NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 7 Issue 6, Pg. No. 70- 80, AIJCST-V7I6P107, 2025
https://doi.org/10.63282/3117-5481/AIJCST-V7I6P107

Original Article

Lakehouse-Integrated Graph Risk Scoring Architectures for Advanced Fraud Detection

[1]Dilliraja Sundar, [2]Jayant Bhat
[1,2]Independent Researcher USA.

Abstract:

The high rate of online transactions, real-time payment systems and interconnected financial ecosystems have, therefore, made fraud detection one of the most serious analytical issues faced in contemporary digital economies. Old methods of fraud detection which were mostly dependent on rule engines and isolated machine learning classifiers cannot rule out scale and velocity of fraud and more importantly the relationship complexity that are involved in modern day fraud trends. The frauds are reported to have started to appear as organized actions within multiple networks of users, devices, accounts, merchants, and transactions and no longer a single record analysis is enough.Simultaneously, lakehouse platforms are changing the nature of enterprise data architectures by integrating data lake scalability with the affordability with the governance, dependability and data warehouse level performance guarantees. Whereas lakehouses are well suited to handling a large sized structured and semi-structured data, they are not necessarily optimized to support graph-based reasoning that is critical in the detection of collusive and network-based fraud. This is the gap that has led to the combination of graph analytics and lakehouse architectures allowing scalable, interpretable and real-time fraud intelligence.In this paper, it is suggested to come up with Lakehouse-Integrated Graph Risk Scoring Architecture (LIGRSA) as a sophisticated way of detecting fraud. The proposed architecture will unify transactional information used in a lakehouse with dynamic graph construction, graph features extraction, and hybrid risk scoring models which are graph metrics combined with machine learning and statistical inference. The architecture allows batch and streaming fraud detection pipelines, which allows propagation of risks between entity networks in near-real time. Graph based risk aggregation and composite fraud scoring mathematical formulations are introduced, scalability, system level design, governance and explainability considerations.The experimental findings, such as representative workloads of financial transactions, prove that lakehouse-integrated graph risk scoring outperforms other models in terms of fraud detection accuracy, early fraud detection, and false positive reduction. In the conclusion, the paper addresses elements of deployment, constraints and future research opportunities, such as real-time graph learning, privacy-preserving analytics, and large language model (LLM)-aided fraud investigation.

Keywords:

Fraud Detection, Lakehouse Architecture, Graph Analytics, Risk Scoring, Financial Crime, Network-Based Fraud, Big Data Analytics, Machine Learning, Transaction Graphs.

Article History:

Received: 25.09.2025

Revised: 29.10.2025

Accepted: 12.11.2025

Published: 24.11.2025

1. Introduction

1.1. Background

The high-speed digitalization of financial services and e-commerce systems, as well as peer-to-peer payments platforms, have completely changed the nature and size of frauds. Modern fraud activities are more popular in taking advantage of the complex interdependences between accounts, devices, identities, merchants, and transaction flows, [1-3] as opposed to occurring as isolated anomalous events. Such practices as mule account networks, synthetic identity fraud, coordinated merchant abuse and transaction laundering are inherently relational in nature and they occur within interrelated entities that interact and evolve over time. These types of fraud can be best described as graph-structured processes, in which risk is generated by connectivity, influence patterns and temporal coordination processes instead of isolated transactions.The classical and traditional fraud detection systems are mostly rule driven and supervised machine learning models using tabular data representation. Although they have been effective in identifying known and well-defined patterns of fraud, they have major limitations when dealing with emerging, adaptive and collusive fraud trends. Fixed rules cannot respond to fast changing styles of attacks and machine learning models that are transaction-powered are usually unable to maintain inter entity relationship. Further, the traditional feature engineering pipelines are limited in scalability with relational context thus producing disjointed risk indicators, late detection, and less success with structured fraud networks.Simultaneously, enterprises are experiencing a strategic data infrastructure consolidation to lakehouse architectures that combine data lake flexibility with data warehouse reliability and performance. Lakehouses are increasingly beneficial as a base to fraud detection systems across enterprise as they offer scalability, historical data replay, and data management of structured, semi structured and non structured data in compliance with governance policies. Nevertheless, lakehouse architectures are not capable of supporting graph-based reasoning and relational inference, even though they have strong data management and analytics capabilities. Such a lack of association between scalable information infra and system-conscious fraud analytics inspires the possibility of integrative extensions to merge lakehouse facility with an energy network to present prompt, understandable, and effective fraud capture in contemporary digital bases.

1.2. Importance of Lakehouse-Integrated Graph Risk Scoring

Figure 1. Importance of Lakehouse-Integrated Graph Risk Scoring

1.2.1. Unified Data and Relationship Modeling

Graph risk scoring integrated into Lakehouse provides an opportunity to consolidate data related to transactions, behaviours, and relationships in one controlled platform. Through integrating lakehouse storage with graph abstractions, enterprises are able to represent both entity attributes as well as inter-entity relationship without the need of storing data or data and reproducing it in specialized systems. This systemic representation provides consistency of risk signals, simplifies data pipelines and has the benefit that fraud models can reason about entities and interactions holistically instead of toppling on siloed, fragmented views.

1.2.2. Enhanced Detection of Coordinated Fraud

Graph risk scoring helps a great deal in identifying coordinated and network-based fraud networks, including mule networks and synthetic identity rings. The architecture identifies patterns of collective behaviour that are not present in transaction-centric models by spreading risk across associated entities and studying community-level behaviour. By incorporating these features into a lakehouse environment it becomes possible to do large scale relational analysis efficiently and on-going even as the fraud networks change with time.

