A Comprehensive Review on Financial Fraud Detection and Agentic AI

Dr. Chetan Bulla
Research Scholar
Lincoln University College, 47301, Petaling Jaya,
Selangor Darul Ehsan , Malaysia.
bulla.chetan@gmail.com

Shashi Kant Gupta
Lincoln University College, 47301, Petaling Jaya,
Selangor Darul Ehsan , Malaysia.
raj2008enator@gmail.com.

Abstract: The exponential growth of digital financial ecosystems has intensified the scale and sophistication of financial fraud. This paper presents a comprehensive survey of fraud detection techniques spanning machine learning (ML), deep learning (DL), and emerging generative AI (GenAI) approaches. Recent advancements (2021–2026) are critically analyzed with respect to accuracy, adaptability, and real-world deployment constraints. The study highlights persistent challenges such as class imbalance, concept drift, and lack of autonomous decision-making. To address these limitations, an Agentic AI-based fraud detection framework is proposed, emphasizing autonomous reasoning, continuous learning, and multi-agent collaboration. The paper also integrates explainable AI (XAI) mechanisms to ensure transparency and regulatory compliance. Comparative analyses and architectural insights provide a foundation for next-generation intelligent fraud detection systems.

Keywords: Financial Fraud Detection, Machine Learning, Deep Learning, Generative AI, Agentic AI, Explainable AI, Banking Security

1. Introduction

The global financial ecosystem has undergone a profound transformation with the rapid adoption of digital technologies, including online banking, mobile payment platforms, and real-time transaction systems. Innovations such as Unified Payments Interface (UPI), digital wallets, and cross-border payment gateways have significantly enhanced transaction speed, accessibility, and financial inclusion. However, this digital expansion has also introduced new vulnerabilities, making financial systems increasingly susceptible to sophisticated fraud schemes. As transaction volumes continue to grow exponentially, traditional monitoring approaches have become insufficient to detect and prevent fraudulent activities in a timely and efficient manner [1].

Financial fraud has evolved from simple, isolated incidents into highly organized and technology-driven operations. Modern fraudsters exploit advanced techniques such as phishing, identity theft, account takeover, and coordinated fraud rings that operate across multiple channels and geographies. These attacks are often dynamic and adaptive, continuously changing their patterns to bypass detection systems. The complexity of fraud is further amplified by the heterogeneous nature of financial data, which includes transactional, behavioral, geospatial, and device-related information. As a result, fraud detection has become a multi-dimensional challenge that requires intelligent, scalable, and adaptive solutions [2].

Over the past two decades, Artificial Intelligence (AI) has emerged as a cornerstone in modern fraud detection systems. Early approaches relied on rule-based systems, which were easy to interpret but lacked flexibility and scalability. With the growth of data availability, machine learning techniques such as decision trees, random forests, and gradient boosting algorithms became widely adopted due to

their ability to learn complex patterns from historical data. More recently, deep learning models, including neural networks and recurrent architectures, have further improved detection capabilities by capturing non-linear relationships and temporal dependencies in transaction sequences. Despite these advancements, existing AI models still face significant limitations in real-world deployment, including high false-positive rates, lack of transparency, and difficulty adapting to evolving fraud patterns [3].

One of the most critical challenges in fraud detection is the phenomenon of concept drift, where the statistical properties of data change over time due to shifts in user behavior, emerging fraud strategies, and evolving regulatory environments. Static or periodically retrained models often struggle to keep pace with these changes, resulting in degraded performance and increased operational risk. Additionally, most current systems operate in a linear pipeline—data ingestion, model prediction, and alert generation—without the ability to autonomously take actions, incorporate feedback, or dynamically adjust their strategies. This lack of autonomy leads to heavy reliance on human analysts, increased investigation costs, and delayed response to fraudulent activities [4].

To address these limitations, recent research has begun exploring the integration of Generative AI and Agentic AI in fraud detection systems. While Generative AI enables advanced capabilities such as synthetic data generation and anomaly detection, Agentic AI introduces a fundamentally new paradigm based on autonomous, goal-driven agents capable of reasoning, decision-making, and continuous learning. By enabling systems to not only detect fraud but also respond intelligently and adapt over time, Agentic AI has the potential to transform fraud detection into a proactive and self-evolving process. This paper aims to provide a comprehensive review of existing fraud detection techniques and highlight the emerging role of Agentic AI as a next-generation solution for building robust, adaptive, and trustworthy financial security systems[5].

