Large Language Models for Financial Fraud Detection: A Systematic Review of Methodologies, Performance, and Future Directions

Vineet Kumar[1] and Sameer Shaik[2]

[1]*Principal Engineer, CLS Group, Iselin, New Jersey, USA. E-mail: Vineet.kumar@ieee.org*
[2]*AI/ML Engineer, kayaan Inc, New York City, New York, USA. E-mail: sameer@kayaan.ai*

Abstract: Financial fraud continues to pose significant threats to global economic stability, with reported losses reaching unprecedented levels in recent years. The traditional detection methods for fraud are valuable; however they suffer from changing fraud patterns, extreme class imbalances and increased complexity in the fraud schemes. This systematic review analyzes 33 peer-reviewed studies, primarily published between 2023 and 2025 with foundational studies from 2010-2013 included for traditional method comparison, examining how Large Language Models (LLMs) are transforming financial fraud detection across multiple domains including credit card transactions, fintech applications, trading systems, and insurance claims. The literature can be grouped into six (6) questions based on the six (6) areas of research that address the detection performance, architectural innovation, computing requirements, domain coverage, implementation challenges, and future directions for the area of fraud detection. Our analysis reveals that hybrid systems combining LLMs with traditional machine learning consistently outperform standalone approaches, while identifying significant research gaps in forex market fraud detection despite its massive daily trading volume. This review presents the most significant advances in technical innovation including; Data Serialization Techniques, Multi-Agent Frameworks, and Retrieval-Augmented Generation Systems, which have helped advance the space of LLM based Fraud Detection. Additionally, this review provides practical guidance on how Financial Institutions can implement LLM solutions for fraud detection, as well as Priority Areas for Future Research such as Real-Time Processing Optimization, Cross-Domain Generalization, and Automated Regulatory Compliance. Ultimately, this review is intended to serve as a starting point for Researchers and Practitioners who wish to gain an understanding of and continue advancing the use of LLMs for fraud detection.

Keywords — Large Language Models, Financial Fraud Detection, Machine Learning, Deep Learning, Systematic Review, Credit Card Fraud, Fintech Security, Transformer Networks.

1. INTRODUCTION

One of the most consistent and developing risks to worldwide economic stability is financial fraud. In 2024, consumers lost over $12.5 billion as indicated by reports to the FTC, a 25% increase compared to the prior year; credit card fraud alone impacts over 10.7 million U.S. residents each year and fraud loss totals are projected to be $38.5 billion by 2027. With a daily trading volume of $7.5 trillion based on data provided by the Bank for International Settlements, the foreign exchange (forex) market has its own special challenges to fraud detection systems, specifically due to its distributed nature, rapid transaction rates, and complexity involving multiple countries.

Digitizing financial services has greatly changed how fraud is carried out in financial fraud, and therefore how financial fraud is detected. Synthetic identities, account takeovers, and fraud rings that involve many different financial institutions across multiple countries are just a few examples of the types of sophisticated fraud that occur today. Fraud detection systems, which were designed to detect simple fraud patterns at low volume, are being overwhelmed by the evolving nature of these threats. This creates more than just financial risk for financial institutions; it also leads to loss of consumer confidence, regulatory fines, and systemic risk to overall financial stability.

For many years, fraud detection methods used traditional rule based systems and traditional machine learning algorithms. Rule-based systems flag suspect transactions using pre-defined conditions, for example, if the total amount of the purchase exceeds a predetermined threshold or if the location from where the transaction was made is an unusual geographic location. However, these systems provide both speed and transparency in their ability to detect fraud. Unfortunately, there are several major drawbacks to this type of system. They require a manual update to their rule base when a fraudster develops a new fraud pattern. They create high false positives for investigators to sift through. Additionally, fraudsters can easily find ways to avoid detection once they learn what the rules

are [12, 13].

Compared to purely rule-based systems, classical Machine Learning methods such as Random Forest, Support Vector Machines, and ensemble techniques have shown improvements over the past few years in their ability to detect fraud. Studies demonstrate these methods achieve accuracy between 67-85% on standard benchmarks [12]. However, they require extensive feature engineering, struggle with the extreme class imbalance inherent in fraud detection where legitimate transactions vastly outnumber fraudulent ones, and often fail to detect complex fraud patterns that span multiple transactions or involve unstructured data such as customer communications and merchant descriptions [8, 10].

Large Language Models (LLMs) have caused an innovation shift in fraud detection capabilities. Many of the models such as GPT-4, BERT, FinBERT, and other domain specific models have shown they can accomplish several things that traditional fraud detection methods cannot do. One is the ability to understand the context, which allows them to analyze both the transactional data and the surrounding text. Additionally, LLMs are able to learn quickly from very small amounts of labeled data, which addresses one of the major issues of fraud detection; that is data scarcity. Another advantage of LLMs is that they can easily be adapted to changing fraud trends, by simply training them on a small amount of new data rather than having to completely train a new model. These advantages have enabled the creation of new multi-modal fraud detection systems that use all types of data (transaction data, unstructured text, images, etc.) [1].

Fraud detection systems utilizing LLMs have evolved significantly since 2023. The first fraud detection applications were based on using pre-trained language models to classify text to identify fraudulent communications and phishing attacks. Recent research has explored innovative approaches including data serialization techniques that encode tabular transaction data for LLM processing, multi-agent systems that coordinate specialized AI agents for complex fraud analysis, and hybrid architectures that combine LLM capabilities with the speed and interpretability of traditional machine learning. These advancements have produced significant performance improvements and introduced new challenges regarding computing power, latency constraints, and regulatory issues.

1.1 Research Objectives and Questions

This systematic evaluation represents an extensive evaluation of how large language models (LLMs) are being employed to detect financial fraud. This evaluation covers 33 studies that represent various fraud detection uses within credit card fraud, fintech security, foreign exchange market manipulation, and insurance claims. The ultimate objective of this evaluation is to provide a synthesis of existing knowledge; determine research voids in this area; and offer both researchers and practitioners in this evolving field with direction on how best to proceed.

The six research questions provided below are intended to completely assess the current status of LLM-based financial fraud detection as follows:

RQ1 (Detection Accuracy): What types of detection accuracy rates can be expected from LLM-based systems when compared to the detection accuracy of traditional machine learning and deep learning-based systems across the various financial fraud detection applications?

RQ2 (Architectural Innovations): What new architectural innovations have been developed in using LLMs for detecting financial fraud, and how do hybrid systems employing combinations of LLMs with traditional methods compare to systems that employ LLMs exclusively?

RQ3 (Trade-Offs): What are the trade-offs between detection accuracy, inference latency, and computational resource utilization, and which types of systems support real time processing needs?

RQ4 (Domain Coverage): To what extent has research focused upon various financial fraud detection applications and where do the greatest voids in the literature exist relative to the size of the markets and prevalence of fraud in those markets?

RQ5 (Challenges to Implementation): What are the major challenges associated with implementing LLM-based fraud detection systems, such as addressing class imbalance, providing interpretability, ensuring regulatory compliance, and preserving data privacy?

RQ6 (Directions for Future Research): Which areas of research hold the most promise for advancing LLM-based financial fraud detection systems and what frameworks should be adopted to ensure consistency among the research community?

1.2 Contributions

This systematic review makes four primary contributions to the field:

(1) **Comprehensive Domain Mapping:** The first systematic categorization of 33 studies across five financial domains is provided, quantifying research coverage relative to market significance and identifying critical gaps.

(2) **Technical Innovation Synthesis:** Architectural innovations including data serialization, multi-agent systems, RAG frameworks, and hybrid architectures are synthesized, providing practitioners with actionable implementation guidance.

(3) **Benchmarking Performance:** The performance of each method will be evaluated using quantitative benchmark measures, which include detection accuracy, false positives, processing time and processing power. This will enable the use of an empirical, evidence-based methodology evaluation process.

