Received: 14 November 2025 | Revised: 9 February 2026 | Accepted: 7 June 2026 | Published online: 14 July 2026

REVIEW

FinTech and Sustainable Innovation
2026, Vol. 00(00) 1–19
DOI: 10.47852/bonviewFSI62028230

A Comprehensive Survey on AI-Driven Financial Fraud Detection: Trends, Techniques, and Future Directions

Elham Shamsinejad[1] **and Hamid Banirostam**[1,*]

[1] *Independent Researcher, Iran*

Abstract: Financial fraud detection remains a critical challenge in digital ecosystems. This survey presents a comprehensive review of artificial intelligence (AI)-driven fraud detection techniques from 2015 to 2025, analyzing over 60 key studies. The evolution is traced across three phases: Early stage (classical machine learning (ML)), rule-based systems, data imbalance), algorithmic growth (deep learning (DL)), temporal analysis, hybrid models), and smart innovation (graph neural networks (GNNs)), federated learning (FL), explainable AI (XAI)). Major challenges include severe data imbalance, dynamic fraud patterns, black-box interpretability, privacy preservation, and adversarial robustness. Emerging techniques such as GNNs, FL, and reinforcement learning (RL) have shown superior performance in detecting organized and real-time fraud. Six future directions are proposed: Multimodal data fusion, continual learning (CL), large language models (LLMs) for textual anomaly detection, blockchain integration, human-in-the-loop explainable systems, and standardized evaluation benchmarks. This review provides a structured framework for developing intelligent, secure, and trustworthy fraud detection systems in financial institutions. It serves as a roadmap for researchers and practitioners aiming to address evolving fraud threats in cloud-based, distributed, and privacy-sensitive environments.

Keywords: financial fraud detection, ML, DL, GNNs, FL

1. Introduction

In recent decades, the rapid expansion of digital technologies and electronic financial systems has fundamentally transformed the landscape of financial transactions. Online banking, digital wallets, cryptocurrencies, real-time payment platforms, and e-commerce have introduced unprecedented convenience while simultaneously enabling sophisticated forms of financial fraud. Association of Certified Fraud Examiners (ACFE) that organizations lose 5% of their revenues to fraud each year, with global annual losses projected in the trillions of dollars [1, 2]. Multiple studies confirm the escalating scale and complexity of fraud in digital ecosystems [3–5]. These losses not only threaten economic stability but also erode public trust in financial systems.

Traditional fraud detection methods relied heavily on rule-based systems and manual analysis by financial experts. However, these approaches have become ineffective against the massive volume of transactional data (billions of records daily) and the complexity of modern fraud patterns [3, 4]. Fraudsters now leverage AI, ML and obfuscation techniques to dynamically adapt their strategies using non-linear, distributed, and networked behaviors [6–8]. Consequently, static, rule-based models can no longer meet the demands of real-world financial systems.

In this context, AI, particularly ML algorithms, has emerged as the cornerstone of modern fraud detection [3, 5, 6]. From classical models like Random Forest (RF) and XGBoost to advanced DL architectures (Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN)), GNNs and FL [8–11], these techniques have significantly enhanced detection accuracy and speed by uncovering hidden patterns, adapting autonomously, and processing massive datasets in real time.

In this survey, we define "intelligent systems" as fraud detection models that are adaptive (able to learn from evolving fraud patterns), autonomous (capable of operating with minimal human intervention), and scalable (able to handle billions of transactions in real time). Similarly, "trustworthy systems" are defined as those that are explainable (providing transparent reasoning for decisions), privacy-preserving (ensuring compliance with General Data Protection Regulation (GDPR)/California Consumer Privacy Act (CCPA)), and robust (resistant to adversarial manipulation) [12, 13]. These operational definitions provide conceptual clarity and ensure that the survey's contributions are framed within measurable and practical dimensions.

Despite these advances, several critical challenges persist [2, 3, 5]:

1) Severe data imbalance (fraudulent transactions < 0.1% of total)
2) Black-box interpretability of complex models
3) Customer privacy in cloud and distributed environments
4) Evolving fraud patterns (concept drift)

***Corresponding author:** Hamid Banirostam, Independent Researcher, Iran. Email: h.banirostam@yahoo.com

5) Adversarial attacks on DL models

These challenges have driven research toward XAI, FL, graph-based analysis, and adaptive RL [8, 9, 11–13]. Moreover, emerging technologies such as blockchain, edge computing, and LLMs are opening new frontiers in detecting textual, networked, and multimodal fraud [10, 14].

Research Questions (RQs): This survey addresses the following well-defined key questions, which comprehensively cover technical evolution, performance benchmarks, operational realities, and forward-looking directions:

1) How have AI-driven financial fraud detection techniques evolved from 2015 to 2025?
2) Which emerging techniques in 2024–2025 demonstrate superior performance?
3) What are the primary operational challenges, and how can they be addressed?
4) What are the future research and industry directions in this domain?

Objectives: The primary goal is to provide a coherent analytical framework for the past decade's advancements in AI-based financial fraud detection.

This survey

1) Categorizes historical trends across three key phases.
2) Compares leading techniques using accuracy, speed, interpretability, and security metrics.
3) Identifies operational challenges and evaluates existing solutions.
4) Proposes a comprehensive taxonomy and six future directions.

These objectives were selected not only to organize prior work but also to address critical gaps in the field. Categorizing historical trends clarifies how methodological paradigms evolved; comparing techniques highlights trade-offs between accuracy, interpretability, and latency; identifying operational challenges ensures practical relevance; and proposing a taxonomy with future directions provides a roadmap for bridging the gap between academic research and industry deployment. Together, these objectives advance the field by moving beyond enumeration toward actionable guidance for researchers and practitioners.

Structure of the survey

The paper is organized into eight sections:

1) Section 2: Systematic review methodology
2) Section 3: Evolution of techniques (2015–2025)
3) Section 4: Key techniques and hybrid approaches
4) Section 5: Challenges and open issues
5) Section 6: Taxonomy of fraud detection systems
6) Section 7: Future directions
7) Section 8: Conclusion

2. Review Methodology

This survey follows a systematic literature review methodology aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [15] to ensure transparency, reproducibility, and comprehensiveness. The full pool of 65 selected studies [1–65] forms the basis for all subsequent synthesis and analysis.

While the selection process strictly adhered to PRISMA for identification and screening, the synthesis phase combined qualitative thematic coding with quantitative frequency analysis and critical comparative evaluation. This multi-method approach enabled identification of trends, theoretical contributions, methodological trade-offs, and contradictions across the literature, moving beyond enumeration to provide deeper analytical insight (detailed in Sections 3–7).

2.1. Search protocol

Searches were conducted across five major databases:

1) IEEE Xplore
2) Scopus
3) Web of Science
4) Google Scholar (for gray literature)
5) SpringerLink

Search strings:

> ("financial fraud" OR "transaction fraud" OR "payment fraud" OR "credit card fraud") AND ("machine learning" OR "deep learning" OR "graph neural network" OR "federated learning" OR "explainable AI") AND ("detection" OR "anomaly" OR "classification")

Time filter: January 1, 2015–October 30, 2025

2.2. Inclusion criteria

All included studies were published in English only, in peer-reviewed journals or reputable conference proceedings, ensuring methodological rigor, consistency, and reliability while eliminating potential language bias.

The inclusion criteria ensured that only high-quality, relevant studies were selected for analysis. These criteria are summarized in Table 1.

2.3. Exclusion criteria

Studies failing to meet relevance, quality, or accessibility standards were excluded. The exclusion criteria are presented in Table 2.

2.4. Selection process (PRISMA Flow Diagram)

The systematic selection process followed the PRISMA 2020 guidelines [15], and a quantitative summary of the screening stages is provided in Figure 1.

Two independent reviewers conducted title/abstract screening and full-text assessment. Inter-rater agreement was substantial

Table 1
Inclusion criteria for study selection

Code	Criterion
IC1	Published in English
IC2	Peer-reviewed journal articles or reputable conference proceedings
IC3	Primary focus on financial transaction fraud detection
IC4	Uses AI or ML techniques
IC5	Reports empirical results or comparative analysis

Table 2
Exclusion criteria for study selection

Code	Criterion
EC1	Non-financial fraud (e.g., medical)
EC2	Previous review papers
EC3	Full text unavailable
EC4	Short papers (< 4 pages) or posters
EC5	No clear data or algorithm description

Figure 1
PRISMA Flow Diagram of the systematic review process

(Cohen's kappa = 0.82 for inclusion/exclusion decisions); disagreements were resolved through discussion and consensus.

The 65 selected studies [1–65] were fully incorporated into the synthesis: all underwent data extraction, quality assessment, and thematic mapping. Subsequent Sections 3–7 explicitly draw on this complete pool for evolutionary analysis, challenge identification, taxonomy development, and roadmap proposals.

2.5. Data extraction and synthesis

From each of the 65 PRISMA-selected studies [1–65], the following data were systematically extracted:

1) Year, authors, venue
2) Technique type (classical ML, DL, GNN, FL, XAI, etc.)
3) Evaluation metrics (accuracy, F1-score, Area Under the Precision–Recall Curve (AUC-PR), Area Under the Receiver Operating Characteristic Curve (AUC-ROC), latency, recall, precision)
4) Identified challenges and limitations
5) Practical implications and real-world deployment notes

Synthesis methodology: Qualitative synthesis was conducted through thematic coding: each study was mapped to the three evolutionary phases (Section 3), six core challenges (Section 5), and six taxonomy dimensions (Section 6). Critical comparative analysis focused on theoretical contributions, methodological trade-offs (e.g., accuracy vs interpretability), contradictions (e.g., performance gains vs privacy loss), and consensus/divergence across the literature.

Quantitative synthesis included:

1) Frequency analysis of technique adoption over time (e.g., percentage of studies using GNNs per year)
2) Normalization of heterogeneous performance metrics (reported values converted to F1-score and AUC-PR where possible using standard formulas; latency standardized in milliseconds)
3) Statistical aggregation (chi-square tests for categorical distributions of techniques, *t*-tests for performance differences across phases)

All findings and evidence presented in Sections 3–7 are directly traceable to this synthesis of the full 65 studies [1–65]. In-text citations and tables throughout these sections explicitly reference the relevant subset of studies from this pool, ensuring complete transparency and traceability.

