Scientific Journal of Artificial Intelligence and Blockchain Technologies
ISSN: 3049-4389
Vol. 3, Issue 3, Jul – Sep 2026 || PP. 7 – 13 https://sjaibt.org

Role of Artificial Intelligence in Enhancing Fraud Detection Capabilities Across Financial Platforms

Prashant Vitthalrao Nandanvare
Assistant Professor, Department of English
KES Arts, Commerce and Science College Kalwan
Savitribai Phule Pune University, Pune
ORCID id - https://orcid.org/0009-0004-3948-6711

Date of Submission: 01-07-2026 Date of Acceptance: 15-07-2026 Date of Publication: 01-08-2026

Abstract— Financial fraud has escalated in scale and sophistication alongside the digitalisation of banking, payments, insurance, and cryptocurrency platforms, exposing the limitations of static, rule-based detection systems. Artificial intelligence (AI), encompassing machine learning, deep learning, and graph-based methods, has emerged as the dominant paradigm for identifying fraudulent activity across these platforms. This paper presents a systematic review and conceptual synthesis of the literature on the role of AI in enhancing fraud detection capabilities across financial platforms. Drawing on foundational data-mining frameworks and recent systematic reviews published between 2011 and 2026, the review traces the field's evolution from early statistical and rule-based approaches to supervised machine learning, deep neural architectures, and relational graph neural networks (GNNs) capable of modelling collusive and networked fraud. The synthesis indicates that AI substantially improves detection accuracy and adaptability relative to traditional methods, particularly for credit card fraud, insurance fraud, and money laundering, but that persistent challenges, severe class imbalance, concept drift, limited labelled data, and the trade-off between predictive performance and interpretability, continue to constrain real-world deployment. The paper concludes that AI offers substantial and demonstrable value for fraud detection, but that its effectiveness in practice depends on addressing data quality, regulatory interpretability requirements, and the operational realities of real-time financial systems.

***Keywords*— artificial intelligence, fraud detection, machine learning, financial platforms, graph neural networks, anomaly detection**

Introduction

Financial platforms, spanning retail banking, card payments, online lending, insurance, and increasingly decentralised cryptocurrency ecosystems, generate transactional data at a volume and velocity that has outpaced the capacity of manual review and static rule-based controls. Fraud schemes have correspondingly grown more diverse and adaptive, ranging from stolen-card transactions and synthetic identity creation to coordinated money-laundering networks and decentralised finance exploits such as flash-loan attacks. Traditional fraud detection relied heavily on fixed thresholds, blacklists, and expert-defined rules; while transparent and easy to audit, such systems are rigid, generate high false-positive rates, and struggle to adapt as fraudulent behaviour evolves. These limitations have driven sustained interest, across both academia and industry, in artificial intelligence (AI) as a means of detecting fraud that would otherwise evade rule-based controls.

AI-based fraud detection encompasses a broad family of techniques, supervised machine learning classifiers, unsupervised anomaly detection, deep neural networks, and, more recently, graph neural networks that explicitly model the relational structure of financial transactions. Each of these approaches offers a different balance of predictive accuracy, adaptability, computational cost, and interpretability. The appeal of AI lies in its capacity to learn complex, non-linear patterns from historical data and to adapt as new information

becomes available, allowing detection systems to identify fraud typologies that were not explicitly anticipated by system designers. At the same time, the deployment of AI in financial fraud detection raises distinct challenges: fraud is a rare-event problem characterised by extreme class imbalance, fraud labels are frequently delayed or incomplete, fraudulent behaviour shifts over time in response to detection efforts, and regulatory environments increasingly demand that automated decisions be explainable and auditable.

This paper addresses these issues by systematically reviewing the academic literature on the role of AI in enhancing fraud detection capabilities across financial platforms. It aims to trace the evolution of AI-based approaches from early data-mining techniques to contemporary deep learning and graph-based methods, to synthesise evidence on their comparative effectiveness, and to identify the persistent barriers that continue to limit their real-world deployment. The paper proceeds as follows: first, a review of the theoretical and empirical literature is presented; second, the theoretical framework underpinning the synthesis is outlined; third, the review methodology is described; fourth, the findings emerging from the synthesis are presented; and finally, the implications of these findings are discussed, followed by concluding recommendations for research and practice.