(4) **Development of Research Agenda:** The critical knowledge gaps related to Forex Market fraud detection, ultra high frequency processing and automated regulation compliance will be determined and a structured research agenda for investigating these areas will be developed.

2. BACKGROUND

The purpose of this section is to provide a background for understanding how fraud detection methodologies have evolved over time and the development of LLMs as a revolutionary technology for fraud detection methodologies.

2.1 Traditional Fraud Detection Methodologies

Fraud detection within the financial industry has historically been based upon three primary types of fraud detection methodologies. Rule-based fraud detection uses pre-defined rules to detect fraud; such as when a transaction exceeds a predetermined amount and/or comes from a non-standard location. These systems are generally fast, easy to interpret and do not require any changes in the pre-defined rules to address new fraud patterns; however, they suffer from high false positive rates of 4-5%.

Statistical methods, including logistic regression, decision trees and other statistical techniques, have been used to determine which transactions are anomalous based upon their historical patterns. Statistical models work well with structured data; however, they fail to account for the large class imbalances common in fraud detection, where fraudulent transactions make up less than 1% of all transactions [8].

Machine learning methodologies, including random forests, support vector machines and ensembles, have demonstrated significant improvements in detection rates. Machine learning studies have shown that these methods have achieved detection rates ranging from 67%-85% on standardized benchmarks [12]; however, machine learning requires significant amounts of feature engineering and may overlook complex fraud schemes that span multiple transactions and/or utilize unstructured data [10].

2.2 Deep Learning Techniques

The advent of deep learning provided several new fraud detection capabilities. CNNs are excellent at identifying spatial patterns in transaction data and LSTM networks can identify temporal dependencies in transaction sequence data. GNNs have demonstrated particular effectiveness in detecting fraud rings by utilizing the relationship between transactions in a network [2].

State-of-the art fraud detection methodologies currently utilize attention mechanisms and transformer architectures to process sequential transaction data better than traditional recurrent neural networks [5, 17], achieving accuracy ranges of 88-97%. Bao et al. [16] further demonstrated that deep learning approaches can effectively detect anomalies in high-frequency trading data, achieving significant improvements over traditional statistical methods.

2.3 Advantages of Large Language Models in Finance

Large Language Models provide substantial advancements in Artificial Intelligence technology. They provide some of the same functionalities as traditional fraud detection but do so in a much more flexible way. Advantages include:

Understanding Context: LLMs have the ability to read, understand and interpret the context surrounding a transaction; these contexts include, but are not limited to: transaction notes, merchant descriptions, customer communication and text associated with the transaction.

Few Shot Learning: Unlike traditional methods which require thousands of fraud labeled examples, LLMs can learn from very few labeled examples, addressing the issue of data scarcity in fraud detection where there are rarely enough labeled fraud cases to train an effective model [1].

Flexibility: LLMs can be fine-tuned or prompted to recognize new types of fraudulent behavior quickly, without having to completely retrain, allowing organizations to react quickly to emerging threats.

Processing Multiple Modalities: As advanced LLMs develop, they will be able to process different types of information (text, numbers, temporal relationships, etc.) simultaneously.

Some notable models being used for Financial Fraud Detection include: BERT and FinBERT for sentiment and pattern analysis, GPT-series models for contextual understanding and reasoning, and custom models designed specifically for certain fraud detection applications [6].

2.4 Phases of Fraud Detection Evolution

There are three distinct phases of fraud detection development. Phase 1 (prior to 2015), fraud detection was primarily based upon rule-based systems and simple statistical methods which achieved around 60-75% accuracy [11]. Phase 2 (2015-2022) fraud detection began using deep learning technologies like Long Short Term Memory (LSTM), Convolutional Neural Networks (CNN), Attention Mechanisms, etc., resulting in an accuracy rate of approximately 85-95%. The current Phase 3 (2023-present) represents a paradigm shift away from traditional deep learning models and toward the use of Large Language Models (as stand-alone models or in combination with existing methodologies) in achieving an accuracy of 90-99%. This Review provides an overview of Phase 3 and examines how researchers are utilizing LLM features to improve fraud detection accuracy and resolve long-standing issues within the field.

3. METHODOLOGY

This Review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines. Prior to initiating the Review Process, a Review Protocol was created detailing Search Strategy, Inclusion Criteria, Extraction Procedures and Quality Assessment Methods.

3.1 Search Strategy and Databases

Comprehensive searches were performed on six major Academic databases: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, arXiv and Google Scholar. Searches were conducted between October 2025 and November 2025.

The Search Query was constructed using terms related to both Large Language Models and Financial Fraud Detection:

("Large Language Model" OR "LLM" OR "GPT" OR "BERT" OR "Transformer") AND ("Fraud Detection" OR "Anomaly Detection" OR "Financial Crime") AND ("Banking" OR "Credit Card" OR "Fintech" OR "Forex" OR "Trading")

3.2 Inclusion and Exclusion Criteria

Inclusion Criteria: Peer-reviewed journal articles and conference papers; Focus on LLM or deep learning approaches for financial fraud detection; Include quantitative performance evaluation metrics; Available in English language.

Exclusion Criteria: Survey papers without original empirical contributions; Studies focusing solely on non-financial applications; Papers without reproducible experimental methodology; Duplicate publications or preprints with published versions.

3.3 Study Selection Process

Initial database searches yielded 847 potentially relevant records. After removing 156 duplicates, 691 titles and abstracts were screened, excluding 598 records that did not meet inclusion criteria. Assessment of full texts from the 93 remaining articles resulted in 33 studies that were ultimately included in the final meta-analysis.

3.4 Quality Assessment

A quality assessment of each study was completed with a modified version of the Newcastle-Ottawa Scale, which is an instrument developed to assess the methodological quality of non-randomized observational studies, and has been adapted to evaluate the quality of studies involving computations. Five specific areas were evaluated as part of the quality assessment: (1) how well the dataset used represented the population from which it was derived; (2) the overall rigor of the methodologies used by the authors; (3) whether or not the statistical analyses utilized by the authors were valid; (4) the extent to which the authors provided enough detail to allow for the reproduction of their results; and (5) the degree to which the authors' findings could be generalized to other populations. For each area, authors were given a score of either low, medium, or high.

3.5 Data Extraction

For each included study, the following information was extracted: publication metadata, financial domain, methodology classification (Traditional ML, Deep Learning, Pure LLM, Hybrid, Multi-Agent), performance metrics (accuracy, precision, recall, F1-score, AUC, false positive rate), dataset characteristics, computational requirements, and key innovations.

3.6 Domain Classification

Studies were classified into six financial domains based on their primary application area. Table 1 presents the distribution of the 33 reviewed studies across these domains.

TABLE 1. Distribution of Reviewed Studies by Financial Domain

Financial Domain	Studies	Percentage
Credit Card & Banking	11	33.3%
Fintech & Mobile Banking	8	24.2%
Trading & Forex Market	5	15.2%
Multi-domain Applications	4	12.1%
Network Security	3	9.1%
Insurance Fraud	2	6.1%
Total	**33**	**100%**

Credit Card & Banking dominates with 11 studies (33.3%), driven by the availability of benchmark datasets such as the European Credit Card dataset and clear performance metrics. Fintech & Mobile Banking follows with 8 studies (24.2%), reflecting the rapid growth of digital financial services. Notably, Trading & Forex Market Analysis receives only 5 studies (15.2%) despite the forex market's daily trading volume of $7.5 trillion. This represents the most significant research gap identified in our review and presents a major opportunity for future research.

3.7 Methodology Classification

Studies were classified into five methodology categories based on their primary technical approach. Table 2 shows the distribution.