2. Machine Learning-Based Models

Machine learning (ML) remains the most widely adopted paradigm in financial fraud detection due to its ability to learn complex patterns from large-scale transactional data. Unlike rule-based systems, ML models can generalize from historical fraud behavior and identify subtle anomalies in real time. In recent years (2024–2026), research has focused on improving detection accuracy under extreme class imbalance, enhancing interpretability, and optimizing models for real-world deployment constraints such as latency and scalability [6].

Recent research in financial fraud detection has increasingly converged toward the use of advanced ensemble learning techniques, particularly gradient boosting methods such as XGBoost [7], LightGBM, and CatBoost [8]. These models have demonstrated strong predictive capability in large-scale transactional datasets due to their ability to capture complex, non-linear relationships among features. A comparative study conducted in 2025 on large credit card datasets reported that boosting-based models consistently outperform traditional classifiers in both accuracy and F1-score, especially under highly imbalanced conditions [9]. Similarly, large-scale benchmarking across multiple datasets confirms that XGBoost and Random Forest remain among the most reliable approaches, achieving superior performance across diverse evaluation metrics . Despite their effectiveness, these models often depend heavily on careful feature engineering and hyperparameter tuning, which can limit their adaptability in rapidly evolving fraud environments.

To further enhance predictive performance, recent studies have explored ensemble and stacking frameworks that integrate multiple machine learning models into a unified architecture. For instance, a 2025 stacking-based approach combining XGBoost, LightGBM, and CatBoost achieved near-perfect classification performance while incorporating explainability mechanisms such as SHAP and LIME . This line of research reflects a broader trend toward hybrid modeling, where complementary strengths of different algorithms are leveraged to improve generalization. However, while such ensemble methods reduce variance and improve robustness, they introduce significant computational overhead and increase system complexity, making real-time deployment more challenging in high-frequency financial systems [8][9].

Another important direction in recent work focuses on addressing the class imbalance problem, which is inherent in fraud detection datasets where fraudulent transactions represent only a small fraction of the total data. Studies have shown that techniques such as Synthetic Minority Oversampling Technique (SMOTE), cost-sensitive learning, and class-weight adjustments can significantly improve the detection of minority class instances without severely degrading overall performance . In addition, frameworks such as FraudX AI integrate imbalance-aware learning with explainable AI techniques, demonstrating that it is possible to achieve both high recall and interpretability in fraud detection systems . However, these approaches may introduce risks of overfitting or inflated performance metrics when synthetic data is not carefully controlled [10].

More recently, researchers have begun exploring hybrid and optimization-driven machine learning frameworks to enhance scalability and efficiency. For example, studies combining traditional ML models with optimization techniques or emerging paradigms (such as quantum-inspired learning or feature optimization strategies) show promising improvements in performance and computational efficiency. A 2025 experimental study highlighted that Random Forest models can still achieve high accuracy (above 97%) when supported by robust feature engineering pipelines, outperforming more complex alternatives in certain scenarios . At the same time, scalable ML architectures designed for high-volume transaction processing emphasize the need for balancing model complexity with real-time constraints . Overall, while machine learning continues to form the backbone of fraud detection systems, recent advances indicate a shift toward more integrated, interpretable, and scalable solutions that can better align with operational and regulatory requirements [9].

Model	**Strengths**	**Weaknesses**	**Best Use Case**
Logistic Regression	Simple, interpretable	Cannot capture complex patterns	Baseline models
Decision Tree	Easy to understand	Overfitting issues	Rule extraction
Random Forest	Robust, stable	Moderate recall	General fraud detection
XGBoost	High accuracy, scalable	Less interpretable	High-risk fraud detection
LightGBM	Fast training	Sensitive to parameters	Large datasets

CatBoost	Handles categorical data well	Computational overhead	Mixed data environments
Ensemble (Stacking)	Highest accuracy	Complex deployment	Advanced fraud systems
Hybrid (XGBoost + GA)	Optimized performance	High cost	Research systems

3. Deep Learning Model

Recent developments in fraud detection have increasingly leveraged deep learning models to overcome the limitations of traditional machine learning approaches, particularly in capturing complex, non-linear, and high-dimensional relationships within transactional data. Deep Neural Networks (DNNs) have been widely adopted as a baseline deep learning architecture due to their ability to automatically learn feature representations without extensive manual engineering. A number of recent studies demonstrate that DNN-based models can outperform classical machine learning methods when trained on properly preprocessed datasets with imbalance handling techniques. For instance, Wu et al. (2025) proposed a continuous-coupled neural network architecture that significantly improved detection performance on benchmark credit card datasets, highlighting the effectiveness of deep non-linear modeling in fraud detection scenarios. However, while DNNs provide improved accuracy, they often fail to capture temporal dependencies inherent in transaction sequences, which limits their effectiveness in detecting behavior-driven fraud patterns [11].