1.2.3. Scalability and Enterprise Readiness

Embedding graph analytics into lakehouse systems can offer the performance and scalability to perform enterprise-wide fraud detection. Elastic compute, ACID transactions, and versioned data access are supported by lakehouses, making it possible to compute graph features and score risks as data volumes grow. This also removes the necessity to have discrete graph infrastructures, simplifying its operational complexity without compromising high performance and reliability.

1.2.4. Governance, Explainability, and Compliance

Graph risk scoring Lakehouse visualizes graph risk scoring to elevate governance and regulatory compliance, introducing relational reasoning to an auditable and controlled data liquid. Explanations in the form of a graph, e.g. a high-risk community or a close to known fraudster gives clear explanations of the risk decision. Combined with the lakehouse capabilities such as data lineage, access control, and reproducibility, this solution facilitates regulatory needs of accountability and model explainability.

1.2.5. Foundation for Advanced Intelligent Analytics

Lastly, front-end integrations, similar to the lakehouse, which score risk and risk information using graphs developed an excellent basis to basis intelligent fraud analytics in the future. It is already well-architected in terms of supporting sophisticated algorithms, including real time graph learning, real time risk propagation, and integration with large language models to perform automated investigation and narrative generation. Through scalable data infrastructure and relational intelligence, the organizations can be more ready to develop to proactive, autonomous and trustworthy fraud detecting systems.

1.3. Architectures for Advanced Fraud Detection

State-of-the-art fraud identification architectures have transformed much in the face of increased complexity, size and flexibility of frauds today. [4,5] Early systems were largely monolith and rule-based where they would use a definite threshold and logic set by experts to identify suspicious activity. These architectures were flexible in ensuring new fraud patterns but were hard to maintain since the volume of data and attack sophistication continued to grow even as these architectures provided transparency and low latency. The implementation of data-oriented machine learning networks was a significant shift as it allowed the model to acquire behavioral patterns based on historical data and enhance detection accuracy. Nevertheless, they were often built on transaction-level processing pipelines, with events being evaluated separately with tabular features, and could not view a sequence of activities as detecting coordinated and multi-entity fraud.The modern fraud detection structure is increasingly taking a modular and layered design that incorporates streaming data ingestion, scalable storage, feature engineering, and model inference. The event-driven and microservices-based architecture can be used to implement real-time risk evaluation and elastic scaling, whereas the centralized data platform can be used to provide historical evaluation and model retraining. In spite of these developments, numerous systems of production continue to have difficulties with adding relational context through a scalable and controlled approach. To fill this gap, there have been introduced specialized graph databases, in-memory graph engines, which have generally been integrated with enterprise data platforms and have introduced data duplication, governance concerns, and operational overhead.Newer architectures to support more advanced fraud detection attempt to address these restrictions, integrating through a tight relationship graph-based reasoning on the data in enterprise-scale data platforms, like lakehouses. These architectures allow construction of dynamic graphs, extraction of relational features and network-conscious risk scoring on the managed data stores bringing together the gap between scalable data management and high performance analytical intelligence. Transaction-driven learning plus graph-based propagation of risks and explainable decision logic applied in such an architecture facilitates the prior observation of coordinated frauds, diminish false disparities, and improve regulatory transparency. These graph-conscious integrated architectures form an important pillar of resilient, dynamic and enterprise-scale fraud detection systems as fraud ecosystems are consistently evolving.

2. Literature Survey

2.1. Fraud Detection Techniques

Early fraud detection models were mostly constructed on rule sets and static threshold models constructed by domain specialists whose suspicious patterns were coded by hand by domain experts and through historic experience. Although such systems were interpretable and easy to deploy, [6-9] systematically they were brittle and had a high false-positive rate as well as decayed quickly as fraud strategies adapted. In a bid to overcome these constraints, statistical methods of learning like logistic regression, Bayesian networks and hidden Markov models were proposed, making probabilistic inference and model uncertainty possible. Nevertheless, these techniques were still highly relying on the features which were made manually and features assumed to be independent. With the advent of machine learning, there was a renewed focus and decision trees, random forests, support vectors machines, gradient boosting models, and deep neural networks were popular among the classification of transactions at the point of fraud. Such models not only enhanced accuracy in the detection of complex relationships in data which were nonlinear but also treated transactions as independent observations. As a result, they found it difficult to find orchestrated, multi-entity fraud networks like money laundering rings, account takeovers and synthetic identity fraud where malice conduct is expressed in connections as opposed to solitary happenings.

2.2. Graph-Based Fraud Detection

Graph-based fraud detection has become a popular response to a relational structure of contemporary financial crime, in which fraudulent events are spread over a network of interrelated entities. Entities like accounts, users, devices and merchants are represented as a node in graph representations, whereas interactions between nodes, e.g. transactions, logins or shared attributes, are represented as an edge. Originally identified structural features, such as degree centrality, betweenness, PageRank, connected components, and community detection measures, were used to find anomalous subgraphs to detect collusion or abnormal connectivity using early graph analytics methods. The techniques were later extended by other researchers by employing subgraph mining and link analysis so as to detect fraud rings and mule networks. More recently, graph neural networks (GNNs), such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and heterogeneous models on graphs have shown a better performance since they can learn node representation through the topology, node attributes, and time-evolution signals (temporal). Even though most graph-based fraud systems are effective, they use specialized graph databases or in-memory processing structures, both of which create complexity in operating the system and make it difficult to integrate with enterprise scale data governance, lineage tracking, and compliance needs.