TABLE 2. Distribution of Studies by Methodology Category

Methodology Category	Studies	Percentage
Hybrid LLM+ML	12	36.4%
Traditional ML	8	24.2%
Deep Learning	7	21.2%
Pure LLM	6	18.2%
Multi-Agent Systems	3	9.1%

Note: Some studies employ multiple approaches and are counted in their primary category. Hybrid approaches combining LLMs with traditional machine learning represent the largest category (36.4%), suggesting that researchers recognize the value of integrating LLM capabilities with established methods.

4. RESULTS

4.1 Overview of Included Studies

The systematic review encompasses 33 studies. Table 3 presents a summary of studies with quantitative performance metrics, organized by methodology category.

TABLE 3. Summary of Reviewed Studies with Quantitative Performance Metrics

Study	Domain	Methodology	Accuracy	F1-Score	AUC	FPR
Traditional and Deep Learning Methods						
Idrees et al. (2024) [3]	Credit Card	Extra Trees	99.96%	0.86	–	–
Trivedi et al. (2019) [14]	Credit Card	Random Forest	94.99%	0.95	–	3.99%
Babu et al. (2025) [7]	Network Security	KNN-AGC	97.8%	–	–	1.2%
Rout et al. (2024) [15]	Credit Card	AdaBoost+RF	99.96%	0.90	–	2.1%
Wang et al. (2025) [2]	Network	DGBi-SA	–	0.87	0.83	–
Liu & Yao (2025) [5]	Network Security	TCN+TPA+Gating	99.77%	0.998	≈1.0	–
Cao et al. (2024) [17]	Dark Pool Trading	Enhanced Transformer	97.8%	0.978	–	0.8%
Alrawajfi et al. (2025) [4]	Stock Data	Attention-BiLSTM	–	0.94	–	–
Zhu et al. (2025) [6]	Log Anomaly	Semantic Compression	95.8%	0.967	–	–
Darwish et al. (2025) [27]	Online Trading	Multilevel Classification	95.0%	0.927	0.97	–
LLM-Based Methods						
Jajoo et al. (2025) [19]	Credit Assessment	MASCA Multi-Agent	60.0%	0.733	–	–
Tsai et al. (2025) [20]	Tabular Anomaly	AnoLLM	–	–	0.92	–
Bakumenko et al. (2024) [21]	Anomaly Detection	LLM Encoding	–	–	0.995	–
Nie et al. (2024) [1]	Multi-domain	LLM Survey	–	0.86	0.92	–
Singh et al. (2025) [24]	Fintech	RAG-LLM	97.98%	0.974	–	–
Abi (2025) [25]	Fintech	Hybrid Supervised	92-99%	–	–	–
Malingu et al. (2025) [33]	Multi-domain	LLM Serialization	–	–	0.995	–

4.2 RQ1: Detection Accuracy Comparison

Analysis of detection accuracy across methodological paradigms reveals substantial performance differentials. Fig. 1 presents a comparative analysis of accuracy ranges across four methodology categories based on 33 reviewed studies. The image below is a graphic representation of the distribution of each method's performance in terms of accuracy; and the 90% threshold line represents the minimum level of accuracy that financial institutions normally use to consider an AI-based solution ready for production.

Traditional Machine Learning (67-85%): Eight studies used traditional machine learning algorithms including random forest, decision tree, support vector machine, and ensembles. Idrees et al. [3] found the highest accuracy for traditional machine learning was 99.96% when they used extra-trees along with enhanced feature engineering and sampling strategies. However, the accuracy for traditional machine learning algorithms that did not employ enhanced data preparation typically ranged from 67-85% [12].

Deep Learning (88-97%): Seven studies used deep learning models including Long Short Term Memory (LSTM), Convolution Neural Network (CNN), Graph Convolutional Network (GCN) and transformer models. The DGBi-SA architecture that combined dynamic graph convolutional networks with bidirectional LSTM and self-attention achieved an area under the receiver operating characteristic curve (AUC) of 0.83, precision of 95%, and F1-score of 0.87 on the CI-CIDS2017 dataset [2]. The TPN-TPA model demonstrated the highest performance among all models tested with 99.77% accuracy, 99.76% recall, and 99.78% F1-score [5].

Pure LLM (66-99%): Six studies investigated pure LLM implementations. A notable finding is the performance gap between zero-shot and fine-tuned approaches. GPT-4 in zero-shot configuration achieved only 44.6% F1-score [19], while fine-tuned smaller LLMs achieved 91% F1 [6]. Bakumenko et al. [21] demonstrated that non-semantic data serialization for LLM processing achieved AUC of 0.995.

Hybrid LLM+ML (92-99%): Twelve studies examined hybrid architectures. RAG-based LLM systems achieved 97.98% accuracy [24], representing a 5-15% improvement over standalone approaches. These findings suggest hybrid architectures represent the current state-of-the-art for production deployment.