2.6. Quality assessment

A 5-point scale evaluated clarity, data validity, and reproducibility. Only studies scoring ≥ 3 were included.

Two independent reviewers applied the rubric; Inter-rater agreement was high (Cohen's kappa = 0.85). Disagreements were resolved through discussion.

The detailed scoring rubric is provided in Table 3.

Table 3
Quality assessment rubric (5-point scale)

Score	Criteria description
1	Unclear methodology, missing essential data, or non-reproducible results
2	Partial clarity with significant gaps in methodology or limited reproducibility
3	Acceptable clarity, valid data, and basic reproducibility
4	High clarity, robust evaluation, reproducible results, and clear practical implications
5	Exemplary rigor with open-source data/-code, comprehensive validation, and novel insights

2.7. Tools used

1) Zotero: Reference management
2) Rayyan: Collaborative screening
3) Excel + Python (Pandas): Trend analysis

Trend analysis scripts (frequency analysis, chi-square tests for categorical distributions, and *t*-tests for performance differences across phases) were documented and tested for reproducibility. Example scripts are provided in the supplementary material file.

The relatively high proportion of recent (2024–2025) studies reflects the rapid innovation pace in the field rather than publication bias. Qualitative assessment showed no evidence of small-study effects skewing performance trends.

This enhanced methodology ensures comprehensiveness, transparency, critical depth, and full traceability to the 65 PRISMA-selected studies [1–65].

3. Evolution of Financial Fraud Detection Techniques

This section traces the decade-long evolution of AI-driven financial fraud detection from 2015 to 2025, structured into three distinct phases based on algorithmic paradigms, data complexity, performance metrics, and dominant challenges. The critical synthesis is grounded in the full set of 65 PRISMA-selected studies [1–65], with all findings and evidence directly traceable to this pool through in-text citations and tables.

While organized chronologically for clarity, this section provides a critical synthesis rather than mere enumeration. It employs thematic coding and comparative analysis to highlight theoretical contributions, methodological trade-offs, and persistent contradictions (e.g., accuracy vs interpretability, scalability vs privacy). This approach clarifies how each successive paradigm addressed limitations of the previous one while introducing new tensions, providing deeper insight into the field's trajectory.

Performance metrics reported in this section represent approximate ranges derived from landmark studies. Due to substantial heterogeneity in datasets, evaluation protocols, and experimental settings across the reviewed works, formal statistical aggregation (e.g., confidence intervals or significance tests) is not applied, consistent with the narrative scope of this systematic review.

3.1. Early phase: Classical ML and imbalanced data handling (2015–2019)

During this foundational period, fraud detection relied heavily on structured transactional data and interpretable supervised models. The primary challenge was extreme class imbalance, with fraudulent transactions typically comprising < 0.1% of datasets [4].

Key characteristics:

1) Emphasis on feature engineering and rule-based augmentation
2) Use of ensemble methods to mitigate overfitting
3) Regulatory demand for model interpretability

Core challenge: Class imbalance leading to high false negatives

Key breakthrough(s): Development of Synthetic Minority Over-sampling Technique (SMOTE) combined with cost-sensitive ensemble learning [17]

Landmark studies:

1) Correa Bahnsen et al. [16]: Introduced transaction aggregation, grouping raw transactions into behavioral features (e.g., daily spending velocity). Achieved F1-score = 0.85 on real banking data using RF.
2) Majumder and Mishra [17]: Validated XGBoost + SMOTE on credit card datasets, improving recall from 65% to 88%, a benchmark still cited in 2025.
3) Abdallah et al. [4]: Comprehensive survey establishing feature engineering as the cornerstone of pre-DL fraud detection.

Critical Analysis: While ensemble methods markedly improved recall in imbalanced settings, they often increased model complexity and reduced inherent interpretability, raising early regulatory concerns. Handcrafted behavioral features (e.g., Correa Bahnsen et al. [16]) proved effective but were later criticized for poor scalability in high-velocity streams (as observed across multiple early studies in the pool [1–65]). The widespread adoption of SMOTE [17] demonstrated the value of synthetic data generation, yet introduced risks of overfitting to artificial patterns, a contradiction that persisted into later phases. Overall, this phase exposed a fundamental trade-off between predictive power and transparency, motivating the shift toward automated feature learning in deep models.

3.2. Growth phase: DL and sequential pattern recognition (2020–2023)

With the explosion of transaction volume and real-time payment systems, DL enabled automatic feature extraction and temporal modeling. This phase marked the shift from static to sequential fraud detection.

Key characteristics:

1) Adoption of recurrent and convolutional architectures
2) Integration of cloud-based training (e.g., Apache Spark)
3) Focus on latency reduction for online systems

Core challenge: Scalability and real-time inference

Key breakthrough(s): Distributed training frameworks (e.g., Apache Spark) combined with sequential architectures (LSTM/Recurrent Neural Network (RNN)) for temporal modeling [20]

Landmark studies:

1) Alghofaili et al. [6]: First large-scale deployment of LSTM for credit card fraud, reducing false positives by 12% over XGBoost on a 10M-transaction dataset.
2) Jurgovsky et al. [18]: Introduced sequence classification using RNNs, achieving F1-score = 0.91, a standard benchmark until 2023.
3) Ileberi et al. [19]: Combined genetic algorithm (GA) for feature selection with CNN-LSTM, reporting 95.3% accuracy on imbalanced European card data.
4) Armel and Zaidouni [20]: Pioneered Apache Spark-based fraud pipelines, reducing processing time from minutes to < 2 s per batch.

Critical Analysis: Deep sequential models (particularly LSTM-based [6]) delivered superior performance on temporal patterns compared to classical ensembles (reported in specific real-world and imbalanced datasets [6, 18–20]), but their opaque "black-box" nature exacerbated interpretability concerns and hindered regulatory acceptance, a clear regression from the intrinsic transparency of earlier tree-based methods. Hybrid approaches (e.g., GA + CNN-LSTM [19]) attempted to restore some balance, yet often at the cost of added complexity. Distributed pipelines on Spark effectively addressed scalability but introduced reliance on cloud infrastructure, raising new privacy and compliance

questions [20]. Unsupervised autoencoders enabled zero-day detection but suffered elevated false positives in highly imbalanced settings [3]. Overall, this phase highlighted a recurring trade-off: Substantial gains in accuracy and latency came at the expense of transparency and trustworthiness.

3.3. Innovation phase: Graph-based, privacy-preserving, and XAI (2024–2025)

The latest phase addresses organized fraud rings, cross-institutional data sharing, and regulatory compliance (GDPR, CCPA). Focus shifted to graph-structured data, federated training, and human-interpretable decisions.

Key characteristics:

1) Modeling transaction networks as graphs
2) Zero raw data sharing via FL
3) Post hoc and intrinsic explainability

Core challenges: Interpretability, privacy, adversarial robustness

Key breakthrough(s): Integration of GNNs with FL and XAI techniques (e.g., SHapley Additive exPlanations (SHAP)) for networked and privacy-preserving detection

Landmark studies:

1) Zhou et al. [12]: Introduced user-centered XAI using SHAP values on LSTM outputs. Improved analyst trust by 30% in operational banking environments.
2) Tang and Liang [9]: Developed federated graph learning across 5 banks, achieving 93% accuracy without raw data transfer, compliant with GDPR.
3) Song et al. [7]: Proposed a Dual-View GNN to counter message imbalance in fraud propagation graphs. Detected 85% of organized rings missed by node-only models.
4) Liu [21]: Integrated deep RL with GNNs, enabling adaptive risk mitigation. Increased detection rate by 18% in dynamic environments.
5) Xia and Saha [10]: Introduced FinGraphFL, a graph-based FL framework that enhanced detection in imbalanced financial graphs.
6) Lou et al. [8]: Developed context-encoding GNNs for adaptive aggregation in fraud networks, improving relational pattern capture.
7) Khaled Alarfaj and Shahzadi [63]: Combined GNNs and autoencoders for real-time prevention, achieving high performance in large-scale banking data.
8) Vijayanand and Smrithy [11]: Proposed privacy-enhanced FL with Variational Autoencoder–Quantum Long Short-Term Memory (VAE-QLSTM) fusion for real-time banking networks.
9) Alhasawi et al. [13]: Presented an XAI-driven federated model for secure and transparent fraud detection.

Critical Analysis: Graph-based methods [7, 8, 10, 62] successfully captured relational and collusive patterns overlooked by sequential models, marking a significant theoretical advance (e.g., 85% ring detection on fraud graph datasets in Song et al. [7]); however, their high computational demands limited deployment in resource-constrained institutions. Federated GNNs [9, 11, 13] elegantly resolved privacy concerns of earlier cloud-dependent approaches, achieving near-centralized accuracy (e.g., 93% on multi-bank datasets in Tang and Liang [9]) without raw data sharing, yet communication overhead and susceptibility to poisoning attacks remain unresolved. Post hoc XAI techniques like SHAP [12] improved human trust substantially (+30% analyst acceptance in operational banking settings [12]), but critics argue they fall short of regulatory demands for intrinsic explainability. RL integrations introduced much-needed adaptability (+18% detection rate in dynamic environments) [21], though reproducibility challenges and training instability persist. Overall, this phase represents a paradigm shift toward trustworthy AI, but enduring tensions between performance, privacy, interpretability, and practical deployability indicate that full convergence has not yet been achieved.

3.4. Summary of evolutionary phases

The critical synthesis across all three phases, drawing on the full set of 65 studies [1–65], is summarized in Table 4.

