of Payment" for every user. If you suddenly try to pay ₹50,000 to a new merchant at 3:00 AM from a new city, the AI triggers a 3-to-5 second "Micro-Pause" to warn you.
- Federated Learning: Banks now share "Fraud Signals" (anonymized data about suspicious devices or IPs) through NPCI's central hub without sharing personal customer data, allowing the entire network to learn about a new scam forthwith.

RBI's Digital Payments Intelligence Platform (DPIP)

It mandates as a prototype in late 2025 and fully operational by 2026, the DPIP acts as a central brain for the Indian banking system.

- It consolidates real-time fraud data from the Ministry of Home Affairs (1930), the DoT (Chakshu), and all commercial banks to block fraudulent VPA (Virtual Payment Address) handles across all apps simultaneously.

XI. EVOLUTION OF UPI FRAUD GUIDELINES – TIMELINE

The evolution of these guidelines represents the shifting battleground between scammers and regulators.

Era	Key Guidelines issued by RBI/NPCI	Major Fraud Focus
Foundation (2016-2018)	Master Direction on Digital Security: Mandated 2FA (Two-Factor Authentication) and Device Binding (linking SIM to app).	Simple Phishing
The "Rules" Era (2019-2022)	Customer Protection Circulars: Established "Zero Liability" for users who report fraud within 3 days. Mandatory "Verify Merchant" names.	Collect Request Scams
Technical Hardening (2023-2024)	Inactive ID Deactivation: NPCI mandated deactivating UPI IDs unused for 12 months. Limited "Collect Requests" for P2P.	Dormant Account Misuse
AI & XAI	Authentication Mechanisms	Deepfakes

Era	Key Guidelines issued by RBI/NPCI	Major Fraud Focus
Era (2025-2026)	Framework: Effective April 1, 2026. Mandates dynamic 2FA (Biometrics/App Tokens) and Risk-Based Authentication.	& AI Scams

Consolidation of Evolution (2016–2026)

- Phase 1 (2016-2020): It Focus on Access Control (PINs, Device Binding).
- Phase 2 (2021-2024): It Focus on System Hardening (cooling periods, deactivating dormant IDs).
- Phase 3 (2025-2026): It Focus on Intelligence & Accountability (MuleHunter.AI, XAI-driven transparency, and mandatory compensation).

XII. MAJOR REGULATORY BREAKTHROUGH FOR 2026

A. The "Dynamic 2FA" Mandate (April 2026)

The RBI's latest mandate, Authentication Mechanisms Directions (effective April 2026) marks the end of the "OTP Era."

- The Shift: Banks are now encouraged to move away from SMS OTPs (which are prone to SIM swapping) towards Dynamic Factors like device-bound biometrics or secure in-app prompts.
- Contextual Checks: If a transaction is flagged as "High Risk" by the AI, the guidelines now allow banks to utilize Digi Locker as a secure channel for high-value transaction confirmation.

B. "Explainability" Criterion:

Under the RBI FREE-AI Sutras, banks are legally required to be able to explain *why* an AI flagged or blocked a customer's transaction. This has made Explainable AI (XAI) a core compliance prerequiste rather than just a technical tool.

C. Merchant Accountability

NPCI now stipulate that all new merchants undergo AI-verified KYC before onboarding. Refunds for disputed transactions must now be initiated within T+1 days, drastically improving recovery rates for victims.

XIII. FINDINGS

1. Architectural Scalability & Execution

- Tiered Inference Initiative : To maintain the <100ms latency required for UPI, engines have shifted to a "Hybrid" model: Random Forest/XG Boost for instant screening, and Graph Neural Networks (GNNs) for background relational mapping.
- Sovereign AI Integration: National frameworks (e.g., NPCI's 2026 collaboration with NVIDIA) has shifted to Payments-Native Foundation Models (like Fi MI), to handle population-scale data natively within India.
- Agentic Orchestration: Modern engines use Mixture of Experts (MoE) architectures to mandate high-volume bursts without performance degradation during peak transaction hours.

2. Evolutionary change of Explainable AI (XAI)

- The "Why" Mandate: XAI tools (SHAP/LIME) are now used to generate real-time Reason Codes for every blocked transaction, moving from diagnostic "black boxes" to regulatory-compliant "glass boxes."
- Algorithmic Traceability: Advanced techniques like DeepLIFT allow auditors to trace a block back to specific activated neurons, ensuring non-discriminatory and fair decision-making.
- Trust Drivers: Transparent intervention (explaining *why* a payment was paused) has increased customer retention by replacing generic error messages with contextual security education.