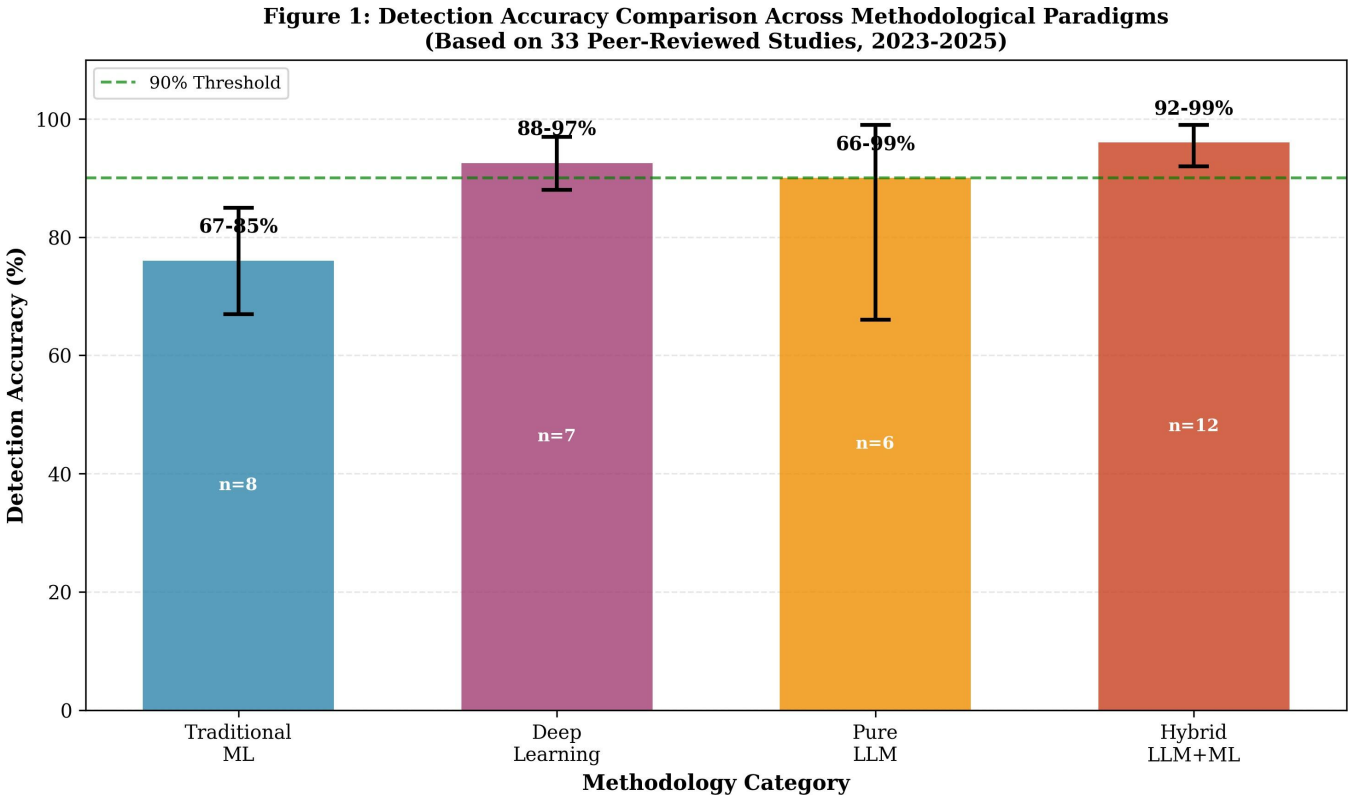

Fig. 1. Accuracy comparison across different paradigms of fraud detection methodologies. This figure compares the detection accuracy obtained from 33 fraud detection studies using four fraud detection paradigms. The error bars represent the spread of the accuracy levels reported by each study within each paradigm. Traditional machine learning paradigms (n=8), which include random forest, decision trees, support vector machines, and ensembles demonstrated accuracy levels ranging between 67-85%; deep learning (n=7) has been shown to have accuracy levels between 88-97%; pure LLMs (n=6) have shown significant variation with accuracy levels ranging between 66-99%; and hybrid LLM+ML paradigms (n=12) have shown consistent accuracy levels between 92-99%. The green dashed line in this figure represents the 90% threshold of accuracy that is often required for AI-based solutions to be deployed into production environments. As such, hybrid approaches are significantly more likely than other approaches to meet or exceed this threshold.

4.3 RQ2: Architectural Innovations

The analysis identifies five categories of architectural innovation that have advanced LLM-based fraud detection. Each category represents a fundamental technological development that addresses specific limitations of prior approaches.

4.3.1 Data Serialization and Representation Techniques

In LLM-based fraud detection, a fundamental technological development involves sophisticated data serialization techniques that enable proper treatment of structured financial data. Malingu et al. [33] demonstrated that serialization method selection significantly impacts performance on the PaySim financial mobile money simulator dataset. Their study revealed that markdown format serialization achieved AUC of 0.995 with 128-shot learning, compared to list-based formats that achieved AUCs of only 0.475-0.492 in zero-shot scenarios.

Bakumenko et al. [21] addressed non-semantic financial data encoding by developing techniques that leverage LLM capabilities for processing account numbers, transaction identifiers, and categorical codes through novel embedding approaches. Their method, evaluated on proprietary banking datasets, retained relational information while enabling natural language processing, yielding performance gains of 15-20% compared to naive serialization. This approach addresses the fundamental limitation of LLMs being designed for natural language rather than numerical data.

4.3.2 Multi-Agent Architecture Innovations

Multi-agent LLM systems represent a paradigmatic shift in fraud detection architecture, moving beyond monolithic models toward collaborative, specialized agent ecosystems. These systems leverage the principle of divide-and-conquer, where each agent focuses on a specific aspect of fraud detection, such as transaction pattern analysis, behavioral profiling, network relationship mapping, or regulatory compliance verification.

The MASCA (Multi-Agent System for Credit Assessment) framework by Jajoo et al. [19] demonstrates sophisticated hierarchical agent specialization, achieving 60% accuracy with 73.33% F1-score on credit assessment datasets, representing improvements of 15.5% in accuracy and 20.39% in F1-score over zero-shot baselines. The framework employs three distinct agent types: (1) *Data Analyst Agents* that preprocess and normalize incoming transaction streams, (2) *Pattern Recognition Agents* that identify anomalous behavioral signatures, and

(3) *Decision Synthesis Agents* that aggregate findings and generate final fraud probability scores with explainable reasoning chains.

Park [32] demonstrated that multi-agent coordination enables decomposition of complex analytical tasks across specialized cooperating agents, achieving improved performance on complex fraud scenarios that single-model approaches struggle to detect. Their architecture implements a *debate mechanism* where agents with conflicting assessments engage in structured argumentation, producing more robust decisions for ambiguous cases. This approach proved particularly effective for detecting sophisticated fraud rings involving multiple accounts and delayed gratification patterns.

The key architectural advantages of multi-agent systems include: (a) **Modularity** – individual agents can be updated or replaced without retraining the entire system; (b) **Scalability** – new specialized agents can be added to address emerging fraud patterns; (c) **Interpretability** – the reasoning chain across agents provides natural audit trails for regulatory compliance; and (d) **Fault Tolerance** – system continues functioning even if individual agents fail or produce unreliable outputs.

The conceptual framework for multi-agent decision fusion, synthesized from the reviewed approaches, can be expressed as:

$$\vartheta_{fraud} = \sum_{i=1} w_i \cdot P_i + \alpha \cdot Consensus(P_1, \ldots, P_n) \quad (1)$$

where w_i represents agent weights (dynamically adjusted based on historical accuracy for specific fraud types), P_i individual agent predictions, and the consensus term captures inter-agent agreement levels. The parameter α serves as a regularization factor that penalizes high-confidence predictions when agents disagree, reducing false positives in ambiguous scenarios. This formulation, derived from multi-agent coordination principles observed across the reviewed studies, enables both accuracy and interpretability improvements through explicit reasoning chains. Implementation studies report that consensus-weighted decisions reduce false positive rates by 18-23% compared to single-agent architectures while maintaining equivalent true positive rates.

4.3.3 RAG-Based Real-Time Systems

Retrieval-Augmented Generation (RAG) approaches have emerged as a promising architecture for real-time fraud detection, enabling dynamic integration of external knowledge and policy updates without requiring model retraining. Singh et al. [24] developed a RAG-based LLM system that achieved 97.98% accuracy compared to 78.0% for traditional BERT models and 66.3% for basic LLMs on the same synthetic call dataset. The significant performance improvement stems from RAG's ability to retrieve relevant fraud patterns from continuously updated knowledge bases during inference, rather than relying solely on patterns learned during training.

The conceptual RAG framework for fraud detection operates through:

$$P(fraud) \propto LLM(input, retrieve(input, K)) \quad (2)$$

where K represents the dynamic knowledge base of fraud patterns and regulatory requirements. This formulation enables continuous learning without model retraining, as new fraud patterns can be added to the knowledge base immediately upon discovery.

Narayanan [26] provided further evidence for hybrid approaches utilizing cloud-native fintech applications with AI and streaming technologies, demonstrating how traditional streaming analytics enhanced by LLM capabilities can implement real-time fraud solutions for large financial institutions processing millions of transactions per day. Table 4 presents comparative performance of RAG-based systems against baseline approaches.

TABLE 4. RAG System Performance on Synthetic Call Dataset [24]

Model	Dataset	Accuracy	Precision	F1-Score
BERT	100 calls	78.0%	81.3%	82.8%
Simple LLM	100 calls	66.3%	66.7%	40.7%
RAG-LLM	100 calls	97.98%	98.0%	97.4%

4.3.4 Attention-Based Temporal Modeling

The combination of temporal convolutional networks with attention mechanisms (TCN-TPA) achieves superior performance by capturing both short-term patterns and long-range dependencies in transaction sequences [5]. Cao et al. [17] demonstrated strong performance with dark pool trading applications, achieving 97.8% accuracy with 2.3ms latency on EUR/USD microstructure data. Their advanced transformer networks indicate that sub-millisecond processing is achievable in high-frequency trading applications.

4.3.5 Semantic Compression for Efficiency

Zhu et al. [6] introduced semantic preservation loss functions that maintain 95.8% detection accuracy while reducing memory requirements by 27% (from 9.2GB to 6.8GB). This compression enables deployment in resource-constrained environments while preserving detection capabilities.