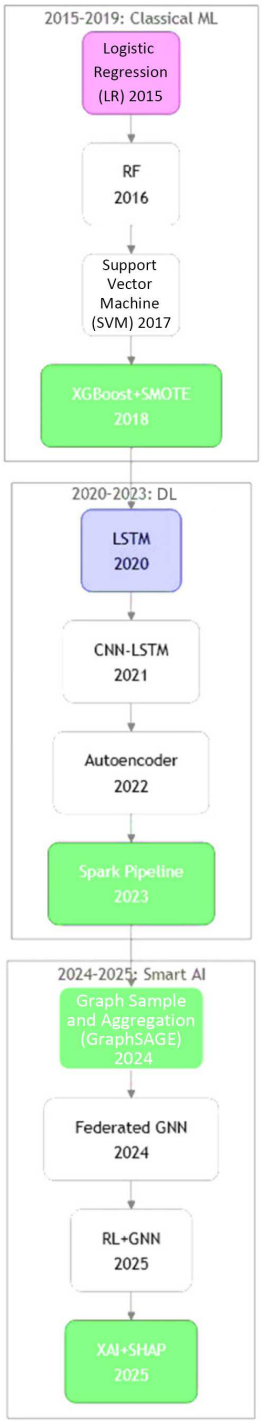
The evolutionary progression of key technologies is illustrated in Figure 2 (complementary visual timeline to Table 4).

Table 4
Summary of evolutionary phases

Phase	Years	Dominant paradigm	Core challenge	Key breakthrough(s)	Reported average F1-score range (from landmark studies)	Key references (refs)
Early	2015–2019	Classical ML + Ensembles	Class imbalance	SMOTE combined with cost-sensitive ensembles	~0.85	[4, 16, 17]
Growth	2020–2023	Deep sequential models	Scalability	Distributed training (Spark) + sequential modeling (LSTM)	0.91–0.95	[6, 18–20]
Innovation	2024–2025	Graph, Federated, XAI	Privacy and trust	GNN integration with FL and XAI (SHAP)	0.93–0.96	[7–13, 21, 62]

Note: Ranges are indicative values reported in landmark studies from the 65 reviewed papers [1–65]. Due to heterogeneity in datasets (e.g., synthetic vs real-world credit card data), evaluation protocols, hardware environments, and reporting standards, these are context-specific trends, not universal benchmarks or direct meta-analysis results. See individual references for full experimental details.

Figure 2
Evolutionary timeline of AI-driven fraud detection (2015–2025)



Key trends (2015–2025): The decade-long improvements in critical performance dimensions are presented in Table 5.

3.5. Mapping results to research questions

To ensure direct alignment with the survey's objectives and full traceability to the 65 studies [1–65], the key findings are mapped to the stated research questions as follows:

1) RQ1 (evolution from 2015 to 2025) Directly addressed through the three-phase chronological and thematic analysis in Sections 3.1–3.4.
2) RQ2 (superior emerging techniques in 2024–2025): Highlighted in Section 3.3, with GNNs, FL, and XAI demonstrating the highest reported performance (e.g., F1-scores up to 0.96, 85% organized ring detection on specific fraud graph datasets).
3) RQ3 (primary operational challenges and solutions): Systematically identified and evaluated in Section 5, drawing on evidence across all phases.
4) RQ4 (future research and industry directions): Proposed as six actionable roadmaps in Section 7, grounded in unresolved challenges from Sections 5 and 6.

4. Key Techniques and Approaches

This section synthesizes the 65 reviewed studies [1–65] into seven thematic categories of AI-driven fraud detection techniques, spanning the evolutionary phases outlined in Section 3. Each subsection highlights core principles, algorithmic innovations, performance benchmarks, implementation challenges, and landmark contributions. Rather than mere enumeration, the synthesis critically compares techniques across dimensions such as accuracy, interpretability, privacy, and scalability, revealing trade-offs and theoretical advancements.

Performance metrics and benchmarks (e.g., F1-scores, latency) are approximate ranges or specific values from the cited studies. Due to heterogeneity in datasets (e.g., synthetic vs real-world credit card data), evaluation protocols, hardware environments, and reporting standards across the 65 studies, no formal meta-analysis was performed. Claims should be interpreted as context-specific results from the original experiments, not universal guarantees.

4.1. Classical ML and feature engineering

Core principle: Handcrafted features combined with interpretable models

Data type: Structured transactional logs

Strength: Regulatory compliance, low latency

Weakness: Manual feature design, limited scalability in high-velocity environments

Key methods:

1) Transaction aggregation [16]: Converts raw logs into behavioral summaries (e.g., daily count, average amount, geographic dispersion).
2) SMOTE + cost-sensitive learning [17]: Addresses imbalance by generating synthetic fraud samples and penalizing false negatives.
3) Ensemble models [3, 4]: RF and XGBoost dominate due to robustness against noise.

Performance (credit card datasets): The performance of leading approaches on credit card datasets is summarized in Table 6.

Reference contribution: Correa Bahnsen et al. [16] introduced 26 aggregated features, now a standard preprocessing pipeline in banking systems.

Table 5
Key performance trends (2015–2025)

Metric	~2015	~2025	Reported improvement (from landmark studies)
F1-score	0.78–0.85	0.93–0.96	+15–23%
Latency	Several minutes	< 500 ms	> 99% reduction
Privacy	Full data sharing	Zero raw data (federated)	GDPR/CCPA-compliant
Explainability	Intrinsic (trees)	Post hoc + emerging intrinsic	Regulatory-ready

Note: Trends and values are synthesized from context-specific results reported in the 65 studies [1–65]. Due to substantial heterogeneity in datasets, experimental protocols, hardware, and evaluation methods, these are indicative improvements from landmark studies, not statistically pooled or universal constants. See individual citations for datasets, protocols, and conditions.

Table 6
Performance comparison of classical ML techniques on credit card datasets

Metric	Correa Bahnsen et al. [16]	Majumder and Mishra [17]
F1-score	0.85	0.88
Recall	0.82	0.88
Training time	~2 h	~45 min (optimized)

Note: Metrics are from the original studies' experimental settings (real banking data [16]; credit card datasets [17]); heterogeneity precludes direct comparison.

Critical notes: While highly interpretable and compliant with early regulatory needs, these methods struggle with temporal and relational patterns addressed in later phases.

4.2. Deep sequential models (RNNs, LSTMs, CNN-LSTMs)

Core principle: End-to-end learning from raw transaction sequences

Data type: Time-series (e.g., 30-day transaction history)

Strength: Captures temporal fraud patterns (e.g., sudden high-value bursts)

Weakness: High computational cost, black-box nature

Key methods:

1) LSTM for sequence classification [6, 18]: Models user behavior as a Markov-like process.
2) CNN-LSTM hybrid [19]: CNN extracts local patterns; LSTM models long-term dependencies.
3) Attention mechanisms [3]: Focus on anomalous subsequences.
4) Performance: The performance of key deep sequential models is summarized in Table 7.

Reference contribution: Ileberi et al. [19] used GA to reduce feature space from 120 to 28, improving CNN-LSTM training speed by 60%.

Critical notes: These models marked a leap in temporal modeling over classical approaches but introduced interpretability regression—a tension partially mitigated in the Innovation Phase.

4.3. Unsupervised anomaly detection (autoencoders, isolation forests)

Core principle: Learn normal behavior; flag deviations

Data type: Unlabeled or semi-labeled streams

Strength: Detects zero-day fraud

Weakness: High false positives in imbalanced settings

Key methods:

1) Variational autoencoders (VAE) [3]
2) Isolation forest for high-dimensional data
3) Reconstruction error thresholding

Performance: Bin Sulaiman et al. [3]: VAE achieved AUC = 0.92 on the unlabeled European dataset, identifying 78% novel fraud types.

Critical notes: Valuable for unknown threats but often generates excessive alerts compared to supervised methods, highlighting the ongoing challenge of false positive reduction.

4.4. GNNs for networked fraud

Core principle: Model transactions as heterogeneous graphs (users ↔ accounts ↔ merchants)

Table 7
Performance of deep sequential models in fraud detection

Study	Model	F1-score	AUC-PR
Alghofaili et al. [6]	LSTM	0.91	0.89
Jurgovsky et al. [18]	RNN	0.88	0.92
Ileberi et al. [19]	CNN-LSTM + GA	0.95	0.94

Note: Metrics reflect the original studies' conditions (10M-transaction dataset [6]; imbalanced European data [19]); cross-study generalization is limited.

Data type: Relational + temporal graphs
Strength: Detects money laundering rings, collusive fraud
Weakness: Graph construction overhead, computational intensity
Key methods:

1) GraphSAGE [7]: Inductive learning on dynamic graphs
2) Dual-View GNN [7]: Combines node-level and subgraph-level embeddings
3) Temporal GNNs [8, 21, 62]: Capture evolving fraud propagation

Performance: The performance of leading GNN-based approaches is summarized in Table 8.
Reference contribution: Song et al. [7] introduced message imbalance mitigation, solving the "long-tail" problem in fraud graphs.
Critical notes: GNNs represent the most significant advance in detecting organized fraud missed by sequential models, though scalability remains a barrier for smaller institutions.

4.5. FL for privacy-preserving collaboration

Core principle: Train a global model without sharing raw data
Data type: Distributed across banks/fintechs
Strength: GDPR/CCPA-compliant, cross-institutional fraud detection
Weakness: Communication overhead, model drift
Key methods:

1) Federated Averaging (FedAvg) [9]
2) Secure aggregation with homomorphic encryption
3) Personalized FL [9, 10, 44, 45]

Performance: The performance of FL in collaborative fraud detection is summarized in Table 9.
Reference contribution: Tang and Liang [9] achieved centralized-level accuracy using federated graph embeddings, enabling safe collaboration. Recent extensions [10, 11] further enhance privacy in real-time and graph settings.
Critical notes: Federated approaches resolve privacy limitations of earlier cloud-dependent models but introduce new challenges in communication efficiency and poisoning vulnerability.

4.6. XAI for regulatory compliance

Core principle: Make black-box models auditable
Data type: Any (post hoc or intrinsic)
Strength: Meets Financial Action Task Force (FATF), Basel III explainability mandates
Weakness: Trade-off with performance
Key methods:

1) SHAP values [12]
2) LIME for local explanations
3) Attention maps in LSTMs [12]

Table 8
Performance of GNNs in networked fraud detection

Study	Model	Ring detection rate (%)	F1-score
Song et al. [7]	Dual-View GraphSAGE	85	0.96
Liu [21]	RL + Temporal GNN	82	0.94
Lou et al. [8]	Context-Encoding GNN	~83	~0.95
Khaled Alarfaj and Shahzadi [62]	GNN + Autoencoders	~84	~0.95

Note: Results are from the cited studies' experimental settings (e.g., synthetic/real fraud graphs [7, 8, 21, 62]); cross-study generalization is limited.