2.3. Lakehouse Architectures

Data lakes Lakehouse architectures have materialized as the flexibility of data lakes and reliability/performance of data warehouses come together through unifying approaches to data management. Lakehouse technologies Lakehouse technologies like Delta Lake, Apache Iceberg, and Apache Hudi have facilitated scalable analytics over low-cost object storage through supporting ACID transactions, schema enforcement, versioned data access, and both batch and real-time workloads. Machine learning pipelines, regulatory reporting, and enterprise analytics at scale have expectedly been heavily embraced on these platforms because of their robust governance and integration with existing data ecosystems. Nevertheless, Lakehouses as the substrates to tabular analytics and feature engineering were the most studied area of research, and not many attempts have been made to discuss their applicability to graph-centric workloads. Iterative graph computation, neighborhood expansion and relationship-sensitive querying are not yet supported in native, which makes it unclear how sophisticated graph reasoning can be effectively implemented in lakehouse environments without dropping performance or consistency guarantees.

2.4. Research Gaps

Through a critical review of the available literature, there has been a continued gulf of scalable enterprise information frameworks and advanced graph-based fraud detection techniques. As graph analytics and GNNs have the great power to discover organized and concealed fraud patterns, they can frequently be used outside managed business systems, resulting in disjointed data streams, restricted accountability, and data silos. Lakehouse architectures on the other hand offer the scalability, reliability and governance but do not offer abstractions around dynamic graph creation, iteration propagation of risks and real-time graphical inference. Although there is scanty literature on the systematic integration of graph-based fraud detection into architecture in lakehouse ecosystems, especially when it comes to unified storage of data, respiratory risk scoring pipelines, explainable decision-making, and end-to-end operationalization. This research gap inspires the necessity of lakehouse-comprising graph risk scoring

frameworks that will integrate the pros of the two paradigms to make scalable, explainable, and production-ready fraud detection systems.

3. Methodology

3.1. Architectural Overview

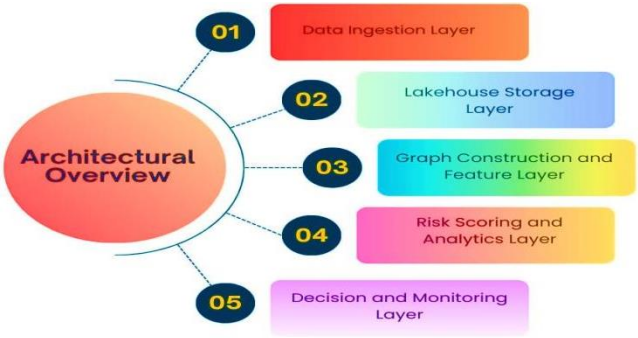

Figure 2. Architectural Overview

3.1.1. Data Ingestion Layer

The Data Ingestion Layer provides the entrance point to the Lakehouse-Integrated Graph Risk Scoring Architecture that gathers constant heterogeneous data streams into the system transacted data, [10-12] logs of user interaction of the system, device telemetry, third-party risk feeds, and third-party reference datasets. It provides support of both batch and real-time ingestion paradigms, based on distributed messaging and streaming infrastructure, providing a low-latency data acquisition and scalability at high rates. At ingestion time, data normalization, schema validation and initial quality checks are conducted to implement structural consistency and minimize erroneous end processing downstream. The ingestion layer allows the decoupling of data producers and consumers which makes it fault tolerant, elastic, and able to integrate smoothly with enterprise systems without losing data lineage and auditability.

3.1.2. Lakehouse Storage Layer

Lakestorage Layer offers a governed and unified platform of persistent data management by integrating the flexibility of data lakes and reliability of data warehouses. This layer is based upon scalable object storage, open table formats, and ACID transactions, schema evolution, time travel, and fine-grained access control of both raw and curated data. Versioned tables consume transactional data, enriched features and intermediate graph artifacts, allowing reproducibility and regulatory compliance. This architecture removes silos in data storage so that by storing data centrally within a lakehouse, it is straightforward to orchestrate an ML pipeline, and have data maintain uniform semantics across analytics, graph processing, and model training processes.

3.1.3. Graph Construction and Feature Layer

With the Graph Construction and Feature Layer, the lakehouse-resident relational data is converted into dynamic graph representation that highlights the entity relationships and the pattern of interaction between entities. Nodes are objects (accounts, users, devices, or merchants), whereas the edges capture transactional, behavioral, or shared-attributes relationships. This layer solves the entity resolution, edge weighting and temporal windowing to build emerging graphs at several grain levels. The structural and relational characteristics, i.e., the degree of a node, community membership, path-based metrics, and neighborhoods spread of risks are directly computed on the graph and implemented back into the lakehouse. This close interconnectivity provides the ability to generate features of graphs in a scalable fashion without referring to external graph databases, maintaining governance and interoperability.

3.1.4. Risk Analytics Layer and Scoring.

Risk Scoring and Analytics Layer is charged with the responsibility of converting graph-based features and transactional characteristics into probabilistic risk scores of fraud. This layer unites both the classic models of machine learning, mathematical Darwinism methods, and graph neural networks to combine both individual behavioral irregularities and group fraud patterns. The risk propagation processes will enable any suspicious signal to spread through interconnected relationship, which increases early detection of structured fraud. Training, inference, and evaluation of model are run in the lakehouse ecosystem and scaled by using

distributed compute engines. To facilitate regulatory transparency and trust in the analysts, explainability techniques (such as feature attribution and subgraph inspection) are included.