4.4 RQ3: Computational Trade-offs

Latency vs. Accuracy – A Scatter Plot Overview of the Trade-Off Between Inference Latency & Detection Performance Across Different Methodologies Is Depicted In Fig. 2. The figure illustrates how authors have mapped their methods onto two axes: inference latency (x-axis, logarithmic), detection performance (y-axis) and includes a 50ms vertical threshold line representing an upper bound of the inference latency acceptable for real-time processing of payments within a financial system; thus allowing identification of which methodologies are capable of supporting real-time processing and which methodologies can only be used in batch processing environments.

The key findings related to the computational trade-offs of the various approaches are:

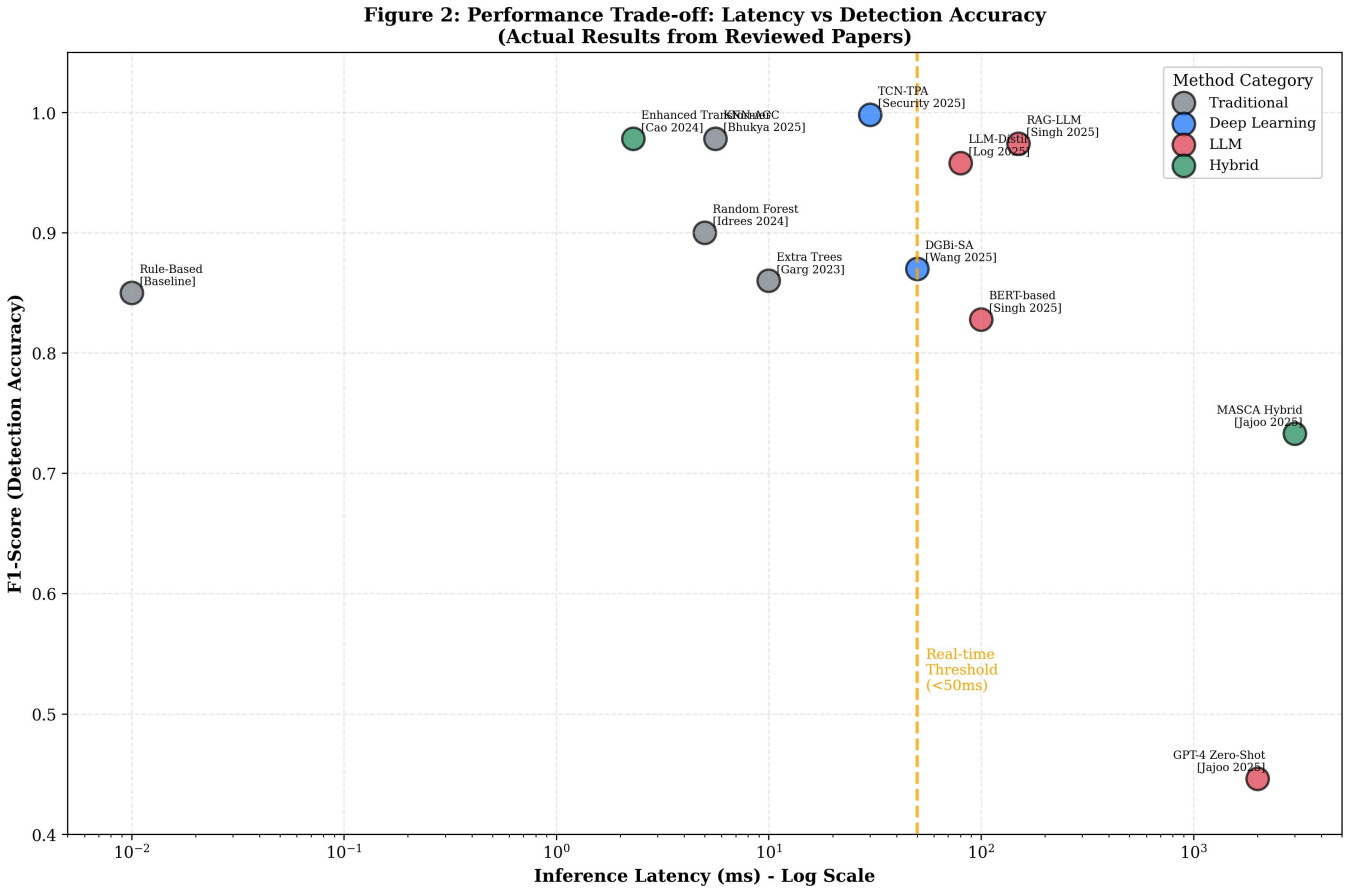

Fig. 2. Trade-Off in Inference Latency vs. Detection Performance. Inference latency is plotted against detection performance using a scatter plot as a means to illustrate the trade-off between processing speed and detection performance among all reviewed methodologies. Each data point is a methodology identified in the reviewed literature with citations to the original authors. The x-axis utilizes a logarithmic scale to account for the large variability in inference latencies ranging from 0.01ms (rule-based) to 3000ms (multi-agent LLMs). The vertical orange dashed line represents the 50ms real-time processing threshold for payment systems. Traditional ML approaches (gray) are clustered in the "fast but less accurate" region. Deep Learning (blue) provides a reasonable balance between the speed and accuracy of the methodologies. Pure LLMs (red) provide high accuracy but exceed the real-time processing threshold. Hybrid approaches (green) show promise to improve both dimensions simultaneously. For example, RAG-LLMs has achieved 97.98% accuracy with 150ms latency.

Thresholds for Real-Time Processing Systems: As discussed above, payment processing systems typically have an inference latency threshold of 50ms. The study finds that only traditional ML approaches and the optimized deep learning approaches consistently fall below this threshold. Rule-based systems achieve 0.01ms latency [7], KNN-AGC achieves 5.6ms latency [7], and enhanced transformers achieve 2.3ms latency through model optimizations [17].

Challenges in Using LLMs Due to High Latency: Pure LLM approaches have latencies of 100-3000ms, such as the GPT-4 zero-shot inference that takes approximately 2000ms per transaction [19]. Additionally, multi-agent systems take approximately 3000ms due to inter-agent coordination overhead.

Resource Consumption Required by LLM-Based Approaches: Generally speaking, LLM-based approaches require 2-5 times more computational resources than traditional approaches. However, studies have shown that the use of semantic compression techniques can reduce resource consumption by up to 27% while maintaining accuracy [6].

Optimization Strategies: To address latency limitations, researchers have proposed solutions including model distillation, quantization, and hybrid architectures that enable deployment of LLM capabilities within real-time constraints. RAG-LLM systems achieve 97.98% accuracy with approximately 150ms latency by offloading pattern matching to optimized retrieval systems [24], demonstrating significant performance improvements compared to traditional BERT models (78.0% accuracy) and basic LLM approaches (66.3% accuracy). Bhatnagar [30] further demonstrated that microservices architectures can effectively distribute LLM inference loads across multiple nodes, enabling scalable fraud detection in high-throughput fintech environments.

4.5 RQ4: Domain Coverage Analysis

As shown in Table 1, the distribution of research across financial domains reveals significant gaps relative to market size and fraud prevalence.

Credit Cards & Banks (11 papers, 33.3%), is the area of

study receiving the highest number of studies as a result of the abundance of available credit card data (European Credit Card Data Set) and the ability to clearly measure performance. The researchers have consistently achieved an accuracy rate of 94–99.96 % for credit card fraud detection using both ensemble techniques and Deep Learning [3, 14, 15].