3. Behavioral & Contextual Intelligence

- Behavioral Identity: Security has moved from "what you know" (PINs) to "how you act." This includes Touch Pressure Signatures, Swipe Velocity, and Typing Cadence to detect social engineering or duress.
- Relational Mapping: GNNs identify Mule Account Clusters by treating transactions as network interfaces , detecting circular fund flows that are invisible to single-transaction analysis.
- positional-Temporal Nuance: Engines now flag "impossible travel" or sudden bursts in velocity that deviate from a user's 3-month historical baseline.

4. Regulatory & Legal Synthesis

- Transparency by Design: The 2026 AI Governance Sutras outlines that transparency is built into the model architecture, not added as a post-hoc layer.

- Privacy-Preserving Learning: Under the DPDP Act, the industry is integrating Federated Learning, allowing banks to learn from global fraud patterns without sharing sensitive Customer PII.
- Human-in-the-Loop (HITL): Regulatory mandates now strictly require a clear escalation path for high-value automated blocks to ensure consumer rights and grievance redressal.

XIV. CONCLUSION

The transition to AI-powered intervention engines in the 2026 UPI ecosystem explains a definitive shift from reactive monitoring to proactive, real-time orchestration. Research concludes that successful fraud detection now depends on a hybrid architectural approach that balances sub-100ms latency with deep relational intelligence. By adopting Graph Neural Networks and behavioral biometrics—such as keystroke dynamics and device handling—these systems have moved beyond simple metadata to detect identity through "how" a user interacts rather than just "what" credentials they provide. This shift has successfully pushed true positive detection rates as high as 97% while significantly reducing the friction of false positives for legitimate users.

Furthermore, the adoption of Explainable AI (XAI) and privacy-preserving technologies has changed regulatory compliance from a constraint into a core performance metric. Under the 2026 AI Governance Sutras and the DPDP Act, the industry has adopted "Transparency by Design," using techniques like SHAP and Federated Learning to ensure every automated block is auditable and ethically sound without compromising sensitive customer data. Ultimately, the future of UPI security lies in this multi-layered framework where high-speed deep learning is synced to human-in-the-loop oversight, ensuring that automated financial interventions remain both instantaneous and transitionally transparent.

REFERENCES

[1] Reserve Bank of India. (2024–2026). *Master direction on digital payment security controls*. https://cms.rbi.org.in

[2] Reserve Bank of India. (2026, February). *Draft framework on digital fraud compensation*.

[3] Reserve Bank of India. (2026). *Authentication framework for digital payments*.

[4] National Payments Corporation of India. (2025, August). *UPI usage guidelines update*.

[5] National Payments Corporation of India. (2025–2026). *Bharat Connect AI-COU rationalization framework*.

[6] Reserve Bank Innovation Hub. (2026). *MuleHunter.AI: AI-based fraud detection system*.

[7] Digital Payments Intelligence Platform. (2026). *DPIP framework for safer UPI*.

[8] Department of Telecommunications. (2026). *Chakshu facility for fraud reporting*.

[9] AI-powered UPI fraud detection. (2025). *International Journal of Innovative Science and Research Technology (IJISRT)*.

[10] Fraud detection in UPI using AI. (2025). *International Journal of Creative Research Thoughts (IJCRT)*.

[11] UPI fraud detection using machine learning. (2025). *Technical study report*.

[12] AI-based detection and behavioral analytics. (2026). *Research study*.

[13] USENIX. (2025). *Security and privacy advice for UPI users*.

[14] National Institute of Standards and Technology. (2021). *NIST AI 100-1: Explainable artificial intelligence*.

[15] European Union. (2024). *EU Artificial Intelligence Act*.

[16] Google Research. (2016). *"Why should I trust you?": Explaining the predictions of any classifier (LIME)*.

[17] PwC India. (2026). *Analysis of RBI digital payment security controls*.

[18] Cashfree. (2025). *Fake payment screenshot scams: Detection and prevention*.

[19] Gridlines. (2025). *How fake UPI scams work*.

[20] Bajaj Finserv. (2025). *Types of UPI frauds*.