Literature Review

Evolution from Rule-Based Systems to Data Mining and Machine Learning

The earliest systematic academic treatment of data-driven financial fraud detection is generally attributed to Ngai et al. (2011), whose classification framework and academic review of the literature examined a substantial body of research published between 1997 and 2008. Their review found that logistic models, neural networks, Bayesian belief networks, and decision trees were the dominant data-mining techniques applied to financial fraud detection, and that these techniques had been most extensively applied to insurance fraud, with corporate fraud and credit card fraud also attracting substantial attention. Ngai et al. further identified a comparative neglect of mortgage fraud, money laundering, and securities and commodities fraud, a gap that has only partially closed in the subsequent literature. This early classification work established the conceptual vocabulary, categorising fraud by domain and detection technique by data-mining family, that continues to structure the field.

Published in the same special issue, Bhattacharyya et al. (2011) undertook one of the first rigorous comparative evaluations of machine learning classifiers specifically for credit card fraud, benchmarking support vector machines, random forests, and logistic regression on real transaction data. Their comparative study demonstrated that ensemble-based methods such as random forests could outperform single classifiers on skewed transaction data, foreshadowing the field's later shift towards ensemble and tree-based methods as a practical baseline for tabular fraud detection tasks.

Machine Learning and Deep Learning Approaches to Fraud Detection

Ali et al. (2022), employing the Kitchenham systematic literature review protocol, synthesised a substantial corpus of machine-learning-based financial fraud detection studies and found that support vector machines and artificial neural networks were the most frequently applied algorithms, with credit card fraud remaining the most extensively studied fraud type. Their review also highlighted those recurring methodological challenges, particularly severe class imbalance in transaction datasets, continued to shape both model selection and reported performance across the reviewed literature.

More recently, Yang et al. (2026) conducted a systematic review of AI-based financial fraud detection studies published between 2015 and 2025, following PRISMA guidelines across a corpus of 122 studies. Their review found that machine learning methods remained the dominant paradigm, accounting for the largest share of reviewed studies, but that deep learning and ensemble methods together now represent a substantial and growing proportion of the literature, reflecting a broader shift towards higher-capacity models capable of capturing non-linear interactions in transaction data. Importantly, Yang et al. observed that supervised models frequently report very high overall accuracy on benchmark datasets, often exceeding 98 per cent, yet the gap between accuracy and F1-score in the same studies illustrates the central technical difficulty of extreme class imbalance: a model can achieve excellent aggregate

accuracy while still failing to capture a meaningful share of genuine fraud cases. This observation reinforces earlier findings by Hilal et al. (2022), whose review of anomaly detection techniques for financial fraud emphasised that evaluation metrics insensitive to class imbalance can systematically overstate the practical value of a detection model.

Graph-Based and Network-Aware Fraud Detection

A distinct and increasingly influential strand of the literature addresses fraud types in which relationships between entities, rather than the attributes of isolated transactions, constitute the primary detection signal. Pourhabibi et al. (2020) conducted a systematic literature review of graph-based anomaly detection approaches, finding that representing financial data as networks of accounts, devices, or counterparties allows detection systems to surface collusive rings, money-laundering chains, and other coordinated fraud patterns that conventional, transaction-level classifiers tend to overlook because such classifiers evaluate each transaction in isolation.

Building on this foundation, Cheng et al. (2024) reviewed the application of graph neural networks (GNNs) specifically to financial fraud detection, proposing a unified framework for categorising existing GNN methodologies. Their review concluded that GNNs are exceptionally effective at capturing complex relational patterns and dynamics within financial networks, and that they significantly outperform traditional, non-relational fraud detection methods on tasks such as money-laundering detection and collusive fraud identification. At the same time, Cheng et al. identified substantial deployment barriers specific to graph-based methods, including the computational cost of aggregating information across large and continuously evolving transaction graphs, the sensitivity of graph construction choices (which entities become nodes, which relationships become edges) to modelling outcomes, and the risk that temporal information can leak across training and evaluation partitions if graphs are not constructed with careful attention to time ordering. Yang et al. (2026) similarly noted that graph-based models, while capable of exploiting relational structure invisible to tabular classifiers, introduce non-trivial engineering overhead through graph construction and neighbourhood aggregation at scale, and that graph supervision is typically sparse and concentrated in regions of the transaction network that have already been investigated, introducing a further source of selection bias into model training.

Challenges: Class Imbalance, Concept Drift, and Interpretability

Across both the classical machine learning literature and the more recent graph-based and deep learning literature, three challenges recur with notable consistency. The first is class imbalance: because genuine fraud represents a small fraction of total transaction volume, models trained without careful attention to this imbalance tend to optimise for the majority, legitimate class, producing models that appear accurate in aggregate while missing a meaningful share of actual fraud. Ali et al. (2022) and Yang et al. (2026) both identify this as one of the most persistent methodological issues in the literature, noting that resampling strategies such as oversampling and undersampling carry their own trade-offs, including the risk of overfitting to synthetic minority-class patterns or discarding informative majority-class behaviour.

