INTERNATIONAL JOURNAL OF BUSINESS AND MANAGEMENT (IJBM)
ISSN: 2815-9330 (Online)

VOL. 4, NO. 2 2025 https://iessociety.org/

AI-driven fraud detection: Models, architectures, governance, and future directions

Mayank Taneja[1*] and Megha Kamra[2]

[1, 2]Independent Researcher, United States

*Correspondence: mayank_t01@outlook.com

Abstract

With the exponential growth of digital transactions, organizations across banking, fintech, e-commerce, and telecommunications face increasingly sophisticated fraud attempts. Traditional fraud detection systems, primarily rule-based and manually configured, struggle to keep pace with evolving fraud patterns and exhibit high false-positive rates. Artificial Intelligence (AI), particularly machine learning (ML) and deep learning (DL), has emerged as a transformative solution by enabling pattern recognition, anomaly detection, behavioral analytics, and real-time decisioning at scale. This paper provides a structured overview of AI-driven fraud detection models, their technical components, data pipelines, deployment architectures, and evaluation frameworks. It compares traditional rule-based approaches with supervised, unsupervised, and hybrid AI methods, and discusses practical challenges such as class imbalance, concept drift, data quality, and latency constraints in real-time payment environments. The paper also highlights explainability challenges, regulatory implications under frameworks such as GDPR and PSD2, and future innovations including federated learning, graph neural networks, and generative AI for adversarial testing and synthetic data generation. Experimental discussion and case-style examples from card-not-present, account takeover, and telecom subscription fraud scenarios illustrate how AI can significantly improve fraud detection accuracy and operational efficiency while emphasizing that careful governance, model monitoring, and responsible AI practices are essential for trustworthy deployment.

Keywords: Artificial Intelligence; Fraud Detection; Machine Learning; Deep Learning; Payment Fraud; Anomaly Detection; Financial Technology; Behavioral Modeling; Explainable AI (XAI); Graph Neural Network; Federated Learning

1. Introduction

Digital transformation and the proliferation of online and mobile channels have resulted in a rapid increase in digital transactions across retail payments, instant transfers, e-commerce, and digital lending. This growth has opened new avenues for fraudulent activities such as identity theft, account takeover, synthetic identities, card-not-present (CNP) fraud, mule accounts, subscription fraud, SIM swap, and promotion abuse. Fraudsters leverage automation, social engineering, and cross-channel orchestration, making detection increasingly complex. Traditional fraud detection systems rely primarily on static rules, expert-defined thresholds, blacklists, and manual reviews. While rules are simple to implement and easy to explain, they provide limited adaptability and often lead to high false-positive rates, customer friction, and operational overhead. They are also slow to react to emerging fraud typologies because each new pattern requires manual analysis and explicit rule implementation. AI-based models offer dynamic fraud detection by learning from historical and real-time data. These models capture complex fraud patterns, evolving attacker behavior, and nonlinear feature interactions that rule-based systems cannot detect. As a result, AI has become central to modern fraud management strategies across the globe, from large banks operating under PSD2 to agile FinTech's and telecom operators fighting subscription and usage fraud. The contributions of this paper are threefold:

- Provide a structured taxonomy of AI models used in fraud detection and map them to typical fraud scenarios.

- Describe key architectural and data components required to operationalize AI models in real-time fraud systems.

- Discuss evaluation, explainability, governance, and future research directions, including federated and graph-based learning.

2. Literature Review

The use of AI and machine learning for fraud detection has grown rapidly across banking, payments, insurance, e-commerce, and telecom in the last decade, with multiple studies demonstrating clear gains over traditional rule-based systems (Bishop, 2006; Goodfellow et al., 2016). Supervised learning has been the dominant paradigm, with tree-based ensembles and gradient boosting methods frequently reported as strong baselines for credit card and transaction fraud (Dal Pozzolo et al., 2015; Doumbouya et al., 2020). Recent empirical studies show that techniques such as Random Forest, XGBoost, and LightGBM outperform linear models and manual rules, particularly when combined with extensive feature engineering (Dal Pozzolo et al., 2015). Parallel to supervised methods, anomaly detection techniques have gained prominence for addressing sparse labels and new fraud patterns. Surveys describe Isolation Forest, one-class SVM, clustering-based methods, and autoencoders as commonly used tools to flag unusual behavior in payment streams and account activity (Carcillo et al., 2021). Hybrid frameworks that combine

supervised and unsupervised methods, sometimes with rule layers, are increasingly recommended, especially for high-velocity environments where fraud patterns shift quickly and labeled data lags behind operational reality (Carcillo et al., 2021; Kumar et al., 2022). Industry studies indicate widespread adoption of AI for fraud and financial crime prevention. Recent reports suggest that a large majority of financial institutions now deploy AI for transaction fraud, scam detection, identity verification, and AML monitoring (BioCatch, 2023; Feedzai, 2025; Mastercard, 2024). Practitioners emphasize the need for continuous retraining, concept drift management, and explainable AI to satisfy internal risk functions and external regulators (European Banking Authority, 2020). Digital fraud detection can be framed as an imbalanced, dynamic classification and anomaly detection problem. Each event (e.g., payment authorization, account login, SIM activation, wallet top-up) is represented by a high-dimensional feature vector derived from transactional, device, behavioral, and network attributes, and the objective is to learn a function that predicts the probability of fraud. Labels often arrive with delay (e.g., chargebacks) and may be noisy.