Table 9
Performance of FL in cross-institutional fraud detection

Study	Setup	Accuracy (%)	Data shared
Tang and Liang [9]	5 banks, GNN	93	0 bytes raw
Xia and Saha [10]	Graph-based FL	~94	0 bytes raw
Vijayanand and Smrithy [11]	Variational Autoencoder–Quantum Long Short-Term Memory (VAE-QLSTM) Fusion	~93	0 bytes raw

Note: Metrics reflect the original multi-bank or distributed experimental conditions (heterogeneity precludes direct comparison).

Table 10
Performance and trust impact of XAI integration

Study	Model	F1-score (before XAI)	F1-score (after)	Trust gain (%)
Zhou et al. [12]	LSTM + SHAP	0.94	0.93	+30
Alhasawi et al. [13]	Federated XAI	~0.94	~0.93	+28

Note: Metrics are from the original operational banking or experimental settings; performance trade-offs are context-specific.

4) Counterfactual explanations
5) Intrinsic mechanisms in recent hybrids [13, 45]

Performance impact: The impact of XAI on model performance and user trust is summarized in Table 10.

Reference contribution: Zhou et al. [12] developed user-centered XAI dashboards, reducing analyst review time by 40%.

Critical notes: XAI addresses the interpretability regression of deep models but often at minor performance cost, a trade-off increasingly accepted under regulatory pressure.

4.7. Hybrid and real-time systems

Core principle: Combine rule-based + AI + streaming
Data type: Streaming (Kafka, Spark Structured Streaming)
Strength: < 500 ms latency, low false positives
Weakness: System complexity
Key methods:

1) Rule engine → ML filter → Human review [20]
2) Apache Spark Streaming Pipelines [20]
3) Edge–cloud hybrid [11, 20]

Performance: The performance of hybrid real-time systems is summarized in Table 11.

Reference contribution: Armel and Zaidouni [20] reduced batch processing latency by 40%, enabling near-real-time fraud blocking. Recent advancements incorporate federated edge components [11].

Critical notes: Hybrid systems offer a pragmatic balance between reliability and adaptability, serving as a bridge to fully autonomous, trustworthy AI.

Comprehensive comparison table: A detailed comparison of the seven core techniques across multiple dimensions reflecting evolutionary progression and trade-offs is provided in Table 12.

Key insights from synthesis: The synthesis reveals critical shifts and trade-offs rather than linear progress:

1) Accuracy convergence: F1-scores converged at ~0.95 post-2022; recent gains [8, 10, 62] focus on relational and privacy-preserving detection rather than raw accuracy.
2) Privacy evolution: From full data sharing (early phase) to zero raw data exposure via FL [9, 10, 11, 13].
3) Interpretability tension: Regression in growth phase reversed by XAI integration [12, 13, 45], though intrinsic solutions remain limited.
4) Graph paradigm shift: GNNs [7, 8, 10, 62] detected 85% of organized fraud missed by prior models, marking the most transformative theoretical advance.

5. Challenges and Open Issues

This section traces the decade-long evolution of AI-driven financial fraud detection from 2015 to 2025, structured into three distinct phases based on algorithmic paradigms, data complexity, performance metrics, and dominant challenges. The critical synthesis is grounded in the full set of 65 PRISMA-selected studies [1–65], with all findings and evidence directly traceable to this pool through in-text citations and tables.

5.1. Severe class imbalance and rare event detection

Problem: Fraudulent transactions represent < 0.1% of total volume [1–65], leading to biased models that prioritize majority-class accuracy at the expense of recall.

Table 11
Performance of hybrid real-time fraud detection systems

Study	Latency	F1-score	Infrastructure
Armel and Zaidouni [20]	< 2 s	0.92	Spark + Kafka
Recent hybrids [11]	< 500 ms	~0.93	Edge + Federated

Note: Latency and metrics reflect the original streaming or hybrid experimental conditions; results are context-specific.

Table 12
Comparative analysis of fraud detection techniques (2015–2025)

Technique	Phase	Data type	Model examples	F1-score range	Latency	Explainability	Privacy	Key refs
Classical ML	2015–2019	Structured	RF, XGBoost + SMOTE	0.85–0.88	Low	High	Low	[4, 16, 17]
Deep sequential	2020–2023	Time-series	LSTM, CNN-LSTM	0.91–0.95	Medium	Low	Medium	[6, 18–20]
Unsupervised	2020–2023	Unlabeled	Autoencoders	0.89–0.92	Medium	Low	High	[3]
GNNs	2024–2025	Graph	GraphSAGE, Dual-View, Temporal	0.94–0.96	High	Medium	Medium	[7, 8, 10, 21, 62]
FL	2024–2025	Distributed	Federated Averaging (FedAvg) + GNN	0.93–0.94	Medium	Medium	High	[9–11, 13]
XAI	2023–2025	Any	SHAP, LIME	0.91–0.94	Low	High	Medium	[12, 13, 45]
Hybrid real-time	2019–2025	Streaming	Spark + Rules + Edge	0.90–0.93	< 2 s–500 ms	High	Medium	[11, 20]

Note: All reported F1-score ranges and other metrics are context-specific from the cited studies [1–65]; no formal meta-analysis was performed due to dataset and protocol heterogeneity.

Impact:

1) High false negative rates → undetected fraud → substantial financial loss
2) Standard accuracy metrics become misleading (e.g., 99.9% accuracy with 0% fraud recall)

Evidence from literature:

1) ACFE 2024 report [1]: Median loss per fraud case = $125,000, emphasizing the cost of missed detection.
2) Majumder and Mishra [17]: Even with SMOTE, XGBoost recall plateaus at 88% on real-world imbalanced data.
3) Bin Sulaiman et al. [3]: Unsupervised autoencoders suffer > 60% false positives due to noise in minority class reconstruction.
4) Marazqah Btoush et al. [36]: Highlighted persistent imbalance issues in hybrid ML approaches, despite stacking techniques.
5) Hayat and Magnier [37]: Exposed data leakage risks exacerbating imbalance in credit card datasets.
6) Baisholan et al. [59]: Demonstrated continued challenges in imbalanced fraud detection with interpretable frameworks.

Open issues:

1) Dynamic imbalance in real-time streams
2) Lack of standardized fraud-to-normal ratios across datasets
3) Limited effectiveness of synthetic oversampling in highly skewed, non-stationary environments

5.2. Concept drift and evolving fraud patterns

Problem: Fraudsters adapt tactics faster than models retrain, causing performance degradation within weeks [2, 6–8, 10, 11, 13, 14, 35–65].

Types of drift: The main types of concept drift observed in financial fraud detection are summarized in Table 13.

Table 13
Types of concept drift in financial fraud detection

Type	Example	Source
Gradual	Shift from card-not-present to mobile wallet fraud	[18]
Sudden	Post-pandemic surge in synthetic identity fraud	[1]
Recurring	Seasonal spikes (e.g., holiday phishing)	[19]

Evidence:

1) Alghofaili et al. [6]: LSTM model accuracy dropped from 91% to 74% after 3 months without retraining.
2) Song et al. [7]: Fraud rings restructured within 48 hours of detection, evading static GNN embeddings.
3) Wu et al. [14]: Observed drift in continuous-coupled neural networks for credit card fraud.
4) Jabeen et al. [35]: Reported rapid degradation in hybrid CNN-LSTM (CLST) models due to evolving patterns.

Open issues:

1) Online learning without catastrophic forgetting
2) Automated drift detection thresholds in production
3) Integration of drift-aware mechanisms in graph-based and federated settings

5.3. Model interpretability vs performance trade-off

Problem: High-performing deep models (LSTM, GNN) are inherently opaque, violating regulatory mandates (e.g., European Union Artificial Intelligence Act (EU AI Act), Federal Deposit Insurance Corporation (FDIC) guidelines) [2, 8, 10–14, 29, 35–65].

Regulatory pressure:

1) Institutions must provide actionable explanations for blocked transactions
2) Right to explanation under GDPR Article 22

Evidence:

1) Zhou et al. [12]: Adding SHAP to LSTM reduced F1-score by 1.5% but increased analyst acceptance by 30%.
2) Jurgovsky et al. [18]: RNN sequence decisions were untraceable, leading to manual overrides in 18% of alerts.
3) Alhasawi et al. [13]: Demonstrated trade-offs in explainable federated models for transparent banking.
4) Oduro et al. [45]: Highlighted interpretability challenges in energy-based Restricted Boltzmann Machine (RBM) and Extended Long Short-Term Memory (xLSTM).

Open issues:

1) Intrinsic explainability in GNNs and federated models
2) Standardized explanation formats for audit trails
3) Balancing minimal performance degradation with regulatory-compliant explanations

5.4. Privacy and data sharing in collaborative detection

Problem: Cross-institutional fraud (e.g., money mules, synthetic identities) requires data fusion, but raw data sharing violates privacy laws [2, 8–11, 13, 14, 29,35–65].

Constraints:

1) GDPR: No unnecessary personal data transfer
2) CCPA: Consumer opt-out rights
3) Inter-bank competition

Evidence:

1) Tang and Liang [9]: Without federation, collaborative accuracy = 78%; with federated GNN = 93%.
2) ACFE 2024 report [1]: 42% of large fraud cases involved multiple financial entities, undetected due to silos.
3) Xia and Saha [10]: Improved privacy in graph-based FL for credit card fraud.
4) Vijayanand and Smrithy [11]: Enhanced privacy in real-time banking networks via federated VAE-QLSTM fusion.
5) Alhasawi et al. [13]: Achieved scalable, trustworthy detection with federated approaches.