The second recurring challenge is concept drift: fraud is an inherently adversarial and adaptive phenomenon, and fraudulent actors continuously adjust their behaviour in response to detection systems, meaning that a model trained on historical data can degrade in performance once fraud tactics shift. Yang et al. (2026) note that many widely used public benchmark datasets cover only short time windows, which limits the field's capacity to meaningfully evaluate how detection models perform under sustained temporal drift, a limitation that also complicates comparison across studies that rely on such benchmarks.

The third recurring challenge concerns interpretability and regulatory compliance. As detection models have grown more complex, moving from logistic regression and decision trees towards deep neural networks and graph neural networks, their internal decision logic has become correspondingly harder to explain in terms that satisfy auditors, regulators, and affected customers. Yang et al. (2026) observe that while post hoc explanation techniques have improved transparency in complex models, such explanations can be unstable under adversarial perturbations or shifts in the underlying data distribution, which limits their reliability in the high-stakes, tightly regulated environments in which fraud detection systems typically operate. This tension between predictive performance and explainability is a recurring theme that distinguishes financial fraud detection from many other machine learning application domains, given the direct regulatory and legal consequences of automated fraud determinations.

Theoretical Framework

This review is guided by two complementary conceptual perspectives drawn from the literature. First, the classification framework proposed by Ngai et al. (2011) provides a structure for organising fraud detection research along two dimensions: the type of financial fraud being addressed (for example, credit

card, insurance, corporate, or money-laundering fraud) and the family of data-mining or AI technique applied to detect it (statistical models, supervised classifiers, unsupervised anomaly detection, or network-based methods). This framework is used in the present review to organise the synthesis of methodological developments across fraud domains, and to highlight where research attention has been concentrated versus comparatively neglected.

Second, the review draws on the relational perspective articulated by Pourhabibi et al. (2020) and extended by Cheng et al. (2024), which conceptualises fraud not solely as a property of individual transactions or accounts but as an emergent pattern within a network of relationships. This relational lens is used to explain why graph-based methods have become increasingly prominent for fraud types, such as money laundering and collusive rings, that are difficult to detect using transaction-level features alone. Together, these two frameworks allow the review to examine not only which AI techniques have been applied to fraud detection, but why particular techniques are suited to particular fraud typologies, and under what conditions their advantages are most pronounced.

Research Objectives

This review pursues the following objectives:

To trace the evolution of AI-based approaches to financial fraud detection, from early data-mining techniques to contemporary deep learning and graph-based methods.

To synthesise empirical evidence on the comparative effectiveness of machine learning, deep learning, and graph-based approaches across major fraud types.

To identify the methodological and operational challenges, including class imbalance, concept drift, and interpretability, that shape the real-world deployment of AI-based fraud detection systems.

To derive evidence-based implications for the design and evaluation of AI-based fraud detection capabilities across financial platforms.

Research Questions

How has the application of artificial intelligence to financial fraud detection evolved from early data-mining approaches to contemporary deep learning and graph-based methods.

To what extent do machine learning, deep learning, and graph-based approaches differ in their effectiveness across major fraud types such as credit card fraud, insurance fraud, and money laundering.

What methodological and operational challenges most significantly constrain the real-world deployment of AI-based fraud detection systems on financial platforms.

Methodology

This paper adopts a systematic literature review and conceptual synthesis design, appropriate for consolidating a body of research that spans more than a decade and multiple distinct methodological traditions, from classical data mining to deep learning and graph neural networks.

Search Strategy

Relevant literature was identified through structured searches of academic databases and indexing services, including IEEE Xplore, ScienceDirect, MDPI journals, and arXiv, alongside targeted searches for foundational and highly cited works in the field. Search terms combined variants of "financial fraud detection," "artificial intelligence," "machine learning," "deep learning," "graph neural network," and "anomaly detection." Searches were restricted to peer-reviewed journal articles, systematic reviews, and rigorously reviewed preprints published in English.

Inclusion and Exclusion Criteria

Studies were included if they examined the application of AI, machine learning, deep learning, or graph-based methods to the detection of fraud within financial platforms, including banking, card payments, insurance, or cryptocurrency contexts, and were published in a peer-reviewed outlet or a well-established preprint venue with subsequent journal publication. Foundational and systematic review papers were prioritised given their capacity to synthesise a broad base of primary studies. Studies focused exclusively on fraud detection in non-financial domains, such as general network intrusion detection without a financial context, were excluded.

