Ai-powered fraud detection and prevention

Dimitriu DAVID
Academy of Economic Studies, Bucharest, Romania
DimitriuDavid24@Stud.Ase.Ro

Abstract. *Given the rapid commoditization of generative artificial intelligence and high-velocity digital transactions, the current rules-based models for detecting fraud have become obsolete. Scientific writings which already exists focuses on highly focused and uni-modal approaches such as graph neural networks for detecting camouflage in specific contexts or hybrid models for clustering data in tabular form. These models, though mathematically sound, lack the required multimodal capabilities and low-latency response times required for execution within digital transaction environments. To overcome these encounters, a new four-pillar architecture is proposed, integrating generative artificial intelligence defense mechanisms, machine learning models, threat intelligence systems, and alert management systems. The research investigates whether integrating various open-source machine learning models will outperform existing academic models for detecting fraud in transaction environments. The methodology involves a comparative analysis of the existing academic models and the proposed architecture in terms of processing latency, explainability, and the extent of the protection offered. Results indicate that while academic models effectively detect fraud rings in digital transactions offline, the proposed multimodal model achieves higher throughput and protects the authentication perimeter of the transaction system from synthetic media and prompt injection attacks. Through this research, a critical gap in the literature between academic models and real-world implementations is bridged. This paper provides a comprehensive structure for enterprise systems to develop multimodal, explainable machine learning models for fraud detection in high-velocity digital transactions.*

Keywords: Fraud detection, artificial intelligence, multimodal architecture, machine learning, digital transactions, explainability.

Introduction

Fraud detection has become an existential imperative for the e-commerce and financial sectors across the world. As the number of e-commerce sites and users has grown exponentially, so too has the number of different forms of cyberattacks that target those platforms. Beyond password-cracking and credential-stuffing attacks, many of today's most sophisticated cyberattacks use audio and video deepfakes to gain access to biometric security portals. Additionally, automated account takeover attacks have begun to utilize malicious automation and generative artificial intelligence with such rapidity that current estimates project that these forms of automated attacks are growing at a faster rate than all human internet traffic combined.

Modern cyberattack vulnerabilities cannot be mitigated by the human oversight that currently exists within most e-commerce and financial platforms. These types of attacks occur in milliseconds, providing no time for any human being to recognize the fraudulent activity before the attack results in financial loss for that company. Furthermore, while there have been various proposed solutions to combat e-commerce fraud, such as the utilization of highly complex graph neural networks to recognize instances of fraud rings within e-commerce transactions, those models are often unimodal and contain latency in their processes that prevent them from being effectively utilized within e-commerce environments that handle a high velocity of transactions.

To combat these growing e-commerce and financial platform security challenges, a new architecture can be developed to secure the modern digital ecosystem. Such an architecture can include the implementation of generative artificial intelligence defense mechanisms to recognize deepfake videos and audio files, the implementation of machine learning to create an intelligence system that recognizes fraudulent activity at low latencies, the implementation of cybersecurity intelligence to recognize topological fraud patterns, and the implementation of an alert management system to respond to any detected fraudulent activity. Each of these elements can be orchestrated together into a single, unified digital security system that utilizes various open-source technologies to create a perimeter defense for the digital ecosystem.

In formulating such a plan, it is first necessary to pose the research question that will guide the research for this project: how can e-commerce and financial data platforms both continuously and efficiently defend themselves against these rapidly growing threats associated with cyberattacks and generative artificial intelligence? Based upon this question, it can be hypothesized that implementing such a multimodal, four-pillar architecture will result in a system that can outperform the existing (and complex) academic fraud detection models. Furthermore, through the implementation of each of these advanced technologies, it is likely that the cybersecurity perimeter for financial and e-commerce platforms will be better defended against fraudulent activity.

Literature review

The evolution from uni-modal algorithms to multimodal enterprise architectures

The current sphere of digital fraud detection is defined by an ongoing arms race between increasingly sophisticated hostile actors and automated systems designed to counter them. In the scientific literature, the working concept of modern fraud detection has shifted from static, rule-based search to dynamic, behavioral anomaly detection (Odeyemi et al., 2024). Therefore, machine learning models are now foundational to digital security, tasked with identifying changes in massive transactional datasets (Ikumapayi & Ayankoya, 2025). However, as threat risks evolve, the academic focus has shifted heavily toward researching topological camouflage—the methods by which fraudsters modify their behavioral networks to mimic the behavior of good users. While the existing research provides deep mathematical configurations for identifying these hidden connections, the proposed solutions persist as mainly theoretical and unimodal. To assess the necessity of the four-pillar architecture proposed in this research, it is critical to evaluate the current state of the art in graph neural networks and scalable data pipelines, and to examine the emerging threat of generative artificial intelligence, ultimately revealing the operational gaps this project seeks to resolve.

A leading theme from recent academic papers is the application of Graph Neural Networks (GNNs) to solve the problem of related fraud and camouflage. Standard machine learning models usually fail when analyzing highly imbalanced datasets through which fraudsters intentionally inject noisy topology patterns to evade detection (Huang et al., 2022). To counter this, researchers have developed hyper-focused algorithms. For example, the CARE-GNN model, which uses deep reinforcement learning to dynamically choose optimal origins for dropping