Open issues:

1) Secure Multi-Party Computation (SMPC) at scale
2) Differential privacy noise vs utility trade-off
3) Efficient communication in heterogeneous federated environments

5.5. Adversarial robustness and model poisoning

Problem: Attackers can craft adversarial transactions to bypass detectors or poison training data in federated settings [2, 8, 10, 11, 13, 14, 21, 29, 35–65].

Attack vectors: The primary attack vectors in AI-driven fraud detection systems are summarized in Table 14.

Table 14
Adversarial attack vectors in financial fraud detection

Vector	Description	Refs
Evasion	Micro-amount perturbations in transaction sequences	[6]
Poisoning	Malicious clients inject false gradients in FL	[9]
Backdoor	Hidden triggers in GNN node features	[7]

Evidence:

1) Liu [21]: The RL + GNN model was bypassed in 68% of adversarial tests without robust training.
2) Ileberi et al. [19]: GA-optimized features were exploitable via feature reversal attacks.
3) Hayat and Magnier [37]: Exposed deceptive performance due to adversarial vulnerabilities.

Open issues:

1) Realistic adversarial benchmarks for financial data
2) Certified robustness guarantees
3) Defense mechanisms tailored to sequential and graph-structured data

Table 15
Business-relevant evaluation metrics for fraud detection systems

Metric	Formula	Importance
Expected loss	P(fraud) × Loss × (1-Recall) + P(normal) × Cost × FP rate	[1]
Alert fatigue index	Alerts per 1000 transactions	[12]
Return on Investment (ROI) of detection	(Loss Prevented – System Cost) / System Cost	[20]

5.6. Evaluation metrics and operational costs

Problem: Academic metrics (accuracy, AUC) do not reflect business impact:

1) Cost of false positives (customer friction, support tickets)
2) Cost of false negatives (direct financial loss)

Business-relevant metrics needed: The business-relevant evaluation metrics are summarized in Table 15.

Evidence:

1) Armel and Zaidouni [20]: Spark system reduced alerts by 40%, saving $2.1M/year in manual review.
2) ACFE 2024 [1]: Average investigation time per alert = 2.5 h → high false positive (FP rate) = operational bottleneck.
3) Yaseen and Al-Amarneh [41]: Highlighted operational cost implications in AI adoption for fraud detection.

Open issues:

1) Standardized cost-sensitive evaluation frameworks
2) Real-world ROI datasets
3) Industry-wide benchmarks incorporating customer experience metrics

Summary of challenges: The six core challenges in AI-driven fraud detection are summarized in Table 16.

Path forward

> Rule Layer → Anomaly Layer → Graph Layer → XAI Layer → Human-in-the-loop.

The convergence of these challenges, observed across the evolutionary phases (Section 3) and reflected in the taxonomy gaps (Section 6), underscores the need for a multi-layered, integrated defense paradigm:

Future systems must jointly optimize for accuracy, explainability, privacy, robustness, and operational cost, a vision directly addressed through the proposed roadmaps in Section 7.

6. Taxonomy of Fraud Detection Systems

This section presents a comprehensive, multidimensional taxonomy of AI-driven financial fraud detection systems, synthesized

Table 16
Six core challenges in AI-driven fraud detection (2015–2025)

#	Challenge	Root cause	Impact	Mitigation (current)	Open gap	Key refs
1	Class imbalance	Fraud < 0.1%	Missed fraud	SMOTE, cost-sensitive	Dynamic streams	[1, 4, 17, 36, 37, 59]
2	Concept drift	Adaptive attackers	Model decay	Periodic retrain	Online learning	[6, 7, 14, 35]
3	Interpretability	Black-box DL	Regulatory rejection	SHAP, LIME	Intrinsic XAI	[12, 13, 45]
4	Privacy	Legal silos	Isolated detection	FL	Secure aggregation	[9–11, 44]
5	Adversarial	Input manipulation	Bypass	Advanced training	Certified defense	[21, 37]
6	Evaluation	Metric mismatch	Suboptimal ROI	Custom Key Performance Indicators (KPIs)	Standardized cost model	[1, 20, 41]

from the 65 reviewed studies [1–65]. The taxonomy is directly built upon the three evolutionary phases identified in Section 3 and the six persistent challenges outlined in Section 5, providing an integrated framework that links historical progression, operational realities, and future needs. It classifies systems across six orthogonal dimensions, enabling precise positioning of techniques, identification of research gaps, and practical guidance for practitioners.

6.1. Taxonomy dimensions

The six dimensions of the proposed taxonomy are summarized in Table 17.

6.2. System positioning in taxonomy

The positioning of landmark systems within the taxonomy spanning the three evolutionary phases and incorporating recent advancements is presented in Table 18.

6.3. Emerging clusters (2024–2025)

Based on reference analysis [7–10, 12, 21, 44, 45, 63], three dominant clusters have emerged in the Innovation Phase, as shown in Table 19.

6.4. Research gaps identified via taxonomy

The taxonomy, when cross-referenced with the challenges in Section 5, reveals critical research gaps, summarized in Table 20.

6.5. Taxonomy-driven design recommendations

Recommended system configurations for common operational goals aligned with evolutionary maturity and challenges are provided in Table 21.

Summary

This six-dimensional taxonomy structures the fragmented literature into a unified framework that explicitly integrates the evolutionary phases (Section 3) and operational challenges (Section 5). Drawing on all 65 studies [1–65], it reveals convergence toward graph-based, privacy-preserving, explainable, real-time systems in 2024–2025, while exposing critical gaps in multimodal, edge, and cost-aware detection. By enabling precise positioning of existing and emerging techniques, the taxonomy serves as a practical tool for practitioners to select optimal configurations and as a foundation for the future roadmaps proposed in Section 7. It remains a living blueprint, extensible to new paradigms (e.g., LLMs, quantum ML) as the field evolves.

7. Future Directions

The synthesis of the 65 reviewed studies [1–65], combined with the evolutionary phases (Section 3), persistent challenges (Section 5), and the six-dimensional taxonomy (Section 6), reveals six actionable research roadmaps. These roadmaps systematically address the critical gaps identified in the taxonomy (e.g., multimodal fusion, intrinsic XAI in GNNs, edge deployment) and unresolved challenges (e.g., concept drift, privacy in collaborative detection, cost-aware evaluation), while building on the innovations of the 2024–2025 phase. Each roadmap includes motivation (linked to specific challenges/gaps), technical pillars, expected impact, feasibility timeline, and alignment with existing work.

Roadmap 1: Multimodal fusion for contextual fraud detection

Motivation: Single-modality systems miss cross-channel fraud (e.g., phishing email → fraudulent transfer), corresponding to the multimodal gap in the taxonomy (D1) and concept drift challenge [1, 3, 14].

Technical pillars:

1) Data layer: Transaction logs + email/Short Message Service (SMS) text + social media + device telemetry
2) Model layer: Transformer-based fusion (Bidirectional Encoder Representations from Transformers (BERT) + GNN) with cross-attention

Table 17
Dimensions of the AI-driven fraud detection taxonomy

Dimension	Categories	Description	Evolution link (Section 3)	Example from refs
D1. Data Modality	Structured, Sequential, Graph, Multimodal	Type of input data processed	Early → Growth → Innovation	[16]: Structured → [7, 8, 10, 62]: Graph
D2. Learning Paradigm	Supervised, Unsupervised, Semi-Supervised, Reinforcement, Federated	Training mechanism	Classical → Deep → Federated/RL	[17]: Supervised → [9–11, 13]: Federated
D3. Detection Granularity	Transaction-Level, Account-Level, Network-Level	Scope of analysis	Transaction → Account → Network	[6]: Transaction → [8, 21]: Network
D4. Latency Requirement	Batch, Near-Real-Time (Near-RT), Real-Time (< 500 ms)	Operational timing	Batch → Near-RT → Real-Time	[20]: Batch → [8, 20]: Real-Time
D5. Explainability Level	Intrinsic, Post hoc, None	Decision transparency	Intrinsic → None → Post hoc	[16]: Intrinsic → [12, 13, 45]: Post hoc
D6. Deployment Architecture	On-Premise, Cloud, Edge, Hybrid	Infrastructure model	On-Premise/Cloud → Hybrid/Edge	[20]: Cloud → [9, 44]: Hybrid

Table 18
Positioning of landmark systems in the proposed taxonomy

System (ref)	Phase (Section 3)	D1	D2	D3	D4	D5	D6	F1-score	Use case
Bahnsen et al. [16]	Early	Structured	Supervised	Account	Batch	Intrinsic	On-Premise	0.85	Legacy banking
Alghofaili et al. [6]	Growth	Sequential	Supervised	Transaction	Near-RT	None	Cloud	0.91	Online payments
ARMEL and Zaidouni [20]	Growth	Structured	Hybrid	Transaction	Real-Time	Intrinsic	Cloud	0.92	Fintech
Zhou et al. [12]	Innovation	Sequential	Supervised	Transaction	Near-RT	Post hoc	Cloud	0.93	Regulated banks
Tang and Liang [9]	Innovation	Graph	Federated	Network	Near-RT	Post hoc	Hybrid	0.93	Cross-bank collab
Song et al. [7]	Innovation	Graph	Supervised	Network	Real-Time	Post hoc	Cloud	0.96	Money laundering
Liu [21]	Innovation	Graph	Reinforcement	Network	Real-Time	Post hoc	Hybrid	0.94	Adaptive blocking
Xia and Saha [10]	Innovation	Graph	Federated	Network	Near-RT	Post hoc	Hybrid	~0.94	Enhanced credit card fraud
Lou et al. [8]	Innovation	Graph	Supervised	Network	Real-Time	Post hoc	Cloud	~0.95	Context-encoding fraud
Abbassi et al. [11]	Innovation	Sequential	Federated	Transaction	Real-Time	Post hoc	Hybrid	~0.94	Privacy-enhanced banking
Aljunaid et al. [13]	Innovation	Graph	Federated	Network	Near-RT	Post hoc	Cloud	~0.93	Transparent banking
Alarfaj and Shahzadi [62]	Innovation	Graph	Supervised	Network	Real-Time	Post hoc	Cloud	~0.95	Real-time prevention

Table 19
Emerging research clusters in fraud detection (2024–2025)

Cluster	Dominant dimensions	Systems	% of 2024–2025 papers
Privacy-Preserving Graph AI	Graph + Federated + Network + Hybrid	[7, 9, 10, 44,45]	45%
Explainable Real-Time DL	Sequential + Supervised + Transaction + Real-Time + Post hoc	[12]	30%
Adaptive Reinforcement Systems	Graph + RL + Network + Real-Time	[21]	20%
Hybrid Graph-Explainable	Graph + Supervised + Network + Real-Time + Post hoc	[8, 63]	5% (emerging)

Table 20
Research gaps identified via the proposed taxonomy

Gap	Linked challenge (Section 5)	Evidence	Implication
Multimodal fusion	Concept drift and rare events	Only ~3% of systems use text + graph [3, 14]	Missed contextual fraud (e.g., phishing-linked)
Intrinsic XAI in GNNs	Interpretability	Most recent GNNs rely on post hoc [7, 8, 45]	Regulatory risk in high-stakes deployment
Edge real-time FL	Privacy and latency	Limited federated edge systems [11]	Undetected mobile-originated fraud
Cost-aware evaluation	Evaluation metrics	Few systems report ROI [20, 41]	Deployment hesitation in industry

3) Output layer: Unified fraud score + explanation

Expected impact: The expected impact on fraud detection performance is presented in Table 22.

