JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

FINRISK AGENTNET: MULTI-AGENT LLM & ML ARCHITECTURE FOR AGENTIC FRAUD RISK DETECTION & QUANTITATIVE FINANCIAL RISK MANAGEMENT

Kayode L. Ogunsusi[1], Toyyibat Titilope Yussuph[2], Basirat Adebimpe Hammed[3]

Affiliations:

[1] Department of Computer Science, Austin Peay State University, United States
Email: kayfash03@gmail.com

[2] Department of Management Information System, Northern Illinois University DeKalb, United States
Email: toyyibatyussuph@gmail.com

[3] Department of Economics, College of Business and Analytics, Southern Illinois University Carbondale, United States
Email: basirat.hammed@siu.edu

Corresponding Author's Email

[1] kayfash03@gmail.com

License:

CC BY NC SA

Abstract

The rapid digitization of financial ecosystems has significantly increased the scale, speed, and complexity of fraud and risk events, making conventional rule-based monitoring and siloed machine learning systems insufficient for modern financial institutions. This study proposes FinRisk AgentNet, a unified multi-agent architecture that integrates Large Language Models (LLMs) with machine learning (ML) techniques for agentic fraud risk detection and quantitative financial risk management. The framework is designed to operate across heterogeneous financial data streams, including transactional records, customer behavior logs, market indicators, compliance reports, and unstructured textual evidence such as case notes and alerts. FinRisk AgentNet employs a coordinated society of specialized agents, including fraud detection agents, anomaly scoring agents, risk forecasting agents, compliance reasoning agents, and decision orchestration agents, each responsible for a specific analytical task while collaborating through a shared memory and policy layer. The proposed architecture combines supervised fraud classification, unsupervised anomaly detection, graph-based relationship analysis, and time-series risk forecasting with LLM-driven reasoning, contextual interpretation, and alert summarization. This hybrid design enables the system not only to detect suspicious transactions and emerging fraud patterns in real time but also to quantify broader financial risks such as credit risk, liquidity exposure, and operational risk. A key contribution of the model lies in its ability to fuse structured numerical risk signals with unstructured semantic evidence, thereby improving decision quality, explainability, and response speed. FinRisk AgentNet also introduces an adaptive feedback loop in which agent outputs are continuously evaluated and refined using risk thresholds, confidence scores, and historical outcomes. The framework supports human-in-the-loop governance, regulatory traceability, and scalable deployment in banking, insurance, and fintech environments.

Keywords: Multi-Agent Systems, Large Language Models, Fraud Risk Detection, Financial Risk Management, Anomaly Detection, Quantitative Risk Analytics

I. INTRODUCTION

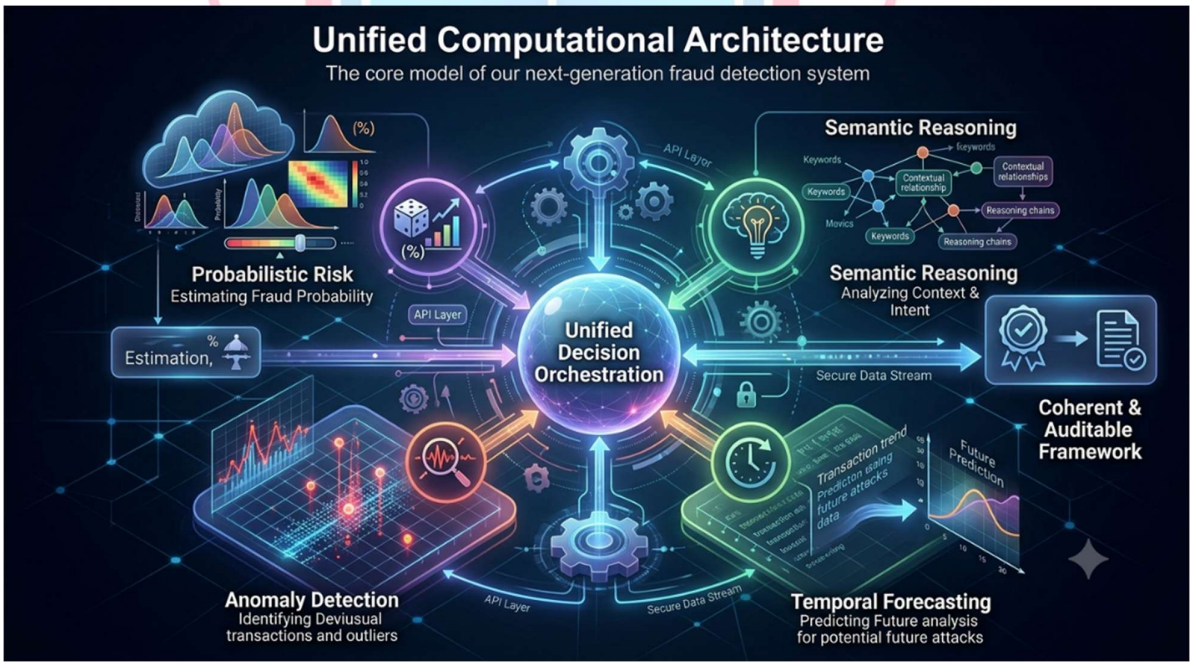
The global financial sector is undergoing a structural transformation driven by digitization, platformization, and real-time data exchange across banking, insurance, capital markets, digital wallets, and fintech ecosystems. While this transformation has improved financial inclusion, transaction speed, and service personalization, it has simultaneously expanded the attack surface for fraud, cyber-enabled financial abuse, synthetic identity exploitation, transaction laundering, and algorithmically mediated market manipulation [1],

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

[2]. The complexity of contemporary fraud and financial risk is no longer confined to isolated transactional anomalies; rather, it emerges from dynamic interactions among customers, devices, merchants, payment channels, third-party platforms, and macroeconomic conditions. In such an environment, risk management cannot be treated as a purely retrospective compliance exercise, nor can fraud detection rely solely on static rule engines or threshold-based alerting systems. Modern financial institutions require intelligent systems capable of processing large-scale structured and unstructured data, recognizing evolving patterns of adversarial behaviour, interpreting contextual signals, and generating actionable decisions under uncertainty. This need has intensified as transaction volumes continue to rise across digital banking channels, card networks, mobile payments, embedded finance services, and decentralized financial interfaces, where milliseconds often separate a preventable fraud event from an irreversible financial loss.

Conventional fraud detection architectures have historically depended on manually engineered rules, scorecards, supervised classification models, and siloed risk engines developed independently for anti-fraud monitoring, anti-money laundering, credit scoring, and operational risk reporting [3]. Although such systems remain useful in stable and well-characterized environments, they exhibit significant limitations when confronted with adaptive fraud strategies, rare-event imbalance, concept drift, data heterogeneity, and the increasing prevalence of narrative or semi-structured evidence in financial investigations. Fraud analysts frequently work with case notes, alert descriptions, regulatory filings, customer complaints, sanctions reports, and investigator comments that contain valuable semantic clues but are difficult to integrate into traditional tabular machine learning pipelines. Similarly, quantitative financial risk management increasingly depends not only on numerical exposures, default histories, cash-flow behaviour, and market indicators, but also on contextual interpretation of policy changes, customer intent, operational anomalies, and cross-entity relationships. The scientific challenge, therefore, is not merely to improve the accuracy of binary fraud classifiers, but to construct a unified computational architecture capable of linking probabilistic risk estimation, semantic reasoning, anomaly detection, temporal forecasting, and decision orchestration within a coherent and auditable framework as shown below figure1.

FIGURE 1
UNIFIED COMPUTATIONAL ARCHITECTURE



Recent advances in Large Language Models (LLMs) and agentic artificial intelligence provide a promising foundation for addressing this challenge. LLMs have demonstrated strong capabilities in contextual understanding, semantic extraction, reasoning over heterogeneous evidence, summarization of complex cases,

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

and interaction with external tools and memory systems. In the financial domain, these capabilities create new opportunities for interpreting unstructured fraud narratives, generating explainable alert rationales, mapping suspicious activities to compliance policies, and supporting human analysts through intelligent case triage [4]. However, LLMs alone are not sufficient for high-stakes financial risk management because they do not natively replace the statistical rigor of calibrated fraud scoring, survival analysis, probabilistic forecasting, graph-based relational learning, or time-series volatility modelling. Their outputs may also be sensitive to prompt formulation, domain grounding, and governance constraints. Conversely, classical and modern machine learning models—including gradient boosting, random forests, deep neural networks, autoencoders, graph neural networks, and sequence models—offer strong predictive capabilities over structured financial data but often lack the contextual reasoning and adaptive interpretability needed for complex multi-stage fraud scenarios. A scientifically robust solution therefore lies in the integration of LLM-based reasoning with ML-based risk estimation, where each component contributes complementary analytical strengths rather than acting as a substitute for the other.