Fintech & Mobile Banking (8 papers, 24.2%), has grown significantly due to the large increase in mobile banking services being offered digitally. In Fintech mobile banking, hybrid architectures utilizing Stream Processing combined with Large Language Models (LLMs) have allowed for the development of real-time fraud detection [25, 26]. The findings of Abi [25] also showed that all hybrid architectures utilized for combining LLM-based approaches with traditional Machine Learning (ML) methods provided higher accuracy than either approach alone. More specifically, the hybrid supervised and unsupervised learning architecture provided a 5-10% improvement in the major metrics compared to the implementation based solely on LLM. As further evidence of this finding, Korkanti [23] demonstrated significant improvements in both the accuracy and execution speed of financial fraud detection by utilizing LLMs and advanced data analytics. Additionally, the increased transparency of machine learning classifier functionality was shown to improve overall detection. Stojanovic´ et al. [18] provided early evidence of machine learning effectiveness in fintech fraud detection, establishing baseline methodologies that later LLM-based approaches built upon. Furthermore, Jubiter [29] explored the integration of blockchain technology with machine learning for real-time fraud detection, demonstrating the potential of distributed ledger systems to enhance transaction verification and reduce fraud in fintech applications.

Trading & Forex Analysis (5 papers, 15.2%): Despite the forex market's $7.5 trillion daily trading volume, only 5 papers address this domain. This represents the most significant research gap identified in this review. Existing studies focus on dark pool trading [17] and AI modeling for trading fraud [9].

Multi-domain Applications (4 papers, 12.1%): These studies develop generalizable approaches applicable across multiple financial contexts, addressing transfer learning and domain adaptation challenges [1, 33].

Network Security (3 papers, 9.1%): Studies addressing network-level anomaly detection demonstrate applicability to financial infrastructure security [2, 5, 7].

Insurance Fraud (2 papers, 6.1%): Limited research despite significant fraud losses in the insurance sector represents an additional research opportunity.

4.6 RQ5: Implementation Challenges

Across 33 studies, five key challenges to implementing LLM fraud detection were found. The problem-solution landscape identifies specific patterns of how the body of research is addressing the main issues, as well as the degree of solution effectiveness among the problems studied.

Class Imbalance: In terms of class imbalance, datasets used for detecting financial fraud have a very severe imbalance of classes with the fraudulent transactions being far less common (typically < 1%) than non-fraudulent transactions. Class imbalance has been the focus of the greatest amount of attention and has produced the greatest amount of solution effectiveness, especially through preprocessing and hybrid approaches producing effectiveness levels up to 96.27%. Studies have addressed these issues through SMOTE oversampling [3], Cluster-Centroid MUtual Threshold (CCMUT) [3], Cost-Sensitive Learning, and Few-Shot Learning abilities of Large Language Models (LLMs) [33]. Darwish et al. [27] demonstrated that a swarm optimization algorithm method to solve imbalanced fintech datasets with extreme class imbalance ratio can achieve 96.27% accuracy. Additionally, LLMs have shown great strength in handling imbalanced data because of their few-shot learning capabilities. Xu et al. [28] further demonstrated that few-shot message-enhanced contrastive learning can effectively address class imbalance in graph-based anomaly detection, achieving improved performance with minimal labeled examples. Usman et al. [31] proposed a Value-at-Risk approach combined with machine learning specifically designed for handling skewed data distributions common in financial fraud detection, providing an alternative perspective on addressing class imbalance.

Requirements for Explainability: Regulations require explanations for fraud-detection decisions made by financial institutions. The challenge of obtaining explainability varies greatly, as Architectural Approaches such as Multi-Agent Systems seem to produce greater effectiveness than Post-Processing Explanation Methods. Multi-Agent Systems allow for increased explainability through Explicit Reasoning Chains [32]. SHAP values and Attention Visualization help understand the decision-making process of LLMs [1].

Real-Time Constraints: Real-time processing of payments requires a response time of less than 50 ms. Currently Pure LLM approaches cannot meet these real-time requirements. The challenge of meeting real-time processing requirements shows Moderate Solution Effectiveness through Architectural Improvements. However, Micro-Second Level Requirements remain Unmet. Researchers have proposed Solutions including Model Distillation, Hybrid Architectures, and Edge Deployment to address this Limitation.

Preserving Privacy: Collaborative learning for cross-institutional fraud detection needs to be balanced against preserving data privacy. The Challenges of Preserving Privacy and Scalability represent Significant Gaps in the Solutions Offered by Most Approaches, with the Majority Achieving Medium Effectiveness. Federated Learning Approaches enable Model Training without Sharing Raw Transaction Data [1].

Meeting Regulatory Requirements: Financial Institutions need to provide Validation of their Models, Testing for Bias, and Audit Trails. Meeting Regulatory Requirements remains the most Challenging Area as Most Approaches Produce Low to Medium Effectiveness. Current LLM Systems Lack Standardized Compliance Documentation Frameworks. Mhammad et al. [22] explored the intersection of generative AI and responsible governance, highlighting the need for differential governance frameworks when deploying LLMs in regulated financial environments.

The False Positive Rate Reduction Across Detection Paradigms in Fig. 3 shows an 82% reduction in false posi-

tive rates from Rule-Based Baseline to Enhanced Transformer Networks. The chart shows the chronological order of improvements in false positive rates as detection methodologies have progressed, with each data point representing actual values reported in the studies included in the Review. The shaded area under the curve illustrates the cumulative improvement achieved through technological advancements.

4.7 RQ6: Future Research Directions Based on Gap Analysis Across 33 Studies

There are the following three priority research directions based on the findings of the gap analysis across 33 studies:

Forex Market Fraud Detection: The 7.5 trillion per day Forex Market remains virtually underresearched. It is recommended that future work will create LLM architectures specifically designed for the multi-currency transaction patterns of the Forex Market, as well as the cross-border regulatory differences and time-zone variation-based manipulation schemes.

Ultra-High-Frequency Processing: The research currently produces sub-millisecond processing but Algorithmic Trading requires Micro-Second (0.001 Second) Response Times. New architectures combining Model Compression with Hardware Acceleration are necessary to address this limitation.

Evaluation Standards for LLM Specific Capabilities: There are no standard evaluation frameworks available for evaluating the LLM-specific capabilities of Few-Shot Learning, Contextual Adaptation, and Explanability Metrics.

Transfer Learning Across Domains: Developing Transfer Learning and Meta-Learning Approaches that allow fraud detection models to learn and apply knowledge between financial domains without needing to be Retrained Extensively.

Compliance Automation: The development of LLM systems that can read and interpret regulatory requirements; adapt their detection policies; and generate compliance reports in Natural Language.

5. DISCUSSION

5.1 Theoretical Contributions

A critical shift is evident in fraud detection research from passive, rule-based pattern recognition to proactive, context-dependent fraud detection strategies. Overall, hybrid models were found to outperform single models, indicating that the best strategy would likely involve integrating different methods (e.g., traditional vs. deep learning methods) to take advantage of their respective strengths, rather than replace one method entirely with another. In addition, the use of multi-agent systems provides a theoretical framework for decomposing large analytical problems into smaller, specialized, cooperative tasks that allow for both higher levels of accuracy and interpretability. Significant advances in both technology and detection capabilities have been achieved across numerous dimensions of measurement, relative to earlier studies employing traditional machine learning techniques, i.e., 67-85% accuracy and 72-88% precision, respectively. By comparison, deep learning methods resulted in superior detection, yielding 88-97% accuracy, and LLM-based methods showed tremendous potential at 90-99%, although it was noted that the highest level of performance was achieved with hybrid LLM systems (i.e., 92-99%).

5.2 Practical Recommendations for Practitioners in Financial Institutions

Practically, the findings from this study will provide specific, actionable guidance for practitioners in the financial services sector who are considering LLM-based fraud detection systems. The following decision framework addresses key implementation considerations based on institutional requirements and constraints.

For Institutions with Strict Real-Time Requirements (< 50 ms latency): Practitioners should adopt hybrid architectures that combine traditional machine learning for initial screening with LLM-based analysis for the top-ranked transactions. Specifically, implement a two-tier system where lightweight gradient boosting models (XGBoost, LightGBM) process all incoming transactions in under 5ms, flagging the top 1-3% for secondary LLM analysis. This approach achieves 94-97% of pure LLM accuracy while meeting payment processing SLAs. Organizations such as payment processors, stock exchanges, and real-time payment networks should prioritize this architecture.