Data Extraction and Synthesis

For each included study, information was extracted on the fraud type addressed, the AI technique or techniques employed, the reported comparative findings, and any identified methodological or deployment challenges. Given the heterogeneity of the reviewed literature, spanning classification frameworks, systematic reviews, and technique-specific surveys, a narrative thematic synthesis approach was used rather than a statistical meta-analysis. Findings were organised

inductively into thematic categories, which subsequently informed the structure of the literature review and results sections of this paper.

Results and Findings

The synthesis of the reviewed literature yields four principal findings regarding the role of AI in enhancing fraud detection capabilities across financial platforms.

First, AI-based methods have progressively displaced purely rule-based and manual fraud detection approaches, with the field moving from early statistical and data-mining techniques identified by Ngai et al. (2011) towards increasingly sophisticated supervised and ensemble learning methods. Ali et al. (2022) found that support vector machines and artificial neural networks remained the most popular algorithms in the reviewed literature, while Yang et al. (2026) documented a clear methodological transition in which deep learning and ensemble methods now account for a substantial share of published studies, alongside a continuing majority of machine learning approaches. This indicates steady, rather than abrupt, methodological progression, with newer techniques supplementing rather than fully displacing established classifiers.

Second, the comparative effectiveness of AI techniques varies meaningfully across fraud types. Bhattacharyya et al. (2011) demonstrated that ensemble methods such as random forests can outperform single classifiers for credit card fraud specifically, a finding consistent with the continued prominence of tree-based and ensemble methods in more recent reviews. For relational fraud types such as money laundering and collusive fraud rings, however, Pourhabibi et al. (2020) and Cheng et al. (2024) found that graph-based methods, which explicitly model relationships between accounts, transactions, and counterparties, substantially outperform transaction-level classifiers that treat each transaction as an independent observation. This suggests that the optimal AI technique is not universal but depends on whether the fraud signal resides primarily in the attributes of individual transactions or in the relational structure connecting multiple entities.

Third, model performance as conventionally reported can substantially overstate real-world detection capability. Yang et al. (2026) observed that supervised models frequently report accuracy figures exceeding 98 per cent on benchmark datasets, yet the corresponding F1-scores reveal that meaningful proportions of genuine fraud cases are still missed, a direct consequence of the extreme class imbalance that characterises fraud data. Hilal et al. (2022) reinforced this point, cautioning that evaluation metrics insensitive to class imbalance can create a misleading impression of model readiness for operational deployment.

Fourth, the deployment of AI-based fraud detection systems, particularly deep learning and graph-based methods, is constrained by challenges that extend beyond predictive accuracy. Cheng et al. (2024) identified the computational cost of large-scale graph aggregation, the sensitivity of graph construction choices to model outcomes, and the risk of temporal information leakage as specific barriers to the real-world deployment of graph neural networks. Yang et al. (2026) further identified concept drift, arising from the adaptive and adversarial nature of fraudulent behaviour, and the instability of post hoc interpretability methods under distributional shift as persistent obstacles to translating strong offline performance into reliable, auditable, real-time detection systems.

Discussion

The findings of this review point to a central conclusion: artificial intelligence has substantially enhanced the technical capacity of financial platforms to detect fraud, but the realisation of this capacity in practice depends on confronting challenges that are distinct from, and in some respects more difficult than, the challenge of achieving strong predictive performance on a benchmark dataset. The field's trajectory, from the statistical and data-mining techniques catalogued by Ngai et al. (2011), through the machine learning methods synthesised by Ali et al. (2022), to the deep learning and graph-based methods reviewed by Cheng et al. (2024) and Yang et al. (2026), demonstrates a genuine and cumulative improvement in the sophistication of available techniques. Each successive generation of methods has extended the range of fraud patterns that can, in principle, be detected, from simple statistical outliers to complex, temporally extended, and relationally structured fraud schemes.

However, the review also makes clear that the gap between benchmark performance and deployable, trustworthy fraud detection remains substantial. The persistence of severe class imbalance across the literature, from Bhattacharyya et al. (2011) through to Yang et al. (2026), indicates that this is not a problem that has been resolved by more sophisticated modelling techniques alone; rather, it requires deliberate attention to evaluation design, using metrics such as precision-recall curves and cost-sensitive measures that better reflect the operational consequences of missed fraud and false alarms. Similarly, the recurring finding that graph-based and deep learning methods introduce meaningful computational and interpretability costs (Cheng et al., 2024) suggests that the choice of AI technique for a given financial platform should be guided not only by offline predictive performance but also by

the platform's latency requirements, regulatory obligations, and capacity to support human review of flagged transactions.