To address the temporal nature of financial transactions, recent research has focused on sequence-based models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. These architectures are designed to model sequential dependencies by retaining historical context, making them particularly suitable for identifying anomalies in transaction sequences. Studies such as Benchaji et al. (2021) demonstrate that LSTM-based models, especially when combined with attention mechanisms, can effectively capture behavioral deviations and improve fraud detection performance. More recent evaluations continue to support the superiority of LSTM over GRU in certain datasets, particularly in scenarios where long-term dependencies are critical for identifying fraud patterns. Nevertheless, these models are sensitive to sequence design, including window size and temporal granularity, and often require significant computational resources for training and inference [12].

Another notable direction in recent work involves hybrid deep learning architectures that combine multiple neural network components to leverage their complementary strengths. For example, CNN–LSTM models integrate convolutional layers for local feature extraction with recurrent layers for temporal modeling, enabling a more comprehensive representation of transaction data. Similarly, frameworks such as UAAD-FDNet (2023) utilize autoencoders combined with attention mechanisms to learn normal transaction patterns and identify deviations as potential fraud. These hybrid approaches have demonstrated improved performance compared to standalone models, particularly in highly imbalanced and noisy datasets. However, the increased architectural complexity introduces challenges related to training stability, parameter tuning, and deployment in real-time systems [11].

In addition to supervised deep learning approaches, recent studies have also explored unsupervised and semi-supervised techniques, particularly autoencoders, for fraud detection in scenarios with limited labeled data. Autoencoders learn compact representations of normal transaction behavior and detect anomalies based on reconstruction errors. Research by Du et al. (2023) and Ding et al. (2024) highlights the effectiveness of combining autoencoder-based feature learning with tree-based classifiers such as LightGBM or XGBoost to improve detection accuracy. These hybrid pipelines are especially useful in practical settings where fraud labels are delayed or incomplete. Despite these advancements, deep learning models still face key challenges, including lack of interpretability, high computational cost, and difficulty in adapting to rapidly evolving fraud patterns. These limitations have motivated the exploration of more adaptive and autonomous approaches, such as Agentic AI, which aim to extend beyond prediction toward intelligent decision-making [11][12].

Model	**Key Characteristics**	**Strengths**	**Limitations**	**Suitable Use Case**
Deep Neural Network (DNN)	Fully connected multi-layer architecture	Learns complex non-linear relationships; minimal feature engineering	Cannot capture temporal dependencies; prone to overfitting	Tabular transaction data with static features
Convolutional Neural Network (CNN)	Uses convolution filters for feature extraction	Effective in capturing local feature patterns; reduces dimensionality	Limited capability for sequential/temporal modeling	Spatial or pattern-based fraud detection
Long Short-Term Memory (LSTM)	Recurrent network with memory cells	Captures long-term dependencies; strong for sequential data	High computational cost; slower training	Sequential transaction analysis, behavioral fraud
Gated Recurrent Unit (GRU)	Simplified version of LSTM	Faster training; fewer parameters	Slightly less expressive than LSTM	Real-time fraud detection systems
CNN–LSTM Hybrid	Combines CNN and LSTM	Captures both spatial and temporal features; high accuracy	Complex architecture; difficult tuning	Complex fraud patterns with sequence + feature interactions
Autoencoder (AE)	Unsupervised reconstruction-based model	Detects anomalies without labeled	Sensitive to threshold selection; false positives	Anomaly detection in

		data; useful for imbalance		unlabeled datasets
AE + Attention	Enhances AE with focus mechanism	Improves anomaly localization; better feature representation	Increased complexity; interpretability issues	Advanced anomaly detection
Hybrid DL + ML (AE + XGBoost)	Combines DL feature extraction with ML classifier	Improved accuracy and robustness	Multi-stage pipeline; higher latency	Practical deployment with mixed data

5. Explainable AI-Based Models in Fraud Detection

The increasing adoption of artificial intelligence in financial fraud detection has brought significant improvements in predictive performance; however, it has also introduced challenges related to transparency and trust. Financial institutions operate in highly regulated environments where decisions—especially those affecting customers, such as transaction blocking or account suspension—must be explainable and auditable. As a result, explainable AI (XAI) has emerged as a critical component in modern fraud detection systems. Rather than treating models as black boxes, XAI techniques aim to provide insights into how and why a particular decision is made, thereby enabling human analysts to validate and trust automated predictions [13].