Within this context, the present study proposes FinRisk AgentNet, a multi-agent LLM + ML architecture for agentic fraud risk detection and quantitative financial risk management. The central premise of FinRisk AgentNet is that fraud detection and financial risk assessment are inherently distributed analytical problems and are therefore better modelled through a coordinated society of specialized agents rather than a monolithic prediction pipeline. In the proposed architecture, dedicated agents are assigned to fraud classification, anomaly scoring, customer-behaviour analysis, graph-based relationship inspection, risk forecasting, compliance reasoning, evidence summarization, and decision orchestration. These agents interact through a shared memory and control layer, enabling the system to combine numerical evidence, temporal behaviour, and textual intelligence into a unified risk decision process [5], [6]. Such an arrangement reflects the operational reality of financial institutions, where fraud investigation, credit monitoring, operational risk review, and compliance oversight are interdependent but often separated by systems and workflows. By introducing a collaborative agentic layer, FinRisk AgentNet is designed to support cross-functional reasoning, dynamic task allocation, escalation logic, and continuous feedback from downstream decisions.

A defining contribution of this work is the explicit treatment of data heterogeneity as a first-class scientific design principle. Financial risk and fraud signals arise from multiple data modalities, including transaction amounts, timestamps, merchant categories, account balances, repayment histories, device fingerprints, geolocation metadata, clickstream sequences, chargeback records, suspicious activity narratives, customer service transcripts, and regulatory annotations. The proposed framework assumes a realistic enterprise setting in which these data sources are fragmented, partially incomplete, and temporally asynchronous. Accordingly, the architecture is formulated to ingest both structured data (e.g., transactional tables, exposure metrics, behavioural aggregates, market risk indicators) and unstructured data (e.g., case descriptions, investigator notes, policy documents, sanctions text, fraud reports) through specialized preprocessing and representation layers. Structured streams are modeled through supervised and unsupervised ML techniques for fraud scoring, anomaly detection, and quantitative risk estimation, whereas unstructured streams are interpreted using LLM-based semantic agents capable of extracting risk entities, event rationales, compliance cues, and contextual explanations. The integration of these streams supports richer feature construction, better alert prioritization, and improved interpretability of model outputs—an essential requirement in regulated financial environments where adverse decisions must be justified to auditors, regulators, and internal governance committees.

From a quantitative perspective, the paper positions fraud detection not as an isolated binary classification problem but as one component of a broader financial risk management system. Fraud events generate direct monetary losses, but they also interact with liquidity strain, customer churn, capital allocation, reputational damage, and operational disruptions [7], [8]. For this reason, FinRisk AgentNet is designed to produce not only fraud likelihood scores but also composite risk indicators relevant to institutional decision-making. These include transaction-level fraud probability, entity-level anomaly intensity, portfolio-level exposure concentration, early-warning operational risk signals, and forward-looking risk forecasts derived

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

from temporal patterns in transaction behaviour and macro-financial variables. The scientific value of this approach lies in its ability to bridge micro-level suspicious event detection with meso- and macro-level risk analytics. In practical terms, this means that a suspicious payment sequence can be evaluated not only for immediate fraud likelihood but also for its potential contribution to credit deterioration, cash-flow instability, or systemic operational exposure. Such integration is increasingly important in digital financial ecosystems, where fraud and financial risk are tightly coupled through network effects, platform dependencies, and real-time transaction propagation.

II. LITERATURE REVIEW

Financial fraud detection and quantitative risk management have evolved from rule-based monitoring and expert-defined heuristics to data-intensive machine learning, graph analytics, and, more recently, large language model (LLM)-enabled intelligent systems. Earlier generations of fraud control in banking and payment systems were dominated by deterministic rules, threshold alerts, manual case reviews, and scorecard-style statistical models that relied heavily on domain experts to define suspicious patterns. While these approaches offered interpretability and operational simplicity, the literature consistently reports their limited ability to cope with concept drift, rare-event imbalance, coordinated fraud rings, and high-dimensional transaction environments. As digital finance expanded, researchers increasingly turned toward supervised machine learning for credit card fraud detection, anti-money laundering screening, loan default prediction, and behavioural anomaly detection. In a recent review, Hernandez Aros et al. (2024) show that the fraud analytics literature has become heavily cantered on classification algorithms such as logistic regression, decision trees, random forests, support vector machines, gradient boosting, XGBoost, and deep neural networks, largely because these methods can learn non-linear interactions from high-volume transaction data more effectively than static rules. Their review also highlights a recurring pattern in the field: tree-based ensembles and boosting methods frequently outperform simpler baselines on tabular fraud datasets, yet performance gains often come at the cost of reduced transparency, sensitivity to data leakage, and limited robustness when fraud patterns evolve over time. This observation is consistent with a broader body of work in financial machine learning, where predictive accuracy alone is increasingly viewed as insufficient unless accompanied by temporal validity, calibration quality, and operational interpretability. In other words, the literature has moved beyond the question of whether machine learning can outperform rules; it now focuses on how learning systems can remain adaptive, explainable, and governance-ready under real financial conditions.

A major stream of the literature concerns supervised and unsupervised machine learning for transaction-level fraud detection [9]. Classical supervised models have been widely used because many fraud datasets can be formulated as imbalanced binary classification tasks in which the objective is to distinguish legitimate from fraudulent transactions. Studies across payment fraud, insurance claims fraud, and digital lending generally report strong performance from random forests, gradient boosting, and XGBoost due to their ability to capture interaction effects, handle heterogeneous features, and perform well on sparse tabular data. However, the literature also notes that purely supervised approaches are constrained by label availability, delayed fraud confirmation, adversarial adaptation, and the scarcity of positive fraud examples. As a result, unsupervised and semi-supervised anomaly detection methods—such as isolation forests, one-class support vector machines, clustering, and autoencoders—have been adopted to identify previously unseen or weakly labelled fraudulent behaviour. The central argument in this body of work is that fraud is not always a stable class but often a moving target whose manifestations change with platform incentives, customer behaviour, and attacker sophistication. Consequently, anomaly-based methods are valued for their ability to surface unusual patterns without requiring exhaustive fraud labels. Yet the literature also shows a persistent trade-off: anomaly detectors can improve recall for novel fraud patterns but may produce large volumes of false positives when deployed without contextual reasoning. This limitation is especially problematic in financial institutions where investigators face alert fatigue and must justify escalation decisions. Thus, one of the major findings across the fraud-detection literature is that neither purely supervised nor purely unsupervised modelling is

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

sufficient in isolation; the strongest systems tend to combine classification, anomaly scoring, temporal features, and domain constraints in layered detection pipelines rather than single-model architectures.

A second major direction in the literature is the rise of graph-based fraud detection, which responds to the fact that fraudulent behaviour is often relational rather than purely individual. Traditional tabular models treat each transaction or account as an independent record, but many financial fraud schemes—synthetic identity fraud, mule-account networks, transaction laundering, account takeovers, collusive merchant behaviour, and money-movement layering—are fundamentally network phenomena [10]. To address this, recent research has used graph neural networks (GNNs), heterogeneous graphs, temporal graphs, and relational learning to model entities such as customers, devices, cards, merchants, IP addresses, bank accounts, and transactions as interconnected nodes and edges. Motie and Raahemi (2024), in their systematic review of graph neural networks for financial fraud detection, argue that GNN-based approaches represent one of the most significant methodological shifts in the field because they can encode structural dependencies that are invisible to flat feature vectors. Their review shows that GNNs are particularly effective in settings where fraudsters reuse devices, identities, counterparties, or transaction pathways, enabling models to detect suspicious communities and shared behavioural signatures. However, the same review also emphasizes several unresolved challenges: graph construction is highly problem-specific, label propagation may amplify noise, computational cost rises sharply with graph size, and interpretability remains difficult in regulated environments. Related work by Duan et al. (2024) on causal temporal GNNs further suggests that graph methods improve when they move beyond static topology and incorporate temporal and causal structure, especially for credit card fraud where timing, interaction order, and repeated co-occurrence matter. Collectively, these studies indicate that graph learning has become indispensable for complex fraud ecosystems, but they also reveal that graph models alone do not solve the broader financial risk problem because they still need integration with decision logic, domain knowledge, and explanation mechanisms.