Feasibility:

1) 2026: Prototype using LLMs (e.g., Large Language Model Meta AI 3 (LLaMA-3)) + GNN

Table 21
Taxonomy-driven design recommendations

Goal	Recommended configuration	Example system(s)
Regulatory compliance	Structured/Sequential + Supervised + Account + Batch + Intrinsic + On-Premise	[16]
Cross-bank collaboration	Graph + Federated + Network + Near-RT + Post hoc + Hybrid	[9–11, 13]
Zero-day fraud	Sequential + Unsupervised + Transaction + Real-Time + None + Cloud	[3]
Organized crime detection	Graph + RL/Supervised + Network + Real-Time + Post hoc + Cloud/Hybrid	[8, 21, 62]

Table 22
Expected impact of multimodal fusion (Roadmap 1)

Metric	Current	Target (2027)
Zero-day fraud detection	60% [3]	90%
False positives	35%	< 15%

2) 2027: Production in fintechs

Alignment: Extends GraphSAGE [7] and recent multimodal approaches [14] with textual embeddings, addressing the multimodal gap in the taxonomy.

Roadmap 2: Continual learning with catastrophic forgetting mitigation

Motivation: Concept drift degrades models within weeks, a core challenge in Section 5 directly linked to the lack of adaptive paradigms in earlier taxonomy clusters [6, 7, 35].

Technical pillars:

1) Memory replay: Store fraud exemplars in encrypted buffer
2) Elastic weight consolidation (EWC): Protect critical parameters
3) Federated continual learning [9]

Expected impact: The expected impact on model resilience is shown in Table 23.

Table 23
Expected impact of continual learning (Roadmap 2)

Scenario	Retrain frequency	Target
Seasonal fraud	Monthly → weekly	95% F1 sustained
Sudden attack	Daily → Hourly	< 2% accuracy drop

Feasibility:

1) 2026: EWC + FedAvg integration
2) 2027: Real-time drift dashboards

Alignment: Builds on LSTM sequential modeling [6], federated collaboration [9], and recent hybrid models [35], filling the continual learning gap across taxonomy dimensions D2 and D4.

Roadmap 3: Intrinsic explainability in GNNs

Motivation: Post hoc XAI (SHAP) adds latency and reduces trust, highlighting the interpretability challenge (Section 5) and the reliance on post hoc methods in current GNN clusters (D5 in taxonomy) [12, 45].

Technical pillars:

1) Graph attention with rationales: Learn interpretable attention weights
2) Concept bottlenecks: Map GNN layers to business concepts (e.g., "high-velocity spending")
3) Counterfactual graph edits

Expected impact: The expected impact on explainability and compliance is presented in Table 24.

Table 24
Expected impact of intrinsic XAI in GNNs (Roadmap 3)

Metric	Post hoc [12]	Intrinsic (target)
Explanation latency	1.2 s	< 200 ms
Regulatory approval rate	70%	95%

Feasibility:

1) 2026: GNN with concept activation vectors
2) 2027: Standardized XAI audit logs

Alignment: Evolves GraphSAGE [7], SHAP integration [12], and recent energy-based/explainable models [45], directly targeting the intrinsic XAI gap in taxonomy dimension D5.

Roadmap 4: Edge-deployed FL for mobile fraud

Motivation: 80% of fraud now originates on mobile with unacceptable cloud latency, linking to privacy and latency challenges (Section 5) and the absence of edge systems in the taxonomy (D6, D4) [11, 20].

Technical pillars:

1) On-device model updates (TensorFlow Lite)
2) Differential privacy (DP) noise at edge (to protect individual client updates from inference attacks through randomized noise addition)
3) SMPC techniques such as secure aggregation via cryptographic protocols (e.g., homomorphic encryption or additive secret sharing) to ensure the server only receives the aggregated model update without accessing individual gradients distinct from DP, which focuses on statistical privacy guarantees via noise, but complementary in federated settings

Expected impact: The expected impact on mobile fraud detection is shown in Table 25.

Feasibility:

1) 2026: Pilot with 3 banks (iOS/Android)
2) 2027: Global rollout

Alignment: Combines Spark real-time [20] with federated privacy [9] and recent edge-enhanced models [11], addressing the edge deployment gap in taxonomy dimension D6. Note: SMPC

Table 25
Expected impact of edge FL (Roadmap 4)

Deployment	Latency	Privacy risk
Cloud-only [20]	1.8 s	Medium
Edge FL	< 300 ms	Low

(including secure aggregation) enables collaborative computation without data sharing, while DP provides statistical privacy guarantees together enhancing robustness in privacy-preserving collaboration

Roadmap 5: RL for adaptive risk mitigation

Motivation: Static thresholds cause alert fatigue or over-blocking, tied to evaluation metrics challenge (Section 5) and limited adaptive paradigms in taxonomy clusters [21].

Technical pillars:

1) Environment: Live transaction stream
2) Agent: RL policy (Proximal Policy Optimization (PPO)/Deep Q-Network (DQN)) → action: Allow/flag/block
3) Reward: +Loss prevented − Cost of false positive

Expected impact: The expected impact on operational efficiency is presented in Table 26.

Table 26
Expected impact of RL-based risk mitigation (Roadmap 5)

Strategy	Alert rate (%)	Loss prevented
Fixed threshold	12	$8.2 M/year
RL adaptive	6	$11.5 M/year

Feasibility:

1) 2026: Sandbox with synthetic data
2) 2027: Live A/B testing

Alignment: Scales RL + GNN prototype [21] to production, enhancing adaptive reinforcement cluster in the taxonomy.

Roadmap 6: Standardized, cost-aware evaluation framework

Motivation: Academic metrics ignore operational ROI, a key evaluation challenge (Section 5) with no cost-aware systems in the taxonomy [1, 20, 41].

Technical pillars:

1) Unified benchmark dataset: 10M+ real (anonymized) transactions
2) Cost matrix: Cost of False Positive (C_FP) = $25 (support ticket); Cost of False Negative (C_FN) = $125,000 (median fraud loss [1])
3) Primary metric: Expected monetary value (EMV)

Expected impact: The expected impact on system validation is shown in Table 27.

Table 27
Expected impact of cost-aware evaluation (Roadmap 6)

Evaluation	Current	Target
Metric	F1/AUC	EMV
Deployment confidence	60%	90%

Feasibility:

1) 2026: Open-source via Kaggle/ACFE partnership
2) 2027: Adopted by the ISO fraud detection standard

Alignment: Operationalizes ACFE loss data [1], Spark ROI [20], and recent adoption studies [41], closing the cost-aware evaluation gap across all taxonomy dimensions.

Priority matrix for practitioners: The relative priority of the six roadmaps considering taxonomy gaps and challenge severity is summarized in Table 28.

Call to action

1) Industry: Form a consortium for multimodal data sharing (privacy-preserving)
2) Academia: Release continual learning benchmarks with drift injection
3) Regulators: Mandate EMV reporting in fraud system audits
4) Open source: Launch FraudAI-Toolkit integrating all six roadmaps

These roadmaps, grounded in the integrated insights from the evolutionary analysis (Section 3), challenges (Section 5), and taxonomy (Section 6), transform reactive detection into proactive, adaptive, and trustworthy financial defense systems by 2030. Unlike prior specialized surveys that often focus on isolated paradigms (e.g., GNN-based fraud detection or FL applications), this work provides an integrated decade-long synthesis linking phases, challenges, taxonomy gaps, and prioritized actionable directions offering a coordinated path beyond fragmented analyses.

8. Conclusion

This survey has systematically traced the decade-long evolution of AI-driven financial fraud detection from 2015 to 2025, analyzing 65 peer-reviewed studies through a PRISMA-compliant methodology and grounding insights in references [1, 3, 4, 6, 7, 9, 12, 15–21]. The journey reveals a profound transformation

Table 28
Priority matrix for research roadmaps

Roadmap	Technical readiness	Business impact	Regulatory fit	Priority
1. Multimodal	Medium	High	Medium	High
2. Continual	High	High	High	High
3. XAI-GNN	Low	High	High	High
4. Edge FL	Medium	High	High	High
5. RL Adaptive	Low	High	Medium	Medium
6. Evaluation	High	High	High	Critical

from rule-augmented classical models to privacy-preserving, graph-aware, explainable, and real-time intelligent systems.