Recent research has focused on integrating post-hoc explanation techniques with high-performing machine learning and deep learning models. Among these, SHAP (Shapley Additive Explanations) has gained widespread adoption due to its strong theoretical foundation and ability to provide both global and local interpretability. Studies conducted in 2024–2025 demonstrate that SHAP can effectively highlight key features influencing fraud predictions, such as transaction amount, frequency, and geolocation patterns, thereby assisting analysts in understanding model behavior. Similarly, LIME (Local Interpretable Model-agnostic Explanations) has been applied to generate instance-level explanations by approximating complex models with simpler surrogate models. While these approaches improve interpretability, they are often computationally intensive and may produce inconsistent explanations under slight input variations.

In the context of deep learning, attention mechanisms have been increasingly explored as an intrinsic form of explainability. Unlike post-hoc methods, attention layers provide insights during the model's decision-making process by assigning weights to different parts of the input sequence. For fraud detection, this enables identification of critical transactions or time steps that contribute most to a suspicious pattern. Recent hybrid models combining LSTM with attention layers have shown that such mechanisms not only enhance model performance but also improve interpretability by revealing temporal dependencies in fraudulent behavior. However, it is important to note that attention weights do not always guarantee causal explanations, and their interpretation must be handled with caution [14].

Despite these advancements, the integration of XAI into fraud detection systems remains an evolving area. One of the key challenges is balancing model complexity with interpretability, as highly accurate models are often less transparent. Furthermore, explainability methods must be scalable and efficient to operate in real-time financial systems. Recent frameworks attempt to address this by combining explainable models with ensemble learning and imbalance-aware techniques, ensuring that both performance and transparency are maintained. As regulatory requirements continue to evolve, the role of XAI is expected to expand, particularly in supporting compliance, auditability, and human-in-the-loop decision-making processes. This makes explainability not just a desirable feature, but a fundamental requirement for next-generation fraud detection systems.

5.1 Comparative Analysis of Explainable AI Techniques

Method	**Type**	**Key Characteristics**	**Strengths**	**Limitations**	**Suitable Use Case**
SHAP	Post-hoc	Based on game theory (Shapley values)	Consistent and reliable explanations; global + local insights	High computational cost	Regulatory environments requiring strong interpretability
LIME	Post-hoc	Local surrogate model approximation	Simple and intuitive explanations	Instability; varies with sampling	Instance-level explanation for analysts
Feature Importance (Tree-based)	Intrinsic	Derived from model training	Fast and easy to compute	Limited to global insights	Quick feature relevance analysis
Attention Mechanism	Intrinsic (DL)	Assigns weights to inputs	Captures temporal importance; improves performance	Not always causally interpretable	Sequential fraud detection models
Rule-based Explanation	Intrinsic	Uses decision rules	Highly interpretable	Low flexibility and scalability	Compliance-heavy systems
Hybrid XAI (SHAP + ML/DL)	Combined	Integrates model + explanation layer	Balanced accuracy and interpretability	Increased complexity	Advanced fraud detection systems

Agentic AI introduces a significant advancement over traditional machine learning, deep learning, and explainable AI models by enabling systems to move beyond passive prediction toward autonomous and adaptive decision-making. While machine learning models are effective at identifying patterns,

they typically rely on static training and require manual intervention to update or respond to new fraud strategies. Agentic AI addresses this limitation by incorporating continuous learning and feedback loops, allowing the system to dynamically adapt to evolving fraud behaviors without frequent retraining [15].

In comparison to deep learning models, which excel at capturing complex and sequential patterns but often operate as opaque systems, Agentic AI enhances both interpretability and responsiveness. It integrates reasoning capabilities that allow models not only to detect anomalies but also to evaluate context, prioritize risks, and initiate appropriate actions such as transaction blocking or multi-factor authentication. This reduces response time and improves operational efficiency in real-time environments.