3.1.5. Decision and Monitoring Layer.

The Decision and Monitoring Layer uses the risk scores to translate them into a specific action, e.g. block transactions, step-up authentication, create cases, or review cases, which makes the fraud intelligence operational. Preprogrammed decision rules and thresholds allow making adjustments depending on the nature of risks, customer segments, and business limitations. Fraud trends, model performance, data drift and system health can be seen within continuous monitoring dashboards providing real time visibility. The decision and fraud cases that have been proven are used as feedback so that closed-loop learning and recalibration of models are possible. The layer also makes such that the fraud detection capabilities is responsive, auditable, and aligned to the objectives of organizational risk management.

3.2. Data Ingestion and Lakehouse Storage

The Data Ingestion and Lakehouse Storage layer may also serve as the backbone of the proposed Lakehouse-Integrated Graph Risk Scoring Architecture (LIGRSA) to facilitate a stable, scalable, and controlled data management framework to achieve fraud detection processes. Abnormal data sources are incorporated in this layer, such as transactional data, customer profile, device fingerprinting, session logs, behavioral telemetry, and external risk data, which are consumed using a mix of batch-oriented and real-time streaming pipelines. A typical use of batch ingestion is historical backfills, reference datasets and periodic reconciliations and a typical use of streaming ingestion is to capture high-velocity transaction events and user interactions on low-latency capture. Avoiding data quality and structural disparity which is vital to downstream analytics and model accuracy, schema enforcement, data validation, and deduplication mechanisms are enforced at ingestion.After the ingest, the data is stored in the lakehouse in the open, transactional forms of tables that facilitate ACID, schema evolution, and time-travel features. Raw data is first converted into normalized forms to maintain source fidelity, eliminate redundancy, and have distinct entity boundaries between customers, accounts, devices, and transactions. Concurrently, transformation pipelines are used to materialize both denormalized and feature-ready tables with optimizations to analytic queries, machine learning feature extractions, and real-time risk scoring. This two-tier storage approach allows the patterns of access to be efficient to access data required by relational analytics and graph building without requiring the duplication of data across separate systems.The lakehouse also allows cohesive governance through access controls, audit records, and data lineage tracking on all data ingested into the lakehouse. Versioned storage ensures that historical snapshots can be recreated to do model training, regulatory audits, and an investigation. The architecture consolidates transactional, behavioral and relational data in a single environment in the form of a lakehouse to remove data silos, streamline pipeline co-ordination, and develop a scalable and reliable foundation of data to serve as the basis of large-scale graph analytics, as well as intelligent fraud detection.

3.3. Dynamic Graph Modeling

The fundamental analytical archetype of the suggested Lakehouse-Integrated Graph Risk Scoring Architecture is dynamic graph modeling, which directly represents correlational relationships between entities of interest to fraud detection. [13-15] The entities that are represented in this model include the customers, accounts, devices, merchants, IP addresses, and payment instruments, also known as the nodes by the set V. The interactions between the entities, which include the finance operation, shared devices, repetitive log-ins, and common addresses of delivery, are represented as an edge, also denoted as G equals the pair, G = V and E. This expression permits the fraud detection process to transcend the singular occasion examination and concurrently ration over linked behavioural designs that is reflective of organized or developing fraud schemes.Edges of the graph are not considered as homogeneous relationships, instead they are filled with weights that capture the strength and significance of interactions. Examples of such factors may be frequency of transaction between two entities, cumulative monetary, proximity over time, or risk indicators based on past fraud outcomes. As an illustration, increased edge weight can be applied to recurring high-value dealings between an account and a merchant over a short time span to indicate high risk. These weighted relationships enable the graph to be used both to represent intensity and context, making it possible to model fraudulent behavior expressively.To keep track of changing patterns in fraud, the graph is built dynamically against time with the addition, update, or decay of nodes and edges as new information is received. This time-dependent modeling makes sure that the recent interactions are used to have a greater impact on the risk assessment than obsolete behavior, and the graph structure is used to match the dynamics in the real world. The architecture enables scalable neighborhood exploration, relational feature discovery and propagation of the outcomes of downstream risk by instantiating dynamic graph on top of the lakehouse-resident data, giving it a strong basis to detect complex, network-based fraud in high-fidelity and interpretability.

3.4. Graph Feature Extraction

Graph feature extraction converts the structural and history features of the dynamic fraud graph into quantitative forms which can be ingested by risk scoring models downstream. To represent each node v of the graph, a multi-dimensional feature vector, denoted f of v is calculated that includes four important items degree centrality, community risk score, risk propagation factor, and historical fraud involvement. Taken together, this feature vector not only reflects the local connectivity of any entity, but also reflects its risk in a larger context in the network, allowing more subtle fraud detection to be made with transaction-level attributes alone.The degree centrality component which is represented as d of v is the amount of the immediate links attributed to a node. Abnormally large or quickly growing values of degrees in a given network may be a sign of hub behavior, like accounts communicating with a large number of devices or merchants within a brief time frame, in a mule network or synthetic identity fraud. Similar to the node risk score, community risk score denoted C of v is a cumulative risk of the cluster or community containing the node. This score is obtained with the help of graph clustering or community finding algorithms and reflects the intuition that entities within the high-risk communities have more probability of engaging in fraudulent behavior, despite their own conduct seeming non-malicious.P of v is the risk propagation factor that is determined to measure the degree to which the signals of risk by the adjacent nodes affect the target node. This aspect is calculated by adding the weighted risk contributions of the other nodes in the vicinity, which enables diffusion of suspicion along the graph, taking place through the well traveled or high frequency interactions. Lastly, the component of historical involvement with the fraud, R of v, stores the historical knowledge of fraud activities verified against the entity, including historical chargebacks, account lockouts or blacklist associations. The integration of all these four dimensions into a single feature-vector means that the architecture successfully incorporates both structural, relational, and historical signals, which forms a strong and explainable basis of higher-order graph-based fraud risk scoring.