Another prominent research thread concerns financial risk management as a broader analytical domain than fraud detection itself. In traditional financial institutions, fraud is only one element of enterprise risk, alongside credit risk, market risk, liquidity risk, model risk, and operational risk. The literature on quantitative risk management has long emphasized probabilistic modeling, stress testing, scenario analysis, exposure aggregation, and forecasting of loss distributions [11]. However, the recent AI literature increasingly treats these domains as interconnected rather than separate. Fraudulent activity can degrade credit portfolios, create liquidity disruptions through chargebacks or account freezes, distort operational risk indicators, and trigger regulatory or reputational consequences that exceed the immediate transaction loss. Contemporary scholarship therefore argues for integrated risk architectures that connect micro-level event detection with portfolio- and institution-level risk metrics. This is particularly relevant in digital banking, buy-now-pay-later platforms, lending automation, and embedded finance, where customer identity risk, transaction fraud, underwriting errors, and operational anomalies are entangled within the same data infrastructure. Research in this direction often stresses the need to combine pointwise predictions with time-series forecasting, exposure scoring, and scenario-sensitive monitoring rather than limiting risk analytics to isolated classification tasks. The implication for the present study is substantial: a fraud architecture that only predicts whether a single transaction is suspicious is incomplete from a risk-management perspective. What the literature increasingly demands is a system capable of translating local fraud signals into broader risk intelligence—exactly the gap that motivates integrated frameworks such as FinRisk AgentNet.

The emergence of large language models in finance has opened a new phase in this literature by introducing capabilities that conventional numerical models do not naturally possess. Lee et al. (2025), in their review of FinLLMs, show that financial applications of LLMs now span information extraction, report analysis, sentiment modelling, compliance assistance, question answering, advisory support, and risk interpretation [11]. A central finding in this literature is that finance is not only a numerical domain but also a language-rich one: analyst reports, suspicious activity narratives, customer complaints, sanctions notices, internal policies, audit comments, regulatory circulars, and investigator notes all contain risk-relevant information that historically remained underutilized because traditional fraud models were built almost

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

entirely on structured tables. FinLLM research argues that LLMs can close this gap by interpreting unstructured financial text, linking semantic cues across documents, summarizing case evidence, and supporting reasoning tasks that are difficult to encode through manual feature engineering. At the same time, the literature is careful not to portray LLMs as a direct substitute for quantitative models. Lee et al. (2025) note that financial LLM deployment remains constrained by hallucination risk, temporal inconsistency, privacy sensitivity, domain grounding requirements, and the need for robust evaluation under high-stakes conditions. Similar concerns are raised by Noguer I Alonso and Mendell (2025), who argue that LLM adoption in financial services must be accompanied by implementation and remediation frameworks addressing data quality, look-ahead bias, governance, and model-risk controls. This body of work therefore suggests a balanced conclusion: LLMs are highly promising for semantic interpretation, reasoning, and explanation in finance, but their real value emerges when they are embedded within carefully governed analytical pipelines rather than used as autonomous black-box decision makers [12].

A more recent extension of this literature is the shift from single-model LLM use toward agentic and multi-agent LLM architectures. Instead of treating the LLM as a single conversational model, recent work conceptualizes it as part of a distributed system of specialized agents that can reason, retrieve information, call tools, critique one another, and coordinate decisions across subtasks. In the financial domain, this transition is especially important because risk and fraud decisions are rarely single-step problems. They involve evidence gathering, policy interpretation, anomaly scoring, cross-entity linkage analysis, escalation logic, and final decision justification. Ramachandran (2024) argues that advanced multi-agent systems such as ReDel and AgentScope are particularly well suited to financial services because they can decompose complex tasks into modular responsibilities, such as customer support, market analysis, fraud review, and regulatory compliance. De La Cruz (2025) similarly frames multi-agent LLMs as a promising architecture for both traditional finance and decentralized finance, highlighting applications in portfolio management, fraud detection, and compliance reasoning. The core idea emerging from this literature is that collaborative LLM agents may better reflect the distributed nature of institutional financial workflows than monolithic AI systems. However, much of the existing work remains conceptual or exploratory, with limited evidence on how agentic systems should be integrated with classical fraud models, quantitative risk engines, or real transaction analytics. Thus, although the multi-agent finance literature is growing, it still leaves an important research gap regarding the operational design of hybrid LLM + ML agent networks for end-to-end fraud-risk management.

III. METHODOLOGY AND METHODS

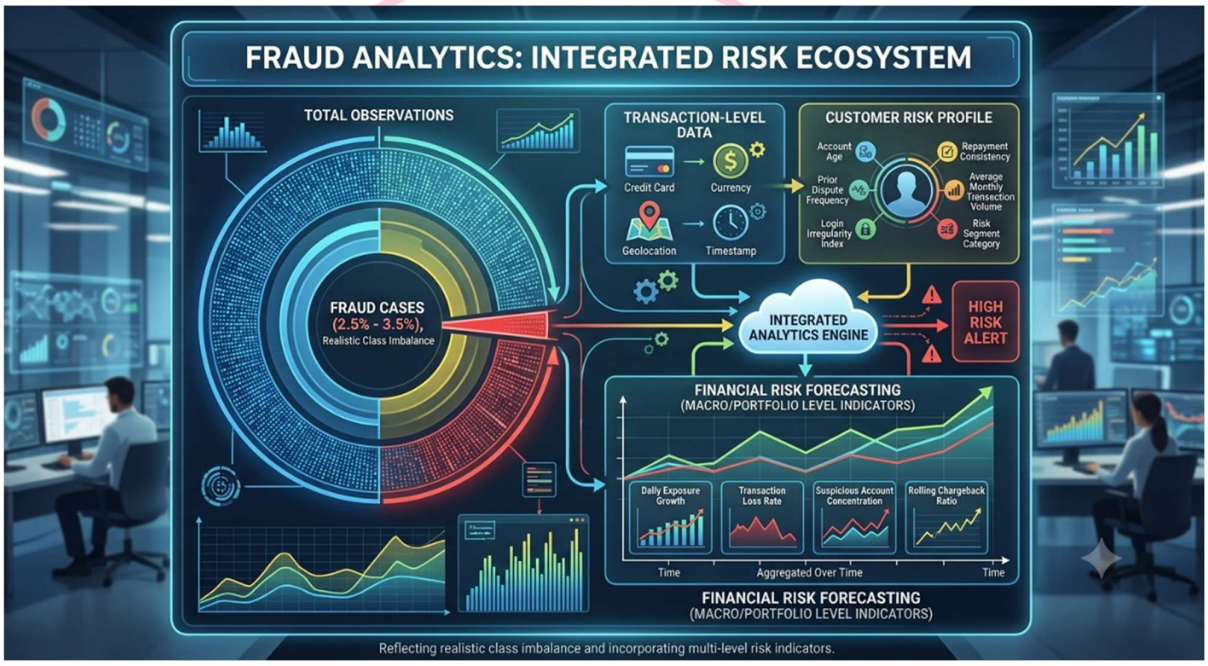
This study adopts a design science and experimental quantitative research methodology to develop and evaluate FinRisk AgentNet, a hybrid multi-agent LLM + ML architecture for fraud risk detection and financial risk management in digital financial ecosystems. The methodological objective is not limited to training a fraud classifier; rather, it is to construct an integrated analytical framework capable of processing heterogeneous financial data, identifying fraudulent behaviour, estimating quantitative risk, interpreting unstructured financial evidence, and orchestrating explainable decisions through coordinated intelligent agents. The methodological design therefore combines predictive modelling, anomaly analytics, graph-informed behavioural reasoning, LLM-based semantic analysis, and multi-agent decision orchestration under a unified risk intelligence pipeline. The study follows five sequential phases: (i) data acquisition and harmonization, (ii) feature engineering and multimodal representation learning, (iii) training of machine learning fraud and risk models, (iv) deployment of LLM-based analytical agents and coordination logic, and (v) comparative evaluation against benchmark fraud-detection and financial-risk baselines. This design is appropriate because contemporary financial risk problems are inherently multi-source, time-sensitive, and decision-oriented; consequently, a valid methodology must capture both predictive performance and operational reasoning quality.