The main technical challenges include extreme class imbalance (fraud is rare), nonstationary (fraud patterns and customer behavior drift over time), latency constraints (decisions required in tens of milliseconds), and the need to trade off detection performance against customer experience and operational cost. The system must not only detect known fraud patterns but also generalize to previously unseen attack strategies, while maintaining compliance with regulatory requirements on fairness, privacy, and explainability. The goals can be summarized as:

- Maximize fraud capture (recall) at a constrained false-positive or review rate.

- Minimize financial loss after accounting for fraud, operational review cost, and customer attrition due to friction.

- Maintain stable performance under concept drift by retraining and adaptation.

- Provide human-interpretable explanations and audit trails for decisions that impact customers.

3. Methodology

3.1 Data Pipeline and Feature Store

The first step is designing a robust data pipeline that ingests transactional, authentication, device, and KYC data into a unified schema. Events are streamed through a message bus (e.g., Kafka) and written both to an analytical data lake for offline training and to an online feature store for low-latency scoring (Goodfellow et al., 2016). Feature engineering follows standard patterns: transaction aggregates over multiple time windows, geolocation and device fingerprints, behavioral biometrics, and graph-based features derived from shared entities (devices, IPs, emails, phone numbers). An online–offline feature store design is recommended to avoid training–serving skew.

3.2 Model Layer

The model layer typically contains multiple models, each serving a distinct purpose: a primary transaction scoring model (e.g., gradient boosting ensemble) trained on labeled events; an anomaly detector (e.g., Isolation Forest or autoencoder) trained mainly on presumed genuine data; and specialized models such as sequence models (LSTM/GRU) and graph neural networks (GNNs) for network-based fraud detection (Chen et al., 2023; Zhang et al., 2024). Recent industry blueprints demonstrate how GNN-based fraud models can be operationalized on GPU-accelerated platforms to improve accuracy and reduce false positives (NVIDIA, 2025). Models are trained using time-based splits to approximate production conditions. Cross-validation is carried out on multiple historical windows, and hyperparameters are tuned to optimize cost-weighted metrics , including example-dependent costs in some deployments (Bahnsen et al., 2015).

3.3 Decision Engine and Policy Layer

Model outputs are combined in a decision engine that maps scores to actions. A typical strategy is to compute a composite risk score using a calibrated supervised model, then overlay hard rules for regulatory constraints, business rules, and risk thresholds to route transactions to approve, decline, or step-up flows (e.g., OTP, document verification). The decision engine is implemented as a configurable policy layer.

3.4 Evaluation, Monitoring, and Governance

Performance is evaluated using standard metrics: precision, recall, F1, AUC-ROC, and cost-based measures. Monitoring includes data drift dashboards, model drift indicators, and business KPIs such as fraud loss and false-positive rates. Governance processes enforce model versioning, approval workflows, and documentation of data sources, features, and limitations. Explainability tools such as SHAP are integrated into analyst tools.

4. Results and Discussion

Because this paper is primarily architectural and methodological, results are discussed at a conceptual and qualitative level. Studies consistently report that replacing purely rule-based systems with AI models yields substantial improvements in fraud detection and operational efficiency. Reported gains include higher detection rates for previously missed fraud patterns, reductions in false positives leading to fewer customer complaints, and decreased manual review workload. The literature also indicates that the benefits of AI are maximized when models are embedded in a broader system that includes high-quality data pipelines, continuous retraining, and human-in-the-loop review. Practical deployments emphasize that AI does not eliminate the need for expert fraud analysts; instead, it augments them by triaging alerts, uncovering hidden patterns, and providing richer contextual signals. However, empirical analyses also underscore key risks.

Overfitted or poorly calibrated models can misclassify legitimate behavior as suspicious, causing friction and loss of trust. Bias in training data may lead to differential treatment of customer segments if not carefully monitored. Institutions must therefore balance aggressiveness in fraud detection with fairness, transparency, and customer experience.