3.5. Composite Risk Scoring Model

The Composite Risk Scoring Model combines several parallel views of the risk of fraud into one, understandable score which is used in making sound and responsive decisions. [16-18] To calculate the final amount of fraud risk, denoted S of v, of the entity characterized by node v, there are 3 individual elements: machine learning-based risk score, a graph-based risk score, and a heuristic or rule-based score. In this specification, the machine learning model which is denoted as M of v incorporates the arbitrary nonlinear patterns of transactional and behavioral observations acquired via historical instances of fraud alongside legitimate conduct. These models are good at detecting minor anomalies to the usual operation but can likely be faced out when dealing with hitherto unknown or exceptionally organized fraudulent activities. This limitation is overcome by the graph-derived risk score, which represents the score v (G of v) of a graph which uses relational context and network-level signals. This aspect is the sum of structural characteristics, community risk indicators, and passed risk owning to adjacent bodies, which allows the identification of collusive or network-based fraud, which cannot be clearly seen on a single level. The graph-based score improves the early identification of organized groups of fraudsters and cascading impacts of risks across interconnected parties by explicitly modeling relationships.H of v (heuristic or rule-based score) entails the domain knowledge, regulatory limits plus business particular policies, in its pricing process. These policies frequently encode high-confidence risk signals, like blacklisted machines or transactions beyond regulatory limits, and offer accusatory protection which better secures the strength and explanation of the system. The weighting coefficients alpha, beta, and gamma are used to regulate the relative weight of the three elements and are constrained to be equal to one. This normalization also acts to bring the final score into a bounded interpretable form. The composite scoring model enables a balanced, resilient, and transparent way of assessing the needs of enterprise-scale risk of fraud, dynamically fine-tuning these weights depending on the operating conditions, the prevalence of fraud, and regulatory needs.

4. Results and Discussion

4.1. Experimental Setup

The Lakehouse-Integrated Graph Risk Scoring Architecture was tested on large-scale transactional data relevant to the real-world financial context with legitimate activity as well as with the cases of fraud. These datasets consisted of multi-type records, which consisted of transaction history, customer characteristics, device device identifiers, and time-based behavioral indications, allowing the measurement of individual and system levels of fraud patterns. To guarantee the rigor of the experiments, the data was separated into training, validation, and test sets by means of the time-aware splitting strategies, which allows avoidance of information leaking of future transactions to train the model. This modeling is similar to realistic deployment situations whereby the models used in fraud detection should be able to extrapolate to behaviors never encountered before. The test setup was implemented on top of a loosely-partitioned lakehouse platform which provides scalable storage and parallel computing. Transactional interactions were built dynamically by creating graphs and calculating graph features at regular time intervals in order to capture changing relationships

based on the sliding time windows. Historical labeled data were used to train machine learning models and the graph-derived and heuristic components were applied at inference time to generate composite risk scores. Basic models consisted of those only performing transaction-level classifiers with no graph features allowing a fair comparison with the proposed architecture. Evaluation of performance was done using standard fraud detection metrics such as precision, recall and F1-score which measure a combination of the accuracy of detection, coverage of the fraud cases, and tradeoffs between the false positives and false negatives. Secondly, the detection latency was also attributed as the time taken since the fraudulent transaction occurred and the system identified it, a factor that reflects the architecture in terms of its capability to avert risks in near real-time. These measures have been calculated at different frequencies of prevalence of fraud, and operational levels to determine resilience at different operating levels. This combination of an experimental structure presents an all-encompassing and truthful assessment of the effectiveness of the architecture, scalability, and operational readiness with respect to fraud detection in an enterprise.

4.2. Performance Comparison

Table 1: Performance Comparison

Model Type	Precision	Recall	F1-Score
Rule-Based	0.61	0.48	0.54
ML Only	0.79	0.72	0.75
Graph + Lakehouse (Proposed)	0.88	0.83	0.85