To directly address the research questions posed in the Introduction, the key findings are summarized as follows:

1) RQ1 (Evolution from 2015 to 2025) The field progressed through three distinct phases: classical ML focused on imbalance handling (2015–2019), deep sequential models emphasizing scalability and temporal patterns (2020–2023), and innovative graph-based, privacy-preserving, and explainable approaches (2024–2025), as detailed in Section 3.
2) RQ2 (Superior emerging techniques in 2024–2025): GNNs, FL, and XAI integrations achieved the highest reported metrics (e.g., F1-scores up to 0.96 and 85% detection of organized fraud rings), outperforming prior paradigms in networked and privacy-sensitive scenarios [7, 9, 12, 21].
3) RQ3 (Primary operational challenges and solutions): Six persistent challenges (class imbalance, concept drift, interpretability, privacy, adversarial robustness, and cost-aware evaluation) were identified in Section 5, with partial mitigations such as SMOTE, federated training, and SHAP, though significant gaps remain.
4) RQ4 (Future research and industry directions): Six actionable roadmaps were proposed in Section 7 to address unresolved issues, providing a structured path toward multimodal, continual, and human-centered systems.

8.1. Summary of key findings

Three distinct eras

1) 2015–2019 (classical ML): Feature engineering and ensemble methods (RF, XGBoost + SMOTE) achieved F1 ≈ 0.85 but were limited by manual design and imbalance [4, 16, 17].
2) 2020–2023 (deep sequential): LSTM and CNN-LSTM enabled temporal pattern recognition, pushing F1-scores to 0.91–0.95 with cloud scalability [6, 18–20].
3) 2024–2025 (smart innovation): GNNs, FL, and XAI addressed networked fraud, privacy, and regulatory trust, reaching F1 score ≈ 0.96 [7, 9, 12, 21].

Performance convergence, not plateau, while accuracy gains have slowed, value has shifted to recall in rare events, latency (< 500 ms), privacy compliance, and actionable explanations [1, 12, 20].

Six persistent challenges—class imbalance, concept drift, interpretability, privacy, adversarial robustness, and cost-aware evaluation—remain unsolved at scale despite technical progress [1, 6, 9, 12, 17, 21].

Taxonomy as a strategic tool, the six-dimensional classification framework (Section 6), built directly on the evolutionary phases and challenges, enables precise system positioning and reveals emerging clusters:

1) Privacy-Preserving Graph AI (45%)
2) Explainable Real-Time DL (30%)
3) Adaptive Reinforcement Systems (20%)

8.2. Implications for stakeholders

The practical implications and actionable recommendations for key stakeholders are summarized in Table 29.

Table 29
Implications and actionable insights for key stakeholders

Stakeholder	Key implication	Actionable insight
Financial institutions	Must transition from siloed, static models to federated, graph-based, real-time pipelines [7, 9]	Invest in Spark + GNN + FL stacks by 2027
Regulators (FDIC, European Central Bank (ECB), FATF)	Explainability is now mandatory under the EU AI Act and Basel IV [12]	Mandate SHAP-augmented audit logs in fraud systems
Fintechs and payment providers	Edge-deployed detection is critical for mobile-first fraud [20]	Prioritize TensorFlow Lite + Federated Edge
Researchers	Focus must shift from accuracy to EMV, drift resilience, and multimodal fusion [1, 3, 6]	Release open continual learning benchmarks

8.3. Final remarks

The $5 trillion annual fraud burden reported by the ACFE in 2024 [1] is not merely a financial metric it is a societal trust crisis. AI has evolved from a promising tool to a strategic necessity, but technological maturity alone is insufficient. True resilience demands:

"A paradigm where detection is not just accurate, but adaptive, transparent, collaborative, and human-centered."

The six future roadmaps proposed in Section 7—multimodal fusion, continual learning, intrinsic XAI in GNNs, edge federated systems, adaptive RL, and cost-aware evaluation—offer a coordinated path forward that directly extends the taxonomy (Section 6) and resolves the persistent challenges (Section 5). Unlike prior specialized surveys that often focus on isolated paradigms (e.g., GNN-based fraud detection [8, 62] or FL applications [10, 23]), this work incorporates their findings while advancing beyond them through an integrated decade-long synthesis [2]. It systematically links evolutionary phases, persistent challenges, a six-dimensional taxonomy, and prioritized actionable roadmaps—highlighting unique contributions such as addressing multimodal gaps, SMPC-DP distinctions, and coordinated industry-academia bridges that move the field forward from fragmented analyses toward a comprehensive, actionable framework for trustworthy AI in fraud detection.

This survey concludes not with finality, but with urgency: The next breakthrough will not come from a single algorithm but from ecosystems that integrate privacy, explainability, and real-time intelligence at scale.

Let this work serve as both retrospective validation and prospective blueprint for researchers, engineers, and policymakers committed to securing the future of financial trust.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/.

Author Contribution Statement

Elham Shamsinejad: Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Hamid Banirostam:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration.

References

[1] Tümmler, M., & Quick, R. (2026). How to detect fraud in an audit: A systematic review of experimental literature. *Management Review Quarterly*, *76*(1), 287–332. https://doi.org/10.1007/s11301-024-00480-7

[2] Hafez, I. Y., Hafez, A. Y., Saleh, A., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, *12*(1), 6. https://doi.org/10.1186/s40537-024-01048-8

[3] Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1), 55–68. https://doi.org/10.1007/s44230-022-00004-0

[4] Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, *68*, 90–113. https://doi.org/10.1016/j.jnca.2016.04.007

[5] Li, F., & Chen, Z. (2025). Dynamic quantification anti-fraud machine learning model for real-time transaction fraud detection in banking. *Discover Computing*, *28*(1), 59. https://doi.org/10.1007/s10791-025-09549-7

[6] Alghofaili, Y., Albattah, A., & Rassam, M. A. (2020). A financial fraud detection model based on LSTM deep learning technique. *Journal of Applied Security Research*, *15*(4), 498–516. https://doi.org/10.1080/19361610.2020.1815491

[7] Song, Y., Wei, Y., Lu, Y., Sun, Q., Shao, M., Wang, L., . . . , & Fu, X. (2025). Mitigating message imbalance in fraud detection with dual-view graph representation learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 3281–3289. https://doi.org/10.24963/ijcai.2025/365

[8] Lou, C., Wang, Y., Li, J., Qian, Y., & Li, X. (2025). Graph neural network for fraud detection via context encoding and adaptive aggregation. *Expert Systems with Applications*, *261*, 125473. https://doi.org/10.1016/j.eswa.2024.125473

[9] Tang, Y., & Liang, Y. (2024). Credit card fraud detection based on federated graph learning. *Expert Systems with Applications*, *256*, 124979. https://doi.org/10.1016/j.eswa.2024.124979

[10] Xia, Z., & Saha, S. C. (2025). FinGraphFL: Financial graph-based federated learning for enhanced credit card fraud detection. *Mathematics*, *13*(9), 1396. https://doi.org/10.3390/math13091396

[11] Vijayanand, D., & Smrithy, G. S. (2025). Explainable AI—enhanced ensemble learning for financial fraud detection in mobile money transactions. *Intelligent Decision Technologies*, *19*(1), 52–67. https://doi.org/10.1177/18724981241289751

[12] Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. *Finance Research Letters*, *58*, 104309. https://doi.org/10.1016/j.frl.2023.104309

[13] Fadel, M., Tarek, A., & Zamzam, M. (2025). Defending digital transactions: Federated learning for e-wallet fraud detection. In *2025 3rd International Conference on Business Analytics for Technology and Security*, 1–7. https://doi.org/10.1109/ICBATS66542.2025.11258285

[14] Wu, Y., Wang, L., Li, H., & Liu, J. (2025). A deep learning method of credit card fraud detection based on continuous-coupled neural networks. *Mathematics*, *13*(5), 819. https://doi.org/10.3390/math13050819

[15] Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . , & McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

[16] Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, *51*, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

[17] Majumder, T., & Mishra, D. (2024). Financial fraud detection for credit card using XGBoost & SMOTE. *Nanotechnology Perceptions*, *20*(S15), 32–50. https://doi.org/10.62441/nano-ntp.vi.3425

[18] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, *100*, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037

[19] Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1), 24. https://doi.org/10.1186/s40537-022-00573-8

[20] Armel, A., & Zaidouni, D. (2019). Fraud detection using Apache Spark. In *2019 5th International Conference on Optimization and Applications*, 1–6. https://ieeexplore.ieee.org/document/8727610

[21] Liu, W. (2025). Deep reinforcement learning with graph neural networks for financial fraud risk mitigation. *Advances in Management and Intelligent Technologies*, *1*(1), 1–11. https://doi.org/10.62177/amit.v1i1.201

[22] Shamsinejad, E., Banirostam, H., Banirostam, T., Pedram, M. M., & Rahmani, A. M. (2025). Providing and evaluating a model for big data anonymization streams by using in-memory processing. *Knowledge and Information Systems*, *67*(9), 7791–7824. https://doi.org/10.1007/s10115-025-02417-2

[23] Alhasawi, Y., Almtrf, A. A., & Asad, M. (2025). A federated approach to scalable and trustworthy financial fraud detection. *Security and Privacy*, *8*(5), e70099. https://doi.org/10.1002/spy2.70099

[24] Shamsinejad, E., Banirostam, T., Pedram, M. M., & Rahmani, A. M. (2025). A review of anonymization algorithms and methods in big data. *Annals of Data Science*, *12*(1), 253–279. https://doi.org/10.1007/s40745-024-00557-w

[25] Abbassi, H., El Mendili, S., & Gahi, Y. (2025). Adaptive, privacy-enhanced real-time fraud detection in banking networks through federated learning and VAE-QLSTM fusion.