A further implication concerns the differential suitability of AI techniques across fraud types. The relational perspective advanced by Pourhabibi et al. (2020) and Cheng et al. (2024) suggests that financial platforms seeking to detect coordinated or networked fraud, such as money laundering or collusive rings, are likely to gain more from investment in graph-based methods than from further refinement of transaction-level classifiers alone. Conversely, for high-volume, low-latency contexts such as real-time card authorisation, the continued prominence of ensemble and tree-based methods documented across the reviewed literature suggests that computationally lighter classifiers may remain the more practical choice, even where deep learning or graph-based alternatives offer marginally superior offline accuracy.

Finally, the recurring emphasis on interpretability across the more recent literature (Yang et al., 2026) signals those methodological advances in predictive performance cannot be considered in isolation from the regulatory and institutional context in which fraud detection systems operate. As AI techniques become more complex, the burden of demonstrating that automated fraud determinations are fair, auditable, and resistant to adversarial manipulation grows correspondingly, and this burden is likely to shape which techniques financial institutions are willing and able to deploy at scale.

Limitations of the Study

This review is subject to several limitations. As a narrative thematic synthesis rather than a statistical meta-analysis, it does not provide a quantified, pooled estimate of the performance improvement that AI techniques offer over rule-based or earlier statistical methods, and the studies included vary considerably in scope, dataset characteristics, and methodological rigour. The review draws substantially on systematic reviews and classification frameworks rather than exhaustively re-analysing every primary study within those reviews, which means that some nuance at the level of individual studies may not be fully captured. Additionally, because publicly available fraud datasets are concentrated in certain domains, particularly credit card fraud, the underlying evidence base itself may over-represent some fraud types relative to others, such as insurance fraud or cross-border money laundering, a limitation that is inherited from the primary literature rather than introduced by this review.

Conclusion and Recommendations

Artificial intelligence has become a central and increasingly indispensable component of fraud detection across financial platforms, enabling the identification of fraudulent patterns that static, rule-based systems are structurally unable to capture. This review has traced the field's evolution from the foundational data-mining classification work of Ngai et al. (2011) and the early comparative evaluations of Bhattacharyya et al. (2011), through the machine-learning-dominated literature synthesised by Ali et al. (2022), to the graph-based and deep learning methods reviewed by Pourhabibi et al. (2020), Cheng et al. (2024), and Yang et al. (2026). Across this trajectory, AI techniques have consistently demonstrated the capacity to detect fraud patterns, from simple statistical anomalies to complex relational schemes, that elude conventional detection methods. At the same time, persistent challenges relating to class imbalance, concept drift, and interpretability mean that predictive sophistication alone does not guarantee effective real-world deployment.

Based on these findings, several recommendations follow. Financial institutions and platform operators should select AI techniques with explicit reference to the structural characteristics of the fraud type being targeted, favouring graph-based methods where relational or collusive fraud is the primary concern, and favouring computationally lighter ensemble methods where low-latency, high-volume screening is required. Evaluation practices should move beyond aggregate accuracy towards metrics, such as precision-recall analysis and cost-sensitive measures, that reflect the true operational consequences of missed fraud and false alarms under conditions of severe class imbalance. Institutions should also invest in mechanisms for detecting and responding to concept drift, given the adversarial and adaptive nature of fraudulent behaviour, rather than relying on periodic retraining alone. Finally, given the tightening regulatory expectations around explainability, future development of AI-based fraud detection systems should treat interpretability not as an afterthought but as a design requirement alongside predictive accuracy. Future research would benefit from standardised, temporally realistic benchmark datasets that span a wider range of fraud types, enabling more rigorous, comparable evaluation of AI techniques under the conditions that most closely resemble real-world deployment.

References

[1] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences, 12(19), 9637. https://doi.org/10.3390/app12199637

[2] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study.

Decision Support Systems, 50(3), 602-613. https://doi.org/10.1016/j.dss.2010.08.008

[3] Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2024). Graph neural networks for financial fraud detection: A review. Frontiers of Computer Science, 19, Article 40474. https://doi.org/10.1007/s11704-024-40474-y

[4] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications, 193, 116429. https://doi.org/10.1016/j.eswa.2021.116429

[5] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569. https://doi.org/10.1016/j.dss.2010.08.006

[6] Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 133, 113303. https://doi.org/10.1016/j.dss.2020.113303

[7] Yang, H., Shukur, Z., & Sahran, S. (2026). A review of artificial intelligence for financial fraud detection. Applied Sciences, 16(4), 1931. https://doi.org/10.3390/app16041931