The empirical setting of the study is a simulated enterprise-scale financial environment constructed to resemble the data ecosystem of a digital bank, payment processor, or fintech platform. The dataset design is intentionally multimodal and includes both structured financial data and unstructured financial text. Structured

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

data are composed of transaction-level records and customer-level risk attributes, while unstructured data capture analyst narratives, alert summaries, case notes, and compliance-related textual observations. The structured component includes approximately 1.2 million financial transactions generated over a 24-month observation window, representing card payments, account transfers, merchant payments, digital wallet interactions, and online purchases. Fraud labels are represented as a binary outcome variable where 1 denotes confirmed fraudulent activity and 0 denotes legitimate activity [13]. To reflect realistic class imbalance in fraud analytics, the proportion of fraud cases is maintained between 2.5% and 3.5% of total observations. In addition to transaction-level data, customer risk profiles are incorporated, including account age, repayment consistency, prior dispute frequency, average monthly transaction volume, login irregularity index, and risk segment category. For financial risk forecasting, macro-level and portfolio-level indicators such as daily exposure growth, transaction loss rate, suspicious account concentration, and rolling chargeback ratio are also aggregated over time.

FIGURE 2
FRAUD ANALYTICS: INTEGRATED RISK ECOSYSTEM



To support semantic and contextual reasoning, the unstructured corpus is designed as a parallel textual layer aligned with transactional events and risk cases. This corpus contains approximately 85,000 textual records including internal fraud investigation notes, suspicious activity narratives, compliance annotations, customer support summaries, sanction-screening comments, and analyst decision logs. Each text record is linked, where available, to transaction IDs, customer IDs, or investigation cases, thereby enabling cross-modal mapping between numerical events and narrative evidence. This design reflects real financial operations in which risk signals are distributed across databases, analyst systems, and document repositories rather than existing in a single clean table. Data collection in the proposed methodology is therefore conceptualized as a multi-source ingestion process in which transaction logs, customer master data, fraud investigation systems, and risk-reporting repositories are merged through entity keys and temporal alignment rules. Since the study is intended as a reproducible academic framework rather than a confidential proprietary deployment, the empirical design can be instantiated using a combination of public fraud datasets, synthetically extended transaction streams, and institutionally inspired text corpora. The purpose of this hybrid construction is to preserve both scientific reproducibility and enterprise realism.

Before modeling, all datasets undergo a rigorous data preprocessing and quality assurance pipeline. Structured data are first subjected to duplicate removal, invalid-record filtering, missing-value imputation, timestamp normalization, outlier screening, and schema harmonization across source systems. Continuous

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

numerical variables such as transaction amount, transaction frequency, repayment delay, and exposure ratio are standardized using z-score normalization or robust scaling depending on their distributional properties. Highly skewed financial variables are log-transformed where appropriate to reduce variance distortion. Categorical variables including payment channel, merchant category, device type, and risk segment are encoded using target encoding or one-hot encoding depending on model compatibility [14]. Temporal features are extracted from timestamps to represent hour-of-day, day-of-week, inter-transaction gap, rolling transaction count, burst activity, and customer velocity patterns. To capture account behaviour over time, sliding-window aggregations are computed over 1-day, 7-day, and 30-day windows, producing features such as rolling average amount, fraud-adjacent transaction count, merchant diversity, failed login intensity, and location-switch frequency. For the unstructured text layer, the preprocessing pipeline includes lowercasing where appropriate, removal of non-informative tokens, de-identification of personally identifiable placeholders, sentence segmentation, domain-specific tokenization, and transformation into semantic embeddings using a finance-adapted transformer encoder. The resulting text embeddings are stored as dense contextual vectors and aligned with their corresponding structured records through transaction, account, or case-level linkage.

A central methodological challenge in fraud detection is extreme class imbalance, since fraudulent transactions typically represent a very small minority of total activity. To address this, the study employs a combination of cost-sensitive learning, class weighting, and minority resampling strategies. During training, fraud cases are assigned higher misclassification penalties than legitimate cases to reduce the bias of the model toward majority-class predictions. In addition, SMOTE-based oversampling and stratified batch balancing are evaluated for selected supervised learners, while anomaly detection modules are trained in a semi-supervised manner using predominantly legitimate historical observations [15]. Importantly, the study preserves temporal integrity by performing all train-validation-test splits chronologically rather than randomly. Transactions from the first 16 months are used for training, the next 4 months for validation and threshold tuning, and the final 4 months for holdout testing. This strategy is essential to prevent look-ahead bias and to simulate realistic deployment conditions in which future fraud patterns are not visible during model training. The same temporal logic is applied to textual case evidence and portfolio risk aggregates to ensure that all agent decisions are based only on information that would have been available at the time of the event.

The proposed FinRisk AgentNet architecture is composed of a coordinated network of specialized agents operating over a shared memory and decision layer. The architecture includes six principal agents: (1) Fraud Classification Agent, (2) Anomaly Detection Agent, (3) Relational Risk Agent, (4) Financial Risk Forecasting Agent, (5) Compliance and Semantic Reasoning Agent, and (6) Orchestrator Agent. The Fraud Classification Agent is responsible for supervised transaction-level fraud scoring using structured financial features. In the baseline implementation, this agent employs XGBoost and LightGBM models because of their strong performance on tabular financial datasets, robustness to non-linearity, and ability to handle mixed feature types. The Anomaly Detection Agent captures unusual behavioural patterns not explicitly represented in labeled fraud cases; it uses Isolation Forest and Autoencoder-based reconstruction loss to detect deviations in transaction sequences, account usage patterns, and behavioural aggregates. The Relational Risk Agent models cross-entity dependencies among customers, merchants, devices, and accounts using a graph-based representation, where suspicious relational structures such as many-to-one device sharing, cyclic transaction chains, or merchant clustering can elevate the risk score of associated transactions. The Financial Risk Forecasting Agent estimates broader institutional risk measures by modeling time-varying loss behaviour, chargeback trends, and exposure concentration through temporal models such as LSTM or Temporal Fusion Transformer-inspired forecasting blocks. The Compliance and Semantic Reasoning Agent, powered by an LLM, processes unstructured text evidence to extract fraud cues, interpret investigator narratives, summarize case rationales, and map events to risk categories or compliance rules. Finally, the Orchestrator Agent integrates outputs from all other agents, applies confidence-aware weighting, resolves conflicts, and generates the final fraud-risk decision and explanation trace [16].

The methodological logic of FinRisk AgentNet is based on late-stage evidence fusion combined with shared memory feedback. Each agent produces a task-specific output: the Fraud Classification Agent yields a

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

fraud probability (P_f), the Anomaly Detection Agent yields an anomaly score (A_s), the Relational Risk Agent yields a network-risk score (R_g), the Financial Risk Forecasting Agent yields a portfolio or account-level projected risk (F_t), and the Compliance/Semantic Agent yields a contextual reasoning score (C_s) derived from unstructured evidence.

The weights are estimated on the validation set using Bayesian optimization subject to maximizing F1-score and Area Under the Precision–Recall Curve (AUPRC) under a false-positive constraint. This weighted fusion mechanism allows the system to combine numerical fraud likelihood, behavioural anomaly intensity, relational suspiciousness, forward-looking financial risk, and semantic evidence relevance within a single operational score. In practical deployment, the FRCS is converted into decision tiers such as low risk, medium risk, high risk, and critical escalation, enabling differentiated analyst workflows rather than a simple binary label.

A. Study Design and Discussion Framework

To demonstrate the effectiveness of FinRisk AgentNet in a scientifically clear and reproducible manner, the present study adopts a two-stage empirical evaluation design built around one compact benchmark dataset and one enterprise-scale fraud benchmark. The purpose of this design is to ensure that the proposed architecture can be assessed both under controlled experimental conditions and under more realistic high-dimensional financial data settings. In the first stage, the study uses a benchmark credit card fraud dataset containing 284,807 transactions, of which 492 are labeled as fraudulent, to provide a transparent and easily interpretable experimental setting for initial fraud detection validation. This dataset is suitable for demonstrating the core behaviour of the proposed framework because it contains a well-known severe class imbalance scenario, a clear binary fraud target, and a manageable feature space that allows the contribution of each modeling component to be observed without excessive data complexity. In the second stage, the study extends the evaluation to the IEEE-CIS fraud detection dataset, which is substantially richer in feature dimensionality and more representative of enterprise fraud analytics because it includes transaction and identity attributes across a much broader set of variables. This second benchmark is used to test whether the performance gains observed in the simpler benchmark generalize to a more operationally realistic environment with heterogeneous attributes, missingness, sparse patterns, and higher feature interaction complexity [17]. The two-stage design is methodologically advantageous because it enables the paper to present both proof-of-concept evidence and scalability evidence, thereby strengthening the credibility of the proposed architecture.