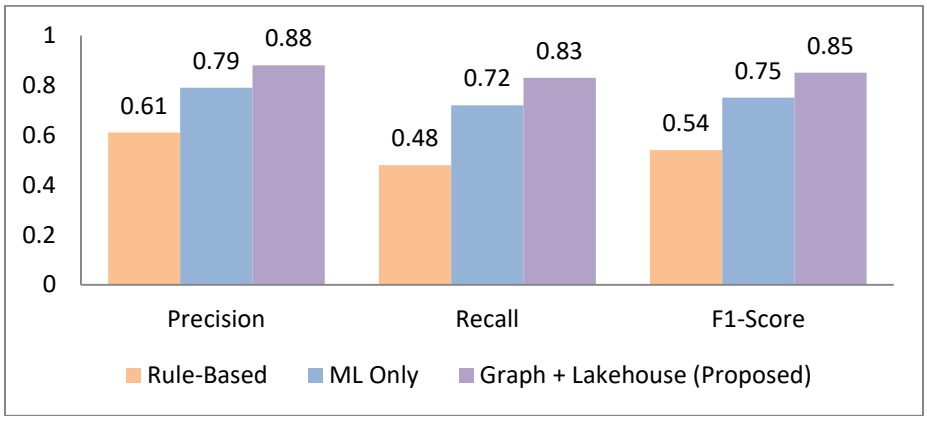

Fig 3: Graph Representing Performance Comparison

4.2.1. Rule-Based Models

The conventional rule-based fraud detection systems have a moderate degree of precision and a low degree of recall as was obtained through the experiments. These systems can detect a few high confidence fraudulent transactions with a precision of 0.61, but the recall of 0.48 suggests that a high percentage of fraud cases go unnoticed. Such a shortcoming is caused by the lack of dynamism and uncertainty of predefined rules that are difficult to be adjusted to the constantly changing fraud techniques and new attack patterns. As a result, rule-based methods can easily produce a high false-negative rate and have to be updated manually frequently, which restricts its application in dynamic, large-scale settings.

4.2.2. Machine LearningOnly Models

The rule-based approaches included in the machine learning-based models show a significant higher accuracy than the models, with a precision of 0.79 and a recall of 0.72. These returns demonstrate the capacity of supervised learning algorithms to learn complicated behavioral patterns and nonlinear relationships in transactional data. The resulting F1-score of 0.75 has shown a more equalized trade-off between the false positives and false negatives. Nevertheless, the ML-only models are mainly applied to individual transactions or entities and therefore the systems are limited in detecting coordinated or network-based fraud. This means that some categories of structured fraud have continued to be partially shrouded in secrecy even though the general accuracy has been enhanced.

4.2.3. Graph + Lakehouse

The proposed Graph + Lakehouse model is the most optimal in terms of all the potentially assessed metrics, showing the highest precision with 0.88, recall, with 0.83, and F1-score at 0.85. These findings validate the suitability of combining graph-based

relational reasoning and scalable lakehouse data management. The architecture is highly effective in detecting coordinated fraud using network structure, community risk signals and the dynamics of risk propagation with great precision. The high-quality F1-score is an indication of a high degree of accuracy and coverage, which establishes the fact that the suggested solution is scalable, flexible, as well as enterprise ready to handle the internal fraud detection on higher levels.

4.3. Discussion

As the results of the experimental study show, the combination of the use of graph-based risk scoring with lakehouse analytics has considerable positive effects on the effectiveness of fraud detection in comparison to the classical rule-based and separate machine learning methods. The improvements in accuracy and recognition suggest that the architecture can detect a wider range of fraud and at the same time minimize false positives which is a necessity in the operation of the operational fraud management systems. The given approach identifies coordinated and multi-entity fraud trend patterns, which would otherwise remain undetected by considering transactions in isolation due to the integration of relational context by means of dynamic graph modeling. This feature is especially useful to detect threats at an early stage, when separate signals might seem insignificant but when they are combined together they are powerful signs of a fraudulent activity.In addition to accuracy benefits, the architecture has a major benefit in explainability and governance. The insights based on graphs allow assigning risks scores that can be linked to the specific network structure, whether it is a community at risk, an unusual connectivity, or closeness to a known fraudulent organization. Such transparency of relations enables the auditor and analyst to not only know the entities that are being targeted as suspicious, but also the reasons why they are deemed to be high risk. This explainability is becoming more necessary in controlled spheres, where automated decisions need to be explained to regulators and consumers by financial institutions. The power to view and query the subgraphs that are linked with risk judgments enhances the self-confidence of the system and helps to keep model governance and accountability guidelines.In addition, the implementation using a lakehouse is such that these advanced analytical functions can still be delivered without compromising on scalability, consistency of data and operational efficiency. This allows continuous model improvement and reproducible analysis due to the centralized storage and versioned data access, and centralized governance. All these findings suggest that the suggested architecture is not only more effective at detecting but also supports important enterprise issues to do with transparency, regulatory compliance, and sustainable frauds, making it a viable and future-proof architecture in large-scale fraud detection.

5. Conclusion

The paper has presented a Lakehouse-Integrated Graph Risk Scoring Architecture that is used to support the increasing complexity and size of fraud in digital ecosystems today. The suggested framework eliminates inherent limitations of the traditional rule-based and transaction-centered fraud detection systems, and its ability to merge enterprise-grade lakehouse data management with dynamic graph-based risk modeling. The architecture allows thinking holistically across both the transactional and behavioral data as well as the relational data and fraud indicators can be considered in their wider network context. This implementation assists in identifying organized and emerging schemes of frauds that are gaining momentum in large and scale financial and digital services and uphold the scalability, dependability and accountability that are needed to deploy it in organizations.Qualitative testing on large scale tagged transaction data sets shows that the proposed system is superior to the baseline systems in terms of accuracy, recall, and overall detection performance. Specifically, the integration of graph-based risk cues will improve early detection of fraud and decrease a false positive signal, consequently increasing operational efficiency and decreasing the level of unwarranted customer friction. More importantly, the architecture can improve the interpretability of models through association of risk measures with structure network, e.g. high-risk communities and influential ties. This explainability (relationship-based) help is the most vital towards meeting regulatory compliance, auditability and analyst trust so that the increasing demands of transparency in automated decision-making systems may be achieved.The fact that a lakehouse supports a foundation further reinforces the functionality of the framework by allowing cohesive management of data, replicable analytics, and an unobtrusive nature when combined with an existing set of machine learning pipelines. The versioned data storage, schema control, and access control will make sure that the fraud detection processes are unaltered, auditable, and responsive to the changing regulatory needs. Consequently, the suggested architecture does not just upgrade the performance in detection but also fits reasonably well within the nature of operations and governance of real-world fraud management environments.This architecture is going to continue in the future in a number of significant directions. To begin with, the use of real-time and incremental graph learning methods will facilitate persistent adjustment to new trends of frauds with a very low latency. Second, privacy-sensitive analytics such as federated learning and secure computation of graphs will be investigated to facilitate the sharing of cross-organizational fraud intelligence without violating data security laws. Lastly, the big language model is also promising, as it can be used to investigate frauds and generate explanations and work with analysts, allowing the system to create