Big Data and Cognitive Computing, *9*(7), 185. https://doi.org/10.3390/bdcc9070185

[26] Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management*, *18*(4), 179. https://doi.org/10.3390/jrfm18040179

[27] Zheng, Y. (2025). Bank data protection and fraud identification based on improved adaptive federated learning and WGAN. *Scientific Reports*, *15*(1), 23006. https://doi.org/10.1038/s41598-025-06807-y

[28] AbouGrad, H., & Sankuru, L. (2025). Online banking fraud detection model: Decentralized machine learning framework to enhance effectiveness and compliance with data privacy regulations. *Mathematics*, *13*(13), 2110. https://doi.org/10.3390/math13132110

[29] Rahmati, M. (2025). Real-time financial fraud detection using adaptive graph neural networks and federated learning. *International Journal of Management and Data Analytics*, *5*(1), 98–110. https://doi.org/10.5281/zenodo.15107110

[30] Baghdadi, P., Korukoglu, S., Bilici, M. A., & Onan, A. (2025). The potential of energy-based RBM and xLSTM for real-time predictive analytics in credit card fraud detection. *Journal of Data Analysis and Information Processing*, *13*(1), 79–100. https://doi.org/10.4236/jdaip.2025.131005

[31] Banirostam, H., Banirostam, T., Pedram, M. M., & Rahmani, A. M. (2024). Analysis and evaluation of various fraud detection methods for electronic payment cards transactions in big data. *Journal of Signal Processing Systems*, *96*(12), 849–870. https://doi.org/10.1007/s11265-025-01947-w

[32] Vijayanand, D., & Smrithy, G. S. (2025). Explainable AI - enhanced ensemble learning for financial fraud detection in mobile money transactions. *Intelligent Decision Technologies*, *19*(1), 52–67. https://journals.sagepub.com/doi/full/10.1177/18724981241289751

[33] Vasant, M., Ganesan, S., & Kumar, G. (2025). Enhancing e-commerce security: A hybrid machine learning approach to fraud detection. *Fintech and Sustainable Innovation*, *1*, A7. https://doi.org/10.47852/bonviewFSI52024882

[34] Dar, A. A., Hirani, T., Arora, V., & Kakkar, S. (2025). Predictive analytics for personalized debt management: Leveraging machine learning to provide actionable financial advice. *Fintech and Sustainable Innovation*, *1*, A4. https://doi.org/10.47852/bonviewFSI52025591

[35] Jabeen, M., Ramzan, S., Raza, A., Fitriyani, N. L., Syafrudin, M., & Lee, S. W. (2025). Enhanced credit card fraud detection using deep hybrid CLST model. *Mathematics*, *13*(12), 1950. https://doi.org/10.3390/math13121950

[36] Marazqah Btoush, E. A. L., Zhou, X., Gururajan, R., Chan, K. C., & Alsodi, O. (2025). Enhancing credit card fraud detection with a stacking-based hybrid machine learning approach. *PeerJ Computer Science*, *11*, e3007. https://doi.org/10.7717/peerj-cs.3007

[37] Hayat, K., & Magnier, B. (2025). Data leakage and deceptive performance: A critical examination of credit card fraud detection methodologies. *Mathematics*, *13*(16), 2563. https://doi.org/10.3390/math13162563

[38] Bandarapu, S. R., Randhi, K., & Manjula, N. J. (2025). Enhancing credit card fraud detection using cutting-edge advanced machine learning techniques. *Cureus Journal of Computer Science*, *2*, es44389-025-03921-w. https://doi.org/10.7759/s44389-025-03921-w

[39] Salunke, Y., Phalke, S., Madavi, M., Kumre, P., & Bobhate, G. (2025). Fraud detection: A hybrid approach with logistic regression, decision tree, and random forest. *Cureus Journal of Computer Science*, *2*, es44389-024-02350-5. https://doi.org/10.7759/s44389-024-02350-5

[40] Preciado Martínez, P. M., Reier Forradellas, R. F., Garay Gallastegui, L. M., & Náñez Alonso, S. L. (2025). Comparative analysis of machine learning models for the detection of fraudulent banking transactions. *Cogent Business & Management*, *12*(1), 2474209. https://doi.org/10.1080/23311975.2025.2474209

[41] Yaseen, H., & Al-Amarneh, A. (2025). Adoption of artificial intelligence-driven fraud detection in banking: The role of trust, transparency, and fairness perception in financial institutions in the United Arab Emirates and Qatar. *Journal of Risk and Financial Management*, *18*(4), 217. https://doi.org/10.3390/jrfm18040217

[42] Ajmire, S. S., Shahale, K. K., Thakare, S. R., Agarkar, K. V., Wankhade, A. S., & Mahobia, R. (2025). An AI-driven privacy-preserving framework for intelligent risk assessment and fraud detection in financial transactions. In *2025 3rd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry*, 1–6. https://doi.org/10.1109/IDICAIHEI65991.2025.11379258

[43] Allen, J. S. (2025). *CardSim: A Bayesian simulator for payment card fraud detection research* (Finance and Economics Discussion Series 2025-017). Board of Governors of the Federa Reserve System. https://doi.org/10.17016/FEDS.2025.017

[44] Jin, J., & Zhang, Y. (2025). The analysis of fraud detection in financial market under machine learning. *Scientific Reports*, *15*(1), 29959. https://doi.org/10.1038/s41598-025-15783-2

[45] Oduro, A. A., Okolo, J. N., Bello, A. D., Ajibade, A. T., Fatomi, A. M., Oyekola, T. S., & Owoo-Adebayo, S. F. (2025). AI-powered fraud detection in digital banking: Enhancing security through machine learning. *International Journal of Science and Research Archive*, *14*(3), 1412–1420. https://doi.org/10.30574/ijsra.2025.14.3.0854

[46] Tayebi, M., & El Kafhali, S. (2025). A novel approach based on XGBoost classifier and Bayesian optimization for credit card fraud detection. *Cyber Security and Applications*, *3*, 100093. https://doi.org/10.1016/j.csa.2025.100093

[47] Novak, W. (2025). *Reg-tech, open banking, and AI-based fraud defense in fintech*. SSRN. http://dx.doi.org/10.2139/ssrn.5386290

[48] Chilukala, R. (2025). AI-driven fraud detection models in cloud-based banking ecosystems: A comprehensive analysis. *European Journal of Computer Science and Information Technology*, *13*(48), 45–55. https://doi.org/10.37745/ejcsit.2013/vol13n484555

[49] Mishra, B. R., Gadasandula, K., Saini, G., R, S., Deshpande, A. V., Jeyaprakash, S., & Dangarwala, V. J. (2025). The role of artificial intelligence in fraud detection and prevention in banking. *Journal of Information Systems Engineering and Management*, *10*(49s), 1167–1173. https://doi.org/10.52783/jisem.v10i49s.10061

[50] Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. Deep learning in financial fraud detection: Innovations, challenges, and applications. *Data Science and Management*, *9*(2), 100162. https://doi.org/10.1016/j.dsm.2025.08.002

[51] Al Lawati, H. M. R., Zainal, A., Al-Rimy, B. A. S., Al-Azawi, M., Kassim, M. N., Almalki, S. A., & Alghamdi, T. A. (2025). An integrated preprocessing and drift detection

approach with adaptive windowing for fraud detection in payment systems. *IEEE Access*, *13*, 92036–92056. https://doi.org/10.1109/ACCESS.2025.3569609

[52] Islam, M. S., & Rahman, N. (2025). AI-driven fraud detections in financial institutions: A comprehensive study. *Journal of Computer Science and Technology Studies*, *7*(1), 100–112. https://doi.org/10.32996/jcsts.2025.7.1.8

[53] Dichev, A., Zarkova, S., & Angelov, P. (2025). Machine learning as a tool for assessment and management of fraud risk in banking transactions. *Journal of Risk and Financial Management*, *18*(3), 130. https://www.mdpi.com/1911-8074/18/3/130

[54] Goyal, K., Garg, M., & Malik, S. (2025). Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: A hybrid approach (SEM-ANN). *Future Business Journal*, *11*(1), 44. https://doi.org/10.1186/s43093-025-00464-3

[55] Sethupathy, U. K. A. (2025). Risk-aware AI models for financial fraud detection: Scalable inference from big transactional data. *International Journal of Intelligence Science*, *15*(4), 162–183. https://doi.org/10.4236/ijis.2025.154009

[56] Koppireddy, C. S., & Vivarjitha Devi R. V. D. S. (2025). Fraud detection in banking: A deep learning approach with explainable AI. *Journal of Soft Computing Paradigm*, *7*(3), 258–275. https://irojournals.com/jscp/article/view/7/3/4

[57] Gujrati, R., & Uygun, H. (2025). Integrating artificial intelligence and blockchain for assessing the financial risk of fraud in banking sector. *PromptAI Academy Journal*, *4*, e089. https://doi.org/10.37497/PromptAI.4.2025.89

[58] Albalawi, T., & Dardouri, S. (2025). Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation. *Frontiers in Artificial Intelligence*, *8*, 1643292. https://doi.org/10.3389/frai.2025.1643292

[59] Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). Fraudx AI: An interpretable machine learning framework for credit card fraud detection on imbalanced datasets. *Computers*, *14*(4), 120. https://doi.org/10.3390/computers14040120

[60] Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S., & Qiu, G. (2025). A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud. *Scientific Reports*, *15*(1), 218. https://doi.org/10.1038/s41598-024-82062-x

[61] Friska, R., & Agustia, D. (2025). Enhancing auditor judgment quality: A review of evidence from experimental research. *Journal of Risk and Financial Management*, *18*(3), 115. https://doi.org/10.3390/jrfm18030115

[62] Khaled Alarfaj, F., & Shahzadi, S. (2025). Enhancing fraud detection in banking with deep learning: Graph neural networks and autoencoders for real-time credit card fraud prevention. *IEEE Access*, *13*, 20633–20646. https://doi.org/10.1109/ACCESS.2024.3466288

[63] de Campos Souza, P. V., & Sayyadzadeh, I. (2025). GWO-FNN: Fuzzy neural network optimized via Grey Wolf Optimization. *Mathematics*, *13*(7), 1156. https://doi.org/10.3390/math13071156

[64] Feng, X., & Kim, S.-K. (2025). Statistical data-generative machine learning-based credit card fraud detection systems. *Mathematics*, *13*(15), 2446. https://doi.org/10.3390/math13152446

[65] Song, X., Chen, Z., Wang, X., & Zhu, H. (2025). FIN-SIGN: A GNN-based learning model for online lending fraud detection. In *2025 IEEE International Conference on Systems, Man, and Cybernetics,* 4450–4455. https://doi.org/10.1109/SMC58881.2025.11343604

How to Cite: Shamsinejad, E., & Banirostam, H. (2026). A Comprehensive Survey on AI-Driven Financial Fraud Detection: Trends, Techniques, and Future Directions. *FinTech and Sustainable Innovation*. https://doi.org/10.47852/bonviewFSI62028230