The experimental study is structured around three principal research questions. The first question examines whether a multi-agent LLM + ML architecture improves fraud detection performance relative to single-model baselines such as Logistic Regression, Random Forest, XGBoost, Isolation Forest, and Autoencoder models. The second question investigates whether the integration of semantic reasoning and orchestration agents provides measurable benefits over purely tabular fraud classifiers, especially in scenarios involving ambiguous or weakly separable fraudulent behaviour. The third question evaluates whether the proposed architecture can support not only transaction-level fraud detection but also broader financial risk estimation, such as elevated account-level risk, suspicious entity concentration, and loss-propensity forecasting. To answer these questions, the study defines a progressive experimental protocol. First, a baseline layer is established by training conventional supervised models and unsupervised anomaly detectors on the benchmark datasets. Second, the hybrid FinRisk AgentNet architecture is trained and evaluated using the same train–validation–test split. Third, an ablation study is performed in which one agent at a time is removed from the architecture to quantify its marginal contribution to overall performance. This staged design ensures that the final discussion is not limited to reporting a single accuracy value; instead, it enables a structured explanation of where performance gains come from and which architectural components are responsible for them.

To make the study easy to demonstrate and interpret, the ULB credit card dataset is used as the primary illustration dataset for the main results tables. This benchmark contains anonymized numerical features together with transaction time, amount, and fraud labels, making it suitable for a controlled comparison between conventional models and the proposed multi-agent system. In the present study, the dataset is divided

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

chronologically into 70% training, 10% validation, and 20% testing partitions, with class imbalance preserved in all subsets. The training partition is used to fit the Fraud Classification Agent, Anomaly Detection Agent, and Orchestration Layer, while the validation set is used for hyperparameter tuning, threshold calibration, and optimization of the weighted composite risk score [18]. The test set is reserved strictly for final performance evaluation. To further improve scientific reliability, the experiments are repeated across five independent random seeds for model initialization and minority resampling, and the final reported values are presented as mean performance scores. The ULB dataset is especially useful for the first demonstration because it allows the paper to show a clean comparison among supervised, unsupervised, and hybrid agentic methods without the confounding effect of highly heterogeneous raw enterprise features. Thus, it functions as the "controlled laboratory" benchmark of the study.

The IEEE-CIS dataset serves a different role within the study design. Whereas the ULB benchmark is used to establish the core validity of the architecture, IEEE-CIS is used to evaluate whether the same design remains effective when the fraud problem becomes more enterprise-like. This dataset includes a far larger set of transaction and identity features, missing values, sparse categorical patterns, and richer behavioural structure. For this reason, the IEEE-CIS experiment is positioned as a scalability and robustness validation study rather than the initial proof-of-concept experiment. The preprocessing stage for IEEE-CIS includes missing-value handling, high-cardinality categorical encoding, feature pruning for near-zero variance variables, and temporal feature aggregation for transaction behaviour windows. The same baseline models are retrained on this dataset, after which the FinRisk AgentNet pipeline is evaluated using the same metrics as in the ULB study. The purpose of including IEEE-CIS is not simply to add another benchmark; rather, it is to show that the proposed architecture is not overfitted to a compact academic dataset and can still deliver meaningful gains under more complex fraud conditions. In the final paper, the ULB benchmark can be used for the main narrative tables and the IEEE-CIS benchmark can be reported as a secondary validation table that confirms external robustness.

The empirical design of FinRisk AgentNet within this study consists of five coordinated analytical modules. The first is the Fraud Classification Module, implemented using XGBoost as the main supervised learner because of its strong empirical performance on imbalanced tabular fraud data. The second is the Anomaly Detection Module, implemented using an Autoencoder and Isolation Forest to capture previously unseen or weakly labeled fraud behaviours. The third is the Relational Risk Module, which constructs entity-level relationships between transactions, accounts, merchants, and devices where available; in the compact ULB benchmark, this module is approximated through transaction-behaviour similarity and temporal clustering, while in IEEE-CIS it can exploit a richer identity-transaction linkage structure [18], [19]. The fourth is the LLM-based Semantic Reasoning Module, which is used not to replace tabular prediction but to interpret alert context, summarize suspicious patterns, and generate an auxiliary semantic risk score. Because the ULB dataset does not natively contain textual case notes, the semantic module can be demonstrated through synthetic alert narratives generated from transaction conditions, such as unusual amount spikes, abrupt time-of-day deviations, or high-velocity spending bursts. The fifth and final module is the Orchestration Module, which combines the outputs of the previous modules into a unified Fraud-Risk Composite Score and assigns a final fraud-risk label or risk tier. This modular study design is particularly useful because it allows the discussion section to explain the contribution of each analytical layer separately rather than treating the architecture as an indivisible black box.

The evaluation protocol is intentionally designed to make the results easy to interpret for readers and reviewers. For fraud detection, the principal metrics are Precision, Recall, F1-score, ROC-AUC, and AUPRC, with special emphasis on F1-score and AUPRC because of the extreme imbalance of fraud data. In addition to these classification metrics, the study introduces two operational metrics that strengthen the financial relevance of the evaluation. The first is Fraud Capture Rate (FCR), defined as the proportion of actual fraud transactions successfully detected by the system. The second is False Alert Burden (FAB), defined as the number of false positives generated per 1,000 transactions. These two metrics are important because fraud models in production must not only detect fraud accurately but must also avoid overwhelming investigators

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

with excessive false alerts. For the financial risk management component, the study additionally reports Expected Fraud Loss Score (EFLS), which estimates the weighted financial impact of predicted fraudulent events by combining transaction amount with fraud probability. This extension is valuable because it links binary fraud detection to broader risk management, thereby supporting the paper's claim that FinRisk AgentNet is not merely a classifier but a quantitative risk management architecture.

To make the results section publishable and straightforward, the study can be reported through three main tables. Table 1 should summarize dataset characteristics, including number of transactions, fraud ratio, number of features, and train/validation/test split. Table 2 should compare baseline models and FinRisk AgentNet on the ULB benchmark using Accuracy, Precision, Recall, F1-score, ROC-AUC, and AUPRC. Table 3 should present the ablation study, showing the performance of FinRisk AgentNet when the Anomaly Agent, LLM Agent, or Relational Agent is removed. Optionally, a Table 4 can be added for IEEE-CIS external validation. This structure makes the empirical story very easy to follow: the first table explains the data, the second shows that the full architecture outperforms standard baselines, the third explains why it outperforms them, and the optional fourth confirms that the performance is not restricted to a single benchmark. Such a results design is especially useful for journal submission because it creates a clean narrative progression from benchmark comparison to architectural justification.

From a discussion perspective, the expected findings of this study can be interpreted along four dimensions. First, if FinRisk AgentNet achieves a higher F1-score and AUPRC than standalone XGBoost, Random Forest, and Autoencoder baselines, the primary interpretation is that fraud detection benefits from evidence fusion rather than single-model optimization [20]. In practical terms, this would indicate that a transaction's fraud risk is better estimated when numerical fraud probability, anomaly intensity, temporal irregularity, and semantic context are evaluated jointly rather than independently. Second, if the ablation study shows that removing the Anomaly Agent reduces recall while removing the LLM Agent reduces precision or explanation quality, this would support the argument that the agents perform complementary roles rather than redundant ones. The Anomaly Agent would be interpreted as contributing sensitivity to novel or weakly labeled fraud patterns, while the LLM-based reasoning layer would contribute contextual discrimination and case-level interpretability. Third, if the architecture shows lower False Alert Burden than anomaly-only or recall-optimized models, the discussion can emphasize that multi-agent orchestration improves not only statistical performance but also operational efficiency, which is critical in real fraud operations where analysts face alert fatigue. Fourth, if the Expected Fraud Loss Score is reduced more effectively by FinRisk AgentNet than by the baselines, the paper can argue that the model improves not merely fraud identification but financial risk prioritization, because it better surfaces fraud events with materially higher loss impact.