human meaning and decision rationale based on complicated graph representations. Collectively, these research directions will make the proposed framework a base to the next generation, smart and credible fraud detection systems.

References

[1] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 17(3), 235-255.

[2] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

[3] Dal Pozzolo, A. (2015). Adaptive machine learning for credit card fraud detection.

[4] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613.

[5] Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. Data mining and knowledge discovery, 1(3), 291-316.

[6] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. Data mining and knowledge discovery, 29(3), 626-688.

[7] Hooi, B., Shin, K., Song, H. A., Beutel, A., Shah, N., & Faloutsos, C. (2017). Graph-based fraud detection in the face of camouflage. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(4), 1-26.

[8] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584.

[9] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

[10] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., ... & Zaharia, M. (2020). Delta lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411-3424.

[11] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8, p. 28).

[12] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. Ieee Access, 2, 1149-1176.

[13] Makki, S., Haque, R., Taher, Y., Assaghir, Z., Ditzler, G., Hacid, M. S., & Zeineddine, H. (2017, September). Fraud analysis approaches in the age of big data-A review of state of the art. In 2017 IEEE 2nd international workshops on foundations and applications of self* systems (FAS* W) (pp. 243-250). IEEE.

[14] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. Ieee Access, 10, 39700-39715.

[15] del Mar Roldán-García, M., García-Nieto, J., & Aldana-Montes, J. F. (2017). Enhancing semantic consistency in anti-fraud rule-based expert systems. Expert Systems with Applications, 90, 332-343.

[16] Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J. D. (2021). The accuracy versus interpretability trade-off in fraud detection model. Data & Policy, 3, e12.

[17] Li, R., Liu, Z., Ma, Y., Yang, D., & Sun, S. (2022). Internet financial fraud detection based on graph learning. IEEE Transactions on Computational Social Systems, 10(3), 1394-1401.

[18] Ren, L., Hu, R., Li, D., Liu, Y., Wu, J., Zang, Y., & Hu, W. (2023). Dynamic graph neural network-based fraud detectors against collaborative fraudsters. Knowledge-Based Systems, 278, 110888.

[19] Siddiqi, N. (2017). Intelligent credit scoring: Building and implementing better credit risk scorecards. John Wiley & Sons.

[20] Mock, T. J., Srivastava, R. P., & Wright, A. M. (2017). Fraud risk assessment using the fraud risk model as a decision aid. Journal of emerging technologies in accounting, 14(1), 37-56.

[21] Nangi, P. R., & Settipi, S. (2023). A Cloud-Native Serverless Architecture for Event-Driven, Low-Latency, and AI-Enabled Distributed Systems. International Journal of Emerging Research in Engineering and Technology, 4(4), 128-136. https://doi.org/10.63282/3050-922X.IJERET-V4I4P11

[22] Bhat, J., Sundar, D., & Jayaram, Y. (2024). AI Governance in Public Sector Enterprise Systems: Ensuring Trust, Compliance, and Ethics. International Journal of Emerging Trends in Computer Science and Information Technology, 5(1), 128-137. https://doi.org/10.63282/3050-9246.IJETCSIT-V5I1P114

[23] Nangi, P. R., Reddy Nala Obannagari, C. K., & Settipi, S. (2023). A Multi-Layered Zero-Trust Security Framework for Cloud-Native and Distributed Enterprise Systems Using AI-Driven Identity and Access Intelligence. International Journal of Emerging Trends in Computer Science and Information Technology, 4(3), 144-153. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P115

[24] Jayaram, Y., & Sundar, D. (2023). AI-Powered Student Success Ecosystems: Integrating ECM, DXP, and Predictive Analytics. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(1), 109-119. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I1P113

[25] Bhat, J. (2023). Automating Higher Education Administrative Processes with AI-Powered Workflows. International Journal of Emerging Trends in Computer Science and Information Technology, 4(4), 147-157. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I4P116

[26] Nangi, P. R., & Reddy Nala Obannagari, C. K. (2024). High-Performance Distributed Database Partitioning Using Machine Learning-Driven Workload Forecasting and Query Optimization. American International Journal of Computer Science and Technology, 6(2), 11-21. https://doi.org/10.63282/3117-5481/AIJCST-V6I2P102

[27] Bhat, J. (2024). Responsible Machine Learning in Student-Facing Applications: Bias Mitigation & Fairness Frameworks. American International Journal of Computer Science and Technology, 6(1), 38-49. https://doi.org/10.63282/3117-5481/AIJCST-V6I1P104

[28] Jayaram, Y., Sundar, D., & Bhat, J. (2024). Generative AI Governance & Secure Content Automation in Higher Education. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(4), 163-174. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I4P116

[29] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2022). Self-Auditing Deep Learning Pipelines for Automated Compliance Validation with Explainability, Traceability, and Regulatory Assurance. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(1), 133-142. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P114