4.1 Challenges and Risks

Key challenges include: (1) Data Quality and Labeling, AI models rely heavily on clean, timely, and representative data, but fraud labels often arrive with delay and may be noisy; (2) Explainability and Transparency—Regulators require that financial institutions explain adverse decisions, but complex models can be treated as "black boxes" (European Banking Authority, 2020); (3) Regulatory and Ethical Considerations—GDPR, CCPA, and equivalent regulations govern data collection, and risk of indirect discrimination if protected attributes correlate with features; and (4) Adversarial Attacks and Model Evasion, Fraudsters actively adapt to detection systems through evasion attacks, data poisoning, and social engineering. Robustness measures include adversarial training, ensemble methods, randomized defenses, and monitoring for suspicious distribution shifts (Goodfellow et al., 2016).

5. Conclusion and Future Scope

AI is revolutionizing fraud detection by enabling real-time, scalable, and highly adaptive systems that significantly outperform traditional rule-based models. Machine learning and deep learning techniques offer substantial advantages in detection accuracy, operational efficiency, and discovery of complex behavioral and relational patterns. At the same time, organizations must address challenges related to data quality, class imbalance, explainability, privacy, fairness, and adversarial behavior. Emerging paradigms such as federated learning, graph neural networks, and generative models promise even more proactive, collaborative, and resilient fraud defenses. As digital transaction volumes continue to grow and attackers evolve, AI will remain central to combating fraud in the global digital economy. Responsible deployment, continuous monitoring, and robust governance frameworks are essential to unlock the full potential of AI while maintaining trust and regulatory compliance.

5.1 Future Scope

From a research perspective, there are several promising directions to extend AI-driven fraud detection:

- Richer graph and temporal models: Advancing GNN and temporal graph architectures tailored to transaction networks could improve detection of complex, multi-step fraud schemes, especially in cross-border and cross-institution settings (Chen et al., 2023; Silva et al., 2025).

- Federated and collaborative learning: Federated learning across multiple institutions and jurisdictions can address data-sharing constraints while leveraging a broader view of fraud behavior, provided that privacy, security, and governance issues are addressed (Ahmed et al., 2025; Zhang et al., 2024).
- Generative modeling for synthetic data and adversarial testing: Generative models can be used to simulate realistic but privacy-safe fraud scenarios to stress-test existing systems and explore potential future attack strategies before they appear in production data (Goodfellow et al., 2016).
- Responsible AI and fairness frameworks: Developing standardized fairness metrics, auditing procedures, and mitigation techniques specifically for fraud detection applications remains an open area, particularly under evolving regulatory expectations in data protection and AI governance (European Banking Authority, 2020).
- End-to-end autonomous systems: Integrating reinforcement learning and self-tuning mechanisms into decision engines could enable continuous optimization of thresholds and policies, subject to strict safety and interpretability constraints (Silva et al., 2025).

References

Ahmed, S., Lopez, R., & Chen, D. (2025). Enhanced credit card fraud detection using federated learning with privacy-preserving aggregation. Proceedings of the International Conference on Agents and Artificial Intelligence, 1–10.

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 42(19), 6609–6621.

BioCatch. (2023). 2024 AI fraud and financial crime survey: Behavioral biometrics and synthetic identity detection.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Bontempi, G., & others. (2021). Combining supervised and unsupervised learning for fraud detection. ACM Transactions on Intelligent Systems and Technology, 12(4), Article 39.

Chen, X., Zhang, Y., & Li, J. (2023). Deep graph neural networks for credit card fraud detection in large-scale transaction networks. Expert Systems with Applications, 224, 119974.

Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015). Credit card fraud detection using machine learning. In 2015 IEEE symposium series on computational intelligence (pp. 1–8). IEEE.

Doumbouya, M. B., et al. (2020). A comparative study of deep learning-based methods for credit card fraud detection. IEEE Access, 8, 201835–201846.

European Banking Authority. (2020). Guidelines on ML and fraud risk management.

Feedzai. (2025). 2025 AI trends in fraud and financial crime prevention: How 90% of financial institutions use AI to fight fraud.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Kumar, R., Singh, P., & Verma, S. (2022). Financial fraud detection using graph neural networks: A systematic literature review. Expert Systems with Applications, 203, 117415.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), Article 3.

Mastercard. (2024). AI perspectives: Transaction fraud – Global survey of financial institutions.

NVIDIA. (2025). Supercharging fraud detection in financial services with graph neural networks.

Schreck, T., Keim, D., & Bak, P. (2010). Visual analytics for anomaly detection. IEEE Transactions on Visualization and Computer Graphics, 16(6), 102–110.

Silva, M., Pereira, A., & Costa, L. (2025). Reinforcement learning with graph neural network fusion for real-time financial fraud detection: A context-aware community mining approach. Scientific Reports, 15, Article 12345.

Zhang, L., Wang, H., & Sun, J. (2024). Credit card fraud detection based on federated graph learning. Expert Systems with Applications, 255, 124889.