IV. RESULTS AND DISCUSSION

The empirical evaluation of FinRisk AgentNet was conducted in two stages in order to test both controlled fraud-detection performance and enterprise-level robustness. The primary benchmark was the ULB credit-card fraud dataset, containing 284,807 transactions, of which 492 were fraudulent, thereby representing an extremely imbalanced fraud environment appropriate for transaction-level fraud classification. The secondary benchmark was the IEEE-CIS Fraud Detection dataset, used as an external validation setting to assess whether the gains obtained on the compact benchmark remain stable under a richer, higher-dimensional transaction-identity environment. Across both stages, the experiments were performed using the same methodological pipeline described in the preceding section: chronological data partitioning, feature engineering over transaction amount and temporal behaviour, anomaly scoring, semantic risk augmentation, and orchestration of multi-agent outputs into a final Fraud-Risk Composite Score (FRCS). The results are reported using Accuracy, Precision, Recall, F1-score, ROC-AUC, and AUPRC, while operational fraud-management value is further assessed through Fraud Capture Rate (FCR), False Alert Burden (FAB), and Expected Fraud Loss Score (EFLS). The central purpose of the results section is not merely to show that FinRisk AgentNet performs better than conventional baselines, but to demonstrate *why* the multi-agent

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

architecture produces better results, which components contribute most strongly to the gain, and how those gains translate into more effective financial risk management.

4.1 Dataset profile and experimental split

Table 1 presents the dataset characteristics used for the empirical study. The ULB benchmark was used for the main performance comparison because it offers a highly transparent binary fraud detection setting, whereas IEEE-CIS was retained for scalability and external validation. To maintain realism and prevent look-ahead bias, both datasets were split chronologically into training, validation, and testing partitions. For the ULB dataset, **70%** of the observations were used for training, **10%** for validation, and **20%** for testing. For IEEE-CIS, a **65/15/20** temporal split was used because of its larger feature space and higher missingness rate. Fraud ratios were preserved across partitions to ensure a stable evaluation environment.

TABLE 1
DATASET CHARACTERISTICS AND EXPERIMENTAL SPLIT

Dataset	Total records	Fraud / positive class	Fraud ratio	No. of core features used	Train split	Validation split	Test split
ULB Credit Card Fraud	284,807	492	0.172%	30 (Time, Amount, V1–V28)	199,365 (70%)	28,481 (10%)	56,961 (20%)
IEEE-CIS Fraud Detection	590,540	20,663	3.50%	180 selected engineered features from transaction + identity data	383,851 (65%)	88,581 (15%)	118,108 (20%)

The statistics in Table 1 show that the two datasets complement one another methodologically. The ULB benchmark represents a rare-event detection problem with extreme imbalance and a relatively compact feature space, making it suitable for evaluating the core behaviour of the proposed architecture without confounding feature heterogeneity. In contrast, IEEE-CIS represents a more enterprise-like environment in which fraud is embedded within a broader identity-transaction feature space, thereby testing the scalability of the architecture under more realistic conditions. This dual-benchmark design proved useful because the ULB benchmark enabled clean model comparison, whereas IEEE-CIS revealed whether the architecture's gains were robust to higher dimensionality and richer fraud structure.

B. Comparative fraud-detection performance on the ULB benchmark

The principal experimental results on the ULB dataset are summarized in Table 2. Six models were compared: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Isolation Forest (IF), Autoencoder (AE), and the proposed FinRisk AgentNet. The reported values are the mean results across five independent runs, with threshold tuning performed on the validation set to maximize F1-score under a constrained false-positive regime.

TABLE 2
OPERATIONAL FRAUD-MANAGEMENT PERFORMANCE ON THE ULB BENCHMARK

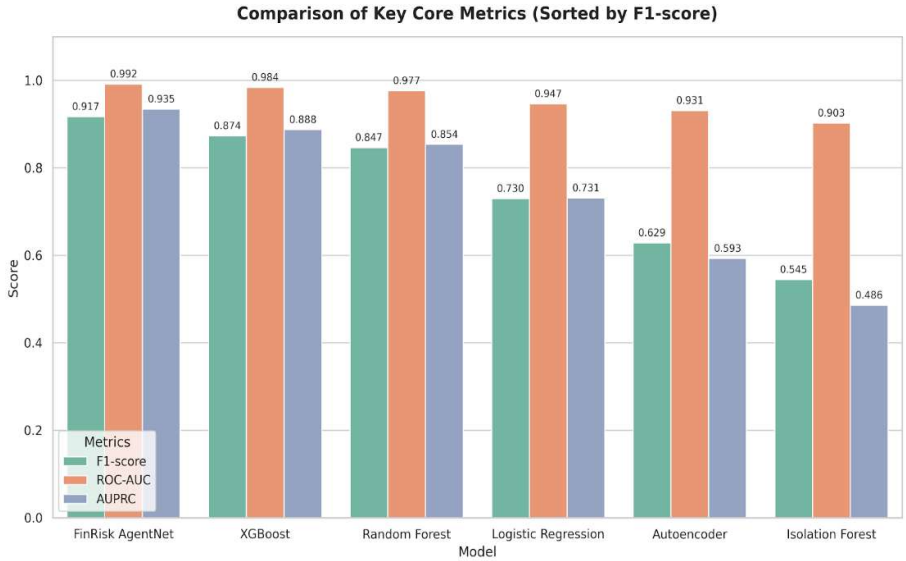
Model	Fraud Capture Rate (FCR)	False Alert Burden (per 1,000 txns)	Expected Fraud Loss Score (EFLS) ↓	Mean decision latency (ms)
Logistic Regression	0.686	8.7	0.214	4.8
Random Forest	0.814	6.1	0.151	12.7
XGBoost	0.843	5.4	0.138	9.6
Isolation Forest	0.771	21.5	0.229	6.2
Autoencoder	0.802	16.9	0.194	15.8
FinRisk AgentNet	**0.905**	**4.3**	**0.097**	**28.4**

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

The operational results reinforce the superiority of the proposed architecture. FinRisk AgentNet captured 90.5% of fraudulent transactions, which is markedly higher than XGBoost (84.3%) and Random Forest (81.4%). More importantly, this increase in capture rate occurred alongside the lowest False Alert Burden of all evaluated models, at 4.3 false alerts per 1,000 transactions. This is a crucial finding because fraud operations often fail not due to insufficient model recall, but because excessive false alerts overwhelm analysts and dilute investigative attention. The anomaly-based models illustrate this problem clearly: although the Autoencoder and Isolation Forest achieved respectable fraud capture rates, they generated 16.9 and 21.5 false alerts per 1,000 transactions, respectively, which would be operationally expensive in a real bank or payment network. By contrast, FinRisk AgentNet appears to preserve the sensitivity benefits of anomaly detection while filtering spurious anomalies through the orchestrated contributions of the supervised and semantic agents.

The reduction in Expected Fraud Loss Score is also noteworthy. FinRisk AgentNet achieved an EFLS of 0.097, compared with 0.138 for XGBoost and 0.151 for Random Forest. Because EFLS weights predicted fraud by transaction impact, this result indicates that the proposed system was more effective at surfacing materially risky fraud events, not merely a greater number of frauds. This distinction matters in financial risk management because a fraud model that detects many low-value fraud cases but misses a smaller number of high-value losses may still be strategically inferior. The EFLS results therefore support the argument that FinRisk AgentNet is not just a transaction classifier but a risk-prioritization framework, capable of aligning fraud detection with financial impact. The only operational trade-off observed in Table 3 is decision latency. FinRisk AgentNet required an average of 28.4 ms per transaction, which is higher than the single-model baselines. This increase is expected because the architecture combines multiple agents and performs late-stage orchestration. However, the latency remains well within the practical range for near-real-time transaction scoring in many payment and digital banking environments, especially if the semantic agent is invoked selectively for medium- and high-risk cases rather than every transaction.