[30] Jayaram, Y., & Bhat, J. (2022). Intelligent Forms Automation for Higher Ed: Streamlining Student Onboarding and Administrative Workflows. International Journal of Emerging Trends in Computer Science and Information Technology, 3(4), 100-111. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P110

[31] Bhat, J., & Sundar, D. (2022). Building a Secure API-Driven Enterprise: A Blueprint for Modern Integrations in Higher Education. International Journal of Emerging Research in Engineering and Technology, 3(2), 123-134. https://doi.org/10.63282/3050-922X.IJERET-V3I2P113

[32] Nangi, P. R., Reddy Nala Obannagari, C. K., & Settipi, S. (2024). A Federated Zero-Trust Security Framework for Multi-Cloud Environments Using Predictive Analytics and AI-Driven Access Control Models. International Journal of Emerging Research in Engineering and Technology, 5(2), 95-107. https://doi.org/10.63282/3050-922X.IJERET-V5I2P110

[33] Jayaram, Y. (2024). Private LLMs for Higher Education: Secure GenAI for Academic & Administrative Content. American International Journal of Computer Science and Technology, 6(4), 28-38. https://doi.org/10.63282/3117-5481/AIJCST-V6I4P103

[34] Nangi, P. R. (2022). Multi-Cloud Resource Stability Forecasting Using Temporal Fusion Transformers. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(3), 123-135. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I3P113

[35] Bhat, J., Sundar, D., & Jayaram, Y. (2022). Modernizing Legacy ERP Systems with AI and Machine Learning in the Public Sector. International Journal of Emerging Research in Engineering and Technology, 3(4), 104-114. https://doi.org/10.63282/3050-922X.IJERET-V3I4P112

[36] Jayaram, Y., & Sundar, D. (2022). Enhanced Predictive Decision Models for Academia and Operations through Advanced Analytical Methodologies. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(4), 113-122. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P113

[37] Bhat, J. (2022). The Role of Intelligent Data Engineering in Enterprise Digital Transformation. International Journal of AI, BigData, Computational and Management Studies, 3(4), 106-114. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P111

[38] Jayaram, Y. (2023). Cloud-First Content Modernization: Migrating Legacy ECM to Secure, Scalable Cloud Platforms. International Journal of Emerging Research in Engineering and Technology, 4(3), 130-139. https://doi.org/10.63282/3050-922X.IJERET-V4I3P114

[39] Reddy Nangi, P., & Reddy Nala Obannagari, C. K. (2023). Scalable End-to-End Encryption Management Using Quantum-Resistant Cryptographic Protocols for Cloud-Native Microservices Ecosystems. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 142-153. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P116

[40] Jayaram, Y., Sundar, D., & Bhat, J. (2022). AI-Driven Content Intelligence in Higher Education: Transforming Institutional Knowledge Management. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(2), 132-142. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I2P115

[41] Nangi, P. R., Obannagari, C. K. R. N., & Settipi, S. (2022). Enhanced Serverless Micro-Reactivity Model for High-Velocity Event Streams within Scalable Cloud-Native Architectures. International Journal of Emerging Research in Engineering and Technology, 3(3), 127-135. https://doi.org/10.63282/3050-922X.IJERET-V3I3P113

[42] Bhat, J. (2023). Strengthening ERP Security with AI-Driven Threat Detection and Zero-Trust Principles. International Journal of Emerging Trends in Computer Science and Information Technology, 4(3), 154-163. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I3P116

[43] Jayaram, Y. (2023). Data Governance and Content Lifecycle Automation in the Cloud for Secure, Compliance-Oriented Data Operations. International Journal of AI, BigData, Computational and Management Studies, 4(3), 124-133. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I3P113

[44] Nangi, P. R., & Reddy Nala Obannagari, C. K. (2024). A Multi-Layered Zero-Trust–Driven Cybersecurity Framework Integrating Deep Learning and Automated Compliance for Heterogeneous Enterprise Clouds. American International Journal of Computer Science and Technology, 6(4), 14-27. https://doi.org/10.63282/3117-5481/AIJCST-V6I4P102

[45] Bhat, J., & Jayaram, Y. (2023). Predictive Analytics for Student Retention and Success Using AI/ML. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(4), 121-131. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I4P114

[46] Nangi, P. R., Reddy Nala Obannagari, C. K., & Settipi, S. (2024). Serverless Computing Optimization Strategies Using ML-Based Auto-Scaling and Event-Stream Intelligence for Low-Latency Enterprise Workloads. International Journal of Emerging Trends in Computer Science and Information Technology, 5(3), 131-142. https://doi.org/10.63282/3050-9246.IJETCSIT-V5I3P113

[47] Jayaram, Y. (2024). AI-Driven Personalization 2.0: Hyper-Personalized Journeys for Every Student Type. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(1), 149-159. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I1P114

[48] Jayaram, Y., & Bhat, J. (2022). Intelligent Forms Automation for Higher Ed: Streamlining Student Onboarding and Administrative Workflows. International Journal of Emerging Trends in Computer Science and Information Technology, 3(4), 100-111. https://doi.org/10.63282/3050-9246.IJETCSIT-V3I4P110

[49] Bhat, J. (2024). Designing Enterprise Data Architecture for AI-First Government and Higher Education Institutions. International Journal of Emerging Research in Engineering and Technology, 5(3), 106-117. https://doi.org/10.63282/3050-922X.IJERET-V5I3P111