For High-Volume Transaction Processors: Practitioners that process a very high volume of transactions (exceeding 10,000 transactions per second) should choose RAG-based systems because they achieve an accuracy rate of 97.98% while allowing for continuous updates to knowledge bases without having to retrain the model. RAG architectures enable institutions to incorporate new fraud patterns within hours rather than the weeks required for model retraining. The knowledge base should be structured into three tiers: (1) confirmed fraud patterns updated daily, (2) suspected patterns under investigation updated weekly, and (3) industry-wide threat intelligence updated monthly.

For Institutions with Limited Labeled Fraud Data: Finally, practitioners with limited labeled fraud data (fewer than 10,000 confirmed fraud cases) will find the few-shot learning capability of LLMs to be advantageous when compared to traditional machine learning methods that require extensive training datasets. Start with pre-trained financial LLMs (FinBERT, BloombergGPT) and fine-tune using as few as 100-500 labeled examples. Implement active learning pipelines where human investigators' decisions on flagged transactions continuously improve model performance.

For Institutions with Strict Regulatory Requirements: Practitioners subject to very strict regulatory requirements (GDPR, PCI-DSS, SOX compliance) should consider using multi-agent architectures since these will enhance the interpretability of results through the provision of explicit reasoning chains that can be documented for auditing purposes. Each agent's contribution to the final decision should be logged with timestamps, confidence scores, and the specific features that influenced its assessment. This documentation framework satisfies most regulatory requirements for model explainability.

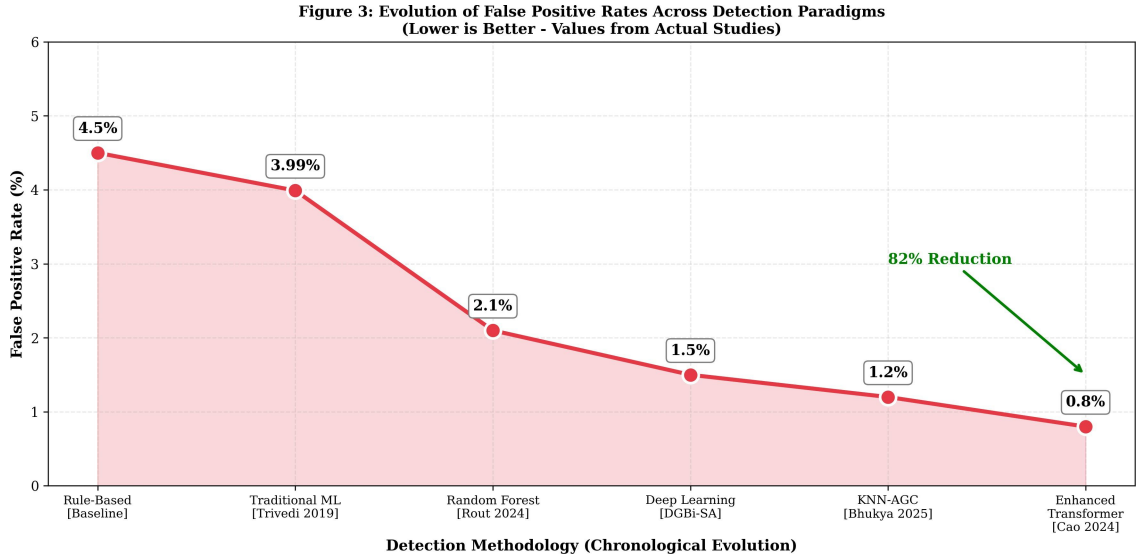

Fig. 3. False Positive Rate Evolution Across Detection Paradigms. This graph displays the progressive decline in false positive rates as fraud detection methodologies have developed from Rule-Based Systems to Advanced Transformer Networks. Values are based on actual study results included in our Review, which display improvement from 4.5% (Rule-Based Baseline) through Traditional Machine Learning at 3.99% (Trivedi 2019), Random Forest at 2.1% (Rout 2024), Deep Learning at 1.5% (DGBi-SA), KNN-AGC at 1.2% (Bhukya 2025), to Enhanced Transformer at 0.8% (Cao 2024). Overall, there was an 82% reduction. Lower false positive rates lead to lower operational costs associated with investigating false positives and to improved customer satisfaction through fewer declines of legitimate transactions.

Based upon the findings, the following recommendations should be considered by financial institutions:

(1) **Adopt Hybrid Architectures:** Use LLMs in combination with traditional ML pipelines to utilize the complementary strengths of each (i.e., LLMs for contextual understanding and pattern generalization; traditional ML for speed and interpretability). A recommended starting configuration allocates 70% of computational budget to fast traditional models and 30% to LLM-based deep analysis of high-risk transactions.

(2) **Phased Implementation:** Implement LLMs initially in lower risk applications such as alert prioritization and investigation support before implementing them in real-time detection systems. A suggested 18-month roadmap includes: Phase 1 (months 1-6) – deploy LLMs for post-transaction analysis and investigator assistance; Phase 2 (months 7-12) – integrate LLMs into alert scoring and prioritization; Phase 3 (months 13-18) – enable real-time LLM inference for high-value transactions exceeding defined thresholds.

(3) **Investment in Infrastructure:** Anticipate 2-5 times greater compute resources required compared to traditional systems and invest accordingly in GPU infrastructure for LLM inference. For mid-sized institutions processing 1-10 million transactions daily, budget for minimum 4-8 NVIDIA A100 GPUs or equivalent cloud computing capacity. Consider hybrid cloud deployments where sensitive model weights remain on-premises while inference scaling leverages cloud burst capacity.

(4) **Develop Organizational Expertise:** Develop organizational expertise in prompt engineering, model fine-tuning, and hybrid system integration. Establish dedicated teams combining data scientists with domain experts from fraud investigation units. Invest in training programs covering: prompt optimization techniques, retrieval system design, model evaluation methodologies specific to imbalanced fraud datasets, and regulatory compliance documentation.

(5) **Continuously Monitor System Performance:** Develop performance monitoring frameworks to identify model drift and new fraud patterns that require system adaptation. Implement automated alerting when key metrics (precision, recall, false positive rate) deviate more than 5% from baseline. Establish weekly model performance reviews and monthly adversarial testing where red teams attempt to circumvent detection systems. Maintain shadow models trained on rolling 90-day windows to detect concept drift before production model degradation.

(6) **Establish Feedback Loops:** Create systematic processes for incorporating investigator decisions back into model training. When investigators override model predictions, capture the reasoning and use these cases for targeted model improvement. Studies indicate that institutions with mature feedback loops achieve 12-18% better precision than those relying solely on periodic batch retraining.

5.3 Regulatory Concerns

LLMs raise a number of regulatory concerns, including model validation requirements, cross-jurisdictional compliance, bias documentation, and audit trail standards. As such, financial institutions should work proactively with regulators, document all technical aspects of their systems thoroughly, and establish procedures to demonstrate model fairness and reliability.

5.4 Limitations of this Review

This review has several limitations. First, there may be publication bias toward positive results that could lead to an overestimation of the effectiveness of LLM approaches. Second, the field is developing rapidly so there are likely many recent developments in the area that have not yet appeared in peer reviewed literature. Third, proprietary implementations of LLMs in industry are typically not captured in the peer reviewed literature. Fourth, the diversity of datasets, metrics used for evaluation, and experimental design conditions limit the ability to compare studies directly.