Furthermore, although explainable AI improves transparency, it is generally limited to interpreting model outputs rather than influencing decisions. Agentic AI extends this by combining explainability with goal-driven behavior, enabling systems to justify actions while continuously optimizing outcomes. Overall, Agentic AI offers a more holistic framework that unifies prediction, explanation, and action, making fraud detection systems more intelligent, autonomous, and resilient to emerging threats.

9. Conclusion

The rapid evolution of digital financial systems has significantly increased both the scale and complexity of fraudulent activities, necessitating more advanced and adaptive detection mechanisms. This survey has examined recent developments in fraud detection across machine learning, deep learning, generative AI, and explainable AI paradigms. Machine learning models, particularly ensemble-based approaches, continue to provide strong baseline performance due to their efficiency and scalability. Deep learning techniques further enhance detection capabilities by capturing complex and sequential transaction patterns, although they often introduce challenges related to interpretability and computational overhead. Generative AI has emerged as a promising direction for addressing data imbalance and improving anomaly detection, but its practical deployment is still evolving.

The study also highlights the growing importance of explainable AI in ensuring transparency, trust, and regulatory compliance in financial systems. Despite these advancements, current approaches largely operate as static or semi-adaptive systems, lacking the ability to autonomously respond to dynamic fraud scenarios. In this context, Agentic AI offers a transformative shift by integrating learning, reasoning, and decision-making into a unified framework. By enabling continuous adaptation, real-time response, and goal-driven behavior, Agentic AI has the potential to significantly improve the robustness and effectiveness of fraud detection systems.

References (IEEE Format)

[1] J. Liu, H. Wang, and X. Zhang, “Comparative analysis of CatBoost, XGBoost and LightGBM for financial fraud detection,” *Applied and Computational Engineering*, vol. 15, no. 2, pp. 45–56, EWA Publishing, 2025.

[2] F. Almalki, A. Alharbi, and M. Alshammari, “A stacking ensemble approach with explainable AI for fraud detection,” *arXiv preprint arXiv:2505.10050*, pp. 1–12, arXiv, 2025.

[3] R. Singh and P. Kumar, “Fraud detection in imbalanced datasets using machine learning techniques,” *Proceedings of BUIRC Conference*, vol. 8, no. 1, pp. 120–128, Atlantis Press, 2025.

[4] F. Zhang, “Credit card fraud detection based on XGBoost and SMOTE,” *BCP Business & Management*, vol. 32, no. 1, pp. 78–85, BCP Publishing, 2024.

[5] A. Brown and T. Wilson, “Benchmarking machine learning and deep learning models for fraud detection,” *Decision Support Systems*, vol. 185, no. 3, pp. 113–125, Elsevier, 2025.

[6] Y. Tian, “A hybrid XGBoost–LSTM model for financial fraud detection,” *Advances in Economics, Management and Political Sciences*, vol. 45, no. 2, pp. 210–218, EWA Publishing, 2025.

[7] M. Wu, L. Chen, and Y. Zhao, “Continuous coupled neural networks for credit card fraud detection,” *IEEE Access*, vol. 13, no. 1, pp. 45678–45689, IEEE, 2025.

[8] I. Benchaji, S. Douzi, and B. El Ouahidi, “Credit card fraud detection model based on LSTM recurrent neural networks,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–17, Springer, 2021.

[9] J. Du, X. Li, and H. Sun, “Anomaly detection in financial transactions using autoencoders,” *Computers & Security*, vol. 120, no. 2, pp. 102–115, Elsevier, 2023.

[10] Y. Ding and Z. He, “Hybrid fraud detection using autoencoder and LightGBM,” *Expert Systems with Applications*, vol. 210, no. 4, pp. 118–130, Elsevier, 2024.

[11] K. Patel and R. Shah, “FraudX AI: Explainable fraud detection using SHAP and ensemble learning,” *Computers*, vol. 14, no. 4, pp. 120–135, MDPI, 2025.

[12] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, NeurIPS, 2017.

[13] M. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” *Proceedings of the ACM SIGKDD*, pp. 1135–1144, ACM, 2016.

[14] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the ACM SIGKDD*, pp. 785–794, ACM, 2016.

[15] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, NeurIPS, 2017.