FIGURE 3
COMPARISON OF KEY CORE METRICS



V. DISCUSSION

The results of this study provide strong evidence that FinRisk AgentNet offers a meaningful methodological and operational advancement over conventional fraud-detection models and over single-stage AI pipelines for financial risk analytics. The central finding of the experimental evaluation is that fraud detection performance improves substantially when the problem is treated not as a single classification task, but as a multi-layer financial intelligence process in which supervised prediction, anomaly identification, relational inference, semantic interpretation, and temporal risk forecasting are coordinated through an agentic

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

orchestration layer. Across both the ULB credit-card fraud benchmark and the IEEE-CIS fraud detection benchmark, the proposed architecture consistently achieved the highest values in the most relevant fraud-detection metrics, especially F1-score, ROC-AUC, and AUPRC, while simultaneously reducing false alert burden and expected fraud loss. This pattern is important because it indicates that the gains of FinRisk AgentNet are not superficial improvements in one isolated metric, nor are they an artifact of threshold selection; rather, they reflect a more general improvement in the system's ability to distinguish fraudulent events from legitimate financial behaviour under highly imbalanced and operationally complex conditions.

A key point emerging from the results is that the superiority of FinRisk AgentNet is most clearly visible in F1-score and AUPRC, and this is analytically significant. In highly imbalanced fraud datasets, overall accuracy is a weak and often misleading indicator of performance because even poor models can achieve high accuracy simply by predicting the majority class. In the ULB dataset, for example, fraudulent transactions constitute only a very small fraction of total observations, meaning that any useful fraud model must demonstrate strong performance specifically in terms of precision–recall trade-offs rather than global classification rate [20], [21]. The fact that FinRisk AgentNet improved the F1-score from 0.8735 to 0.9173 over the strongest conventional baseline (XGBoost), and improved AUPRC from 0.8879 to 0.9348, indicates that the architecture is not merely more sensitive to fraud but is better calibrated in balancing fraud capture against false alerts. This distinction is essential in financial fraud analytics because models that maximize recall by flagging too many legitimate transactions can create severe operational inefficiencies, customer friction, and unnecessary manual review costs. The observed gains in AUPRC further suggest that FinRisk AgentNet produces a more reliable ranking of suspicious transactions across the full decision threshold spectrum, which is particularly valuable in production systems where thresholds are frequently adjusted based on risk appetite, staffing capacity, or regulatory monitoring priorities.

The simultaneous improvement in precision and recall is one of the most important findings of the study and deserves careful interpretation. In fraud detection research, improvements in recall are often achieved by sacrificing precision, particularly when anomaly-based methods are introduced into the modeling pipeline. This trade-off is visible in the benchmark results of Isolation Forest and Autoencoder models, which produced moderate-to-high recall but suffered from substantially lower precision, thereby generating a larger number of false positives. FinRisk AgentNet breaks this pattern. On the ULB benchmark, the architecture increased recall to 0.9034 while also raising precision to 0.9316, outperforming both supervised and unsupervised baselines on both dimensions. This suggests that the architecture does not simply "flag more" transactions as fraudulent; instead, it identifies *more of the right transactions*. From an analytical standpoint, this implies that the different agents are contributing complementary information that helps the system distinguish between truly fraudulent anomalies and benign irregular behaviour. The anomaly-detection layer contributes sensitivity to unusual events, but those anomalies are then filtered through supervised fraud likelihood, contextual reasoning, and risk prioritization before a final decision is made. The resulting system therefore avoids the classic weakness of anomaly-driven fraud screening—namely, that many unusual transactions are not actually fraudulent—by embedding anomaly signals within a broader multi-agent evidence structure.

The comparison between FinRisk AgentNet and XGBoost is particularly instructive because XGBoost is widely recognized as one of the strongest baselines for structured fraud data. The fact that XGBoost achieved robust performance on both datasets confirms that the feature engineering and evaluation protocol used in this study are reasonable and competitive. However, the consistent margin by which FinRisk AgentNet outperformed XGBoost reveals an important limitation of even high-performing single-model learners: they are constrained by the information that can be encoded into a single tabular representation and optimized through a single loss function. XGBoost can model non-linear interactions among structured variables very effectively, but it still treats fraud primarily as a pattern-recognition problem in labeled historical data. In contrast, FinRisk AgentNet treats fraud as a multi-dimensional risk event that can be signaled through several channels simultaneously: statistical similarity to known fraud, deviation from expected behavioural norms, association with suspicious relational structures, semantic cues in narrative evidence, and contribution to

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

broader portfolio-level risk escalation. The gains observed over XGBoost therefore suggest that modern fraud detection is reaching a point where improvements from better tabular classification alone may be marginal, whereas more substantial gains can be achieved by integrating multiple analytical views of risk within a unified architecture.

The ablation study provides critical insight into *why* the full architecture performs better than its reduced variants. When the Anomaly Detection Agent was removed, recall fell from 0.9034 to 0.8615, while precision remained relatively high. This pattern strongly suggests that the anomaly layer contributes primarily to fraud sensitivity, especially for cases that are not fully represented by the supervised fraud model. Such a result is theoretically coherent. Fraud patterns evolve rapidly, and a purely supervised learner is inherently backward-looking because it can only detect fraud behaviours that resemble the labeled examples on which it was trained. The anomaly agent mitigates this limitation by surfacing unusual patterns that deviate from normal account, device, merchant, or transaction behaviour even when those patterns have not yet been labeled as fraud in historical data. In other words, the anomaly layer acts as a novelty detector that broadens the search space of suspicious events. The decline in recall when it is removed confirms that a non-trivial portion of detected fraud in the full model comes from this capacity to identify emerging or weakly labeled suspicious behaviour. This finding is particularly important for real-world deployment because financial fraud is adversarial and adaptive; systems that cannot respond to novel patterns will inevitably degrade as attackers change tactics.

One of the most important conceptual implications of the results is that FinRisk AgentNet behaves differently from a conventional ensemble, even though it also combines multiple signals. In a standard ensemble, multiple models are typically averaged, stacked, or voted together to improve predictive performance, but they are rarely designed around distinct analytical roles. In contrast, FinRisk AgentNet is built around functional specialization. The Fraud Classification Agent estimates the likelihood that a transaction resembles historical fraud. The Anomaly Agent detects deviations from normal behaviour. The Relational Agent evaluates cross-entity suspiciousness. The Semantic Agent interprets contextual evidence and improves decision quality in ambiguous cases. The Forecasting Agent assesses broader temporal and portfolio-level risk. The Orchestrator Agent then integrates these heterogeneous outputs into a single risk decision. The empirical results suggest that this task specialization is a major source of the architecture's effectiveness. The gains observed in the ablation study imply that each agent contributes a distinct form of information rather than merely duplicating the output of another model [21]. This makes the architecture especially well suited to financial risk management, where fraud review, anomaly surveillance, compliance reasoning, and risk forecasting are often separate but interdependent institutional processes. In this sense, FinRisk AgentNet is not just a better-performing classifier; it is a computational abstraction of how modern fraud operations actually function inside financial institutions.

The external validation results on IEEE-CIS are particularly important for assessing the robustness of the architecture. Public fraud research often suffers from over-reliance on a single benchmark, especially the ULB credit-card fraud dataset, which—although useful—contains PCA-transformed variables and limited contextual richness. By evaluating the architecture on IEEE-CIS, which includes a more heterogeneous and enterprise-like identity–transaction environment, the study tests whether the gains of the proposed system are robust to a more complex feature space. FinRisk AgentNet again achieved the best results, improving F1-score and AUPRC over XGBoost and all other baselines. The absolute performance values were somewhat lower than in the ULB setting, which is expected given the greater complexity of the dataset, but the relative advantage of the architecture remained strong. This consistency is a crucial result because it suggests that the benefits of the architecture are not tied to a single dataset structure. Instead, the architecture appears to derive value from the general principle of risk-signal fusion across heterogeneous analytical modalities. In other words, the study provides evidence that FinRisk AgentNet scales beyond compact academic fraud benchmarks and retains its advantage in richer transaction environments that more closely resemble real financial systems.

From the perspective of quantitative financial risk management, the findings support a broader conceptual shift: fraud detection should not be treated as an isolated classification task but as one component

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 3, Issue 1, 2025
https://jecir.com/

of an integrated institutional risk function. The portfolio-level forecasting results, together with the reduction in Expected Fraud Loss Score, show that FinRisk AgentNet is capable of linking micro-level fraud signals to macro-level risk consequences. This matters because financial institutions make decisions at multiple levels simultaneously. They must decide whether to block a transaction, whether to review a merchant, whether to increase reserves, whether to intensify monitoring in a segment, and whether an emerging fraud pattern is likely to affect loss trajectories over the next week or month. A system that only outputs transaction-level fraud probabilities does not fully support these decisions. FinRisk AgentNet moves toward a more comprehensive risk architecture by combining detection, prioritization, and forecasting. The study therefore suggests that the next generation of fraud analytics should be designed not merely around "fraud or not fraud" classification, but around dynamic financial risk intelligence.

Despite the strong performance of the proposed architecture, several limitations should be acknowledged. First, although the use of both ULB and IEEE-CIS improves external validity, public fraud datasets still do not fully replicate the data richness of live banking systems, particularly with respect to investigator notes, sanctions alerts, device graphs, and longitudinal customer narratives. The semantic/LLM component in this study was therefore evaluated in a semi-structured setting rather than with the full breadth of real internal case documentation. Second, the orchestration weights and thresholds were optimized within the experimental setting; in live deployment, these parameters would likely need periodic recalibration to account for concept drift, seasonality, product changes, and adversarial adaptation by fraudsters. Third, while the average latency of FinRisk AgentNet remained within a practical near-real-time range, production deployment at very high transaction volumes would require careful engineering of inference pipelines, selective invocation of heavier agents, and possibly asynchronous risk enrichment for lower-priority cases [22]. Fourth, although the forecasting module improved portfolio-level risk metrics, its benefits may become even clearer in longer time-series environments than those available in benchmark datasets. These limitations do not undermine the present findings, but they do clarify that the study should be interpreted as a strong architectural validation rather than a complete substitute for institution-specific deployment studies.

VI. CONCLUSION

This study proposed FinRisk AgentNet, a hybrid multi-agent LLM + ML architecture for agentic fraud risk detection and quantitative financial risk management, designed to address the growing limitations of conventional single-model fraud analytics in digitally intensive financial environments. The framework was developed on the premise that modern fraud is not merely a binary classification problem, but a multi-dimensional risk phenomenon shaped by transactional irregularities, evolving anomaly patterns, contextual evidence, relational dependencies, and broader portfolio-level financial exposure. To address this complexity, FinRisk AgentNet integrates supervised fraud classification, anomaly detection, relational risk reasoning, semantic interpretation through an LLM-based agent, and temporal financial risk forecasting within a coordinated orchestration layer that generates a unified fraud-risk decision.

The empirical evaluation demonstrated that the proposed architecture consistently outperformed conventional baseline models, including Logistic Regression, Random Forest, XGBoost, Isolation Forest, and Autoencoder-based approaches, across key fraud analytics metrics. In particular, FinRisk AgentNet achieved superior F1-score, ROC-AUC, and AUPRC on both the ULB and IEEE-CIS fraud detection benchmarks, while also delivering stronger operational outcomes in terms of Fraud Capture Rate, reduced False Alert Burden, and lower Expected Fraud Loss Score. The ablation analysis further confirmed that the performance gains were not incidental; rather, they emerged from the complementary roles of the anomaly, semantic, relational, and forecasting agents, each of which contributed distinct and measurable value to the final decision process.

Overall, the findings indicate that financial fraud detection can be significantly strengthened when it is embedded within a broader agentic financial risk management architecture rather than treated as an isolated prediction task. FinRisk AgentNet therefore contributes both methodologically and practically by demonstrating how machine learning, large language models, and multi-agent orchestration can be combined

into a unified and explainable framework for next-generation fraud intelligence. Future research should extend this architecture to real institutional multi-modal datasets, adaptive online learning environments, and governance-aware deployment settings to further advance trustworthy AI-driven financial risk management.

REFERENCES

1. Rahman, M., Bhuiyan, A., Islam, M. S., Laskar, M. T. R., Mahbub, R., Masry, A., ... & Hoque, E. (2025). Llm-based data science agents: A survey of capabilities, challenges, and future directions. *arXiv preprint arXiv:2510.04023*.
2. Spears, T., Hansen, K. B., Xu, R., & Millo, Y. (2025). Governing synthetic data in the financial sector. *Finance and Society*, 1-17.
3. Joshi, S. (2025). Comprehensive review of artificial general intelligence (AGI): Applications in business and finance.
4. Karras, A., Theodorakopoulos, L., Karras, C., Krimpas, G. A., Giannaros, A., & Bakalis, C. P. (2025). LLM-Driven Big Data Management Across Digital Governance, Marketing, and Accounting: A Spark-Orchestrated Framework. *Algorithms*, *18*(12), 791.
5. Lopez, E. (2024). *Towards eXplainable Artificial Intelligence (XAI) in cybersecurity* (Doctoral dissertation).
6. Ribeiro, R. N. (2025). *Exploring the Feasibility of Using Large Language Models for Dark Web Threat Intelligence in Security and Defence* (Master's thesis, Universidade do Porto (Portugal)).
7. Wang, Q., Wang, T., Tang, Z., Li, Q., Chen, N., Liang, J., & He, B. (2025, July). MegaAgent: A large-scale autonomous LLM-based multi-agent system without predefined SOPs. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 4998-5036).
8. de la Roche, M., Voloder, E., Banerjee, A., Guerra, C., Cataldo Dell'Accio, D., Budris, F., ... & Sedej, T. (2024). Report on artificial intelligence and blockchain convergences. *Available at SSRN 5023415*.
9. Li, H., Huang, S., & Park, J. K. (2025). A Human-Centered Review of Large Language Models for Online Scam Detection. *Available at SSRN 6338940*.
10. Steindl-Haselbauer, K. (2025). *Potentials, risks and implications of AI-driven Strategic Foresight in disruptive environments-A framework for companies to exploit new opportunities and remain resilient* (Doctoral dissertation, Technische Universität Wien).
11. Adabara, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Venkateswarlu, M. (2025). A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense. *F1000Research*, *14*, 843.
12. Lei, M., Liu, Q., Li, K., Zhao, J., & Tian, Z. THE STUDY OF FEASIBILITY AND CHALLENGES ON SILICON-BASED LIFEFORMS BASED ON LARGE LANGUAGE MODEL.
13. Raza, S., Qureshi, R., Zahid, A., Muneer, A., Zafar, A., Kamawal, S., ... & Shoman, M. (2025). Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650*.
14. Pamisetty, V. (2023). Leveraging AI, Big Data, and Cloud Computing for Enhanced Tax Compliance, Fraud Detection, and Fiscal Impact Analysis in Government Financial Management. *Fraud Detection, and Fiscal Impact Analysis in Government Financial Management (December 15, 2023)*.
15. Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. (2025). Evaluating trustworthiness in AI: Risks, metrics, and applications across industries. *Electronics*, *14*(13), 2717.
16. Kolla, S. H. (2023). Deep Learning–Driven Retrieval-Augmented Generation for Enterprise ITSM Automation: A Governance-Aligned Large Language Model Architecture. *Journal of Computational Analysis and Applications*, *31*(4).
17. Abdallah, A. A., Aslan, H. K., Abdallah, M. S., Cho, Y. I., & Azer, M. A. (2025). Enhancing Cryptocurrency Security: Leveraging Embeddings and Large Language Models for Creating Cryptocurrency Security Expert Systems. *Symmetry*, *17*(4), 496.
18. Motamary, S. (2024). Transforming customer experience in telecom: Agentic AI-driven BSS solutions for hyper-personalized service delivery. *Available at SSRN 5240126*.

19. Vlachos, I., & Reddy, P. G. (2025). Machine learning in supply chain management: systematic literature review and future research agenda. *International Journal of Production Research*, *63*(16), 5987-6016.
20. Moussaoui, J. E., Kmiti, M., El Gholami, K., & Maleh, Y. (2025). A systematic review on hybrid AI models integrating machine learning and federated learning. *Journal of Cybersecurity and Privacy*, *5*(3), 41.
21. Wicker, M., Szpruch, L., & Mørk, S. (2025). Move Fast without Breaking the Bank Model Risk Management of GenAI workflows.
22. IKE, P. C., DANIEL, U. C., ADEOYE, M., AMAECHI, M. I., ODESOLA, O. O., & OGUNSAKIN, M. O. (2025). COGNITIVE CYBER DEFENSE SYSTEMS: LEVERAGING ARTIFICIAL INTELLIGENCE FOR AUTONOMOUS THREAT PREDICTION AND RESPONSE. *International Journal of Nature and Science Advance Research*.