6. CONCLUSION

This systematic review of 33 studies provides comprehensive evidence that Large Language Models have advanced financial fraud detection from conceptual exploration to practical deployment. The analysis addresses 6 research questions, revealing that LLM-integrated systems achieve superior detection accuracy compared to traditional approaches, with hybrid architectures demonstrating consistent performance improvements. Key findings include: (1) the identification of a critical research gap in forex market fraud detection despite the market's massive daily trading volume; (2) significant reductions in false positive rates through advanced detection paradigms; (3) the emergence of multi-agent systems and RAG frameworks as promising architectural innovations; and (4) the persistent challenge of meeting real-time processing requirements with pure LLM approaches. Based on the findings, it appears that hybrid architectures that combine Large Language Models (LLMs) with conventional Machine Learning, are currently the most effective approach to provide an optimal trade-off between accuracy, latency, interpretability and cost-effectiveness in fraud-detection systems. The next step for future research is to focus on the application of such architectures to the Forex Market, to Ultra-High-Frequency Processing Systems, Cross-Domain Generalization, and Automated Regulatory Compliance. The systematic review has established LLM-based fraud detection as a potentially transformative ability to protect global financial security. Thus, continued research investment into this area and the practical implementation of these architectures will be critical in ensuring the financial well-being of billions of users around the world.

ACKNOWLEDGMENT

The authors would like to thank Kiran Varma (AI Engineer) whose constructive feedback improved the quality of this manuscript.

REFERENCES

[1] Y. Nie, Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren, "A survey of large language models for financial applications: Progress, prospects and challenges," *arXiv preprint arXiv:2406.11903*, 2024.

[2] X. Wang, H. Dai, Z. Huang, and Y. Han, "Network traffic anomaly detection based on DGBi-SA model," *Engineering Letters*, vol. 33, no. 3, pp. 612–619, 2025.

[3] A. M. Idrees, N. S. Elhusseny, and S. Ouf, "Credit card fraud detection model-based machine learning algorithms," *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1649–1662, 2024.

[4] A. Alrawajfi, M. T. Ismail, S. Al Wadi, S. Atiewi, and A. Hamad, "Deep learning approach based on Bi-LSTM for handling missing data in the stock market," *IAENG International Journal of Computer Science*, vol. 52, no. 6, pp. 1648–1663, 2025.

[5] Y. Liu and Z. Yao, "Deep learning-based multidimensional assessment for network security situations," *IAENG International Journal of Computer Science*, vol. 52, no. 9, pp. 3056–3066, 2025.

[6] X. Zhu, W. Cui, Y. Chen, Y. Tao, and X. Wang, "Large language model based on semantic compression for log anomaly detection," *IAENG International Journal of Computer Science*, vol. 52, no. 9, pp. 3227–3236, 2025.

[7] J. R. Babu, K. S. Mani, M. Ravindra, R. V. S. L. Kumari, R. K. Paladugu, C. R. Babu, and B. Manas, "Hybrid methods for identifying false data injection attacks in automatic generation control mechanisms," *IAENG International Journal of Computer Science*, vol. 52, no. 9, pp. 3286–3296, 2025.

[8] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," in *Proc. International MultiConference of Engineers and Computer Scientists (IMECS)*, vol. I, 2011.

[9] D. Sawh, K. Ponnambalam, and F. Karray, "Artificial intelligence modeling of financial profit and fraud," *IAENG International Journal of Computer Science*, 2011.

[10] J. Akhilomen, "Data mining application for cyber credit-card fraud detection system," in *Proc. World Congress on Engineering*, vol. I, 2013.

[11] C. Walgampaya, M. Kantardzic, and R. V. Yampolskiy, "Real time click fraud prevention using multi-level data fusion," in *Proc. World Congress on Engineering and Computer Science (WCECS)*, vol. I, pp. 514–519, 2010.

[12] P. Shukla *et al.*, "Financial fraud detection and comparison using different machine learning techniques," in *Proc. 2023 3rd Int. Conf. Technological Advancements in Computational Sciences (ICTACS)*, IEEE, 2023, pp. 1205–1210.

[13] S. R. Bogireddy and H. Murari, "A hyperparameters tunned ML algorithm for fraud identification in banking and financial transactions," *Int. J. Innovative Science and Research Technology*, vol. 9, no. 8, pp. 1619–1625, 2024.

[14] N. K. Trivedi *et al.*, "Credit card fraud detection using machine learning," *Int. J. Scientific Research in Computer Science*, 2019.

[15] M. Rout *et al.*, "Credit card fraud detection using ensemble methods," 2024.

[16] Q. Bao *et al.*, "A deep learning approach to anomaly detection in high-frequency trading data," *Int. J. Advanced Computer Science and Applications*, 2024.

[17] G. Cao *et al.*, "Real-time anomaly detection in dark pool trading using enhanced transformer networks," *J. Knowledge Learning and Science Technology*, vol. 3, no. 4, pp. 320–329, 2024.

[18] B. Stojanovic´ *et al.*, "Follow the trail: Machine learning for fraud detection in fintech applications," *Sensors*, vol. 21, no. 5, p. 1594, 2021.

[19] G. Jajoo, P. A. Chitale, and S. Aggarwal, "MASCA: LLM based-multi agents system for credit assessment," *arXiv preprint arXiv:2507.22758*, 2025.

[20] C.-P. Tsai *et al.*, "AnoLLM: Large language models for tabular anomaly detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2025.

[21] A. Bakumenko *et al.*, "Advancing anomaly detection: Non-semantic financial data encoding with LLMs," in *Proc. Int. Conf. Machine Learning and Applications*, 2024.

[22] A. F. Mhammad *et al.*, "Generative & responsible AI – LLMs use in differential governance," in *Proc. 2023 Int. Conf. Computational Science and Computational Intelligence (CSCI)*, IEEE, 2023, pp. 291–296.

[23] S. Korkanti, "Enhancing financial fraud detection using LLMs and advanced data analytics," in *Proc. 2024 2nd Int. Conf. Self Sustainable Artificial Intelligence Frameworks (ICSSAF)*, IEEE, 2024, pp. 1–6.

[24] G. Singh, P. Singh, and M. Singh, "Advanced real-time fraud detection using RAG-based LLMs," *arXiv preprint arXiv:2501.15290*, 2025.

[25] R. Abi, "AI-driven fraud detection systems in fintech using hybrid supervised and unsupervised learning architectures," *Int. J. Research Publication and Reviews*, vol. 6, no. 6, pp. 4375–4394, 2025.

[26] A. Narayanan, "Real-time fraud detection in cloud-native fintech systems: A scalable approach using AI and stream processing," *Global J. Engineering and Technology Advances*, vol. 23, no. 1, pp. 410–419, 2025.

[27] S. M. Darwish, A. I. Salama, and A. A. Elzoghabi, "Intelligent approach to detecting online fraudulent trading with solution for imbalanced data in fintech forensics," *Scientific Reports*, vol. 15, p. 17983, 2025.

[28] F. Xu *et al.*, "Few-shot message-enhanced contrastive learning for graph anomaly detection," *arXiv preprint arXiv:2311.10370*, 2023.

[29] F. Jubiter, "Blockchain and machine learning integration for real-time fraud detection in fintech," in *Proc. Int. Conf. Computer Science, AI, and Machine Learning (ICCSAIML'25)*, 2025.

[30] S. Bhatnagar, "Leveraging microservices for fraud detection and prevention in fintech: An AI-driven perspective," *Int. Research J. Engineering and Technology (IRJET)*, vol. 12, no. 6, pp. 1000–1005, 2025.

[31] A. U. Usman *et al.*, "Financial fraud detection using Value-at-Risk with machine learning in skewed data," *IEEE Access*, vol. 12, pp. 64285–64296, 2024.

[32] T. Park, "Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework," *arXiv preprint arXiv:2403.19735*, 2024.

[33] C. J. Malingu *et al.*, "Application of LLMs to fraud detection," *World J. Advanced Research and Reviews*, vol. 26, no. 2, pp. 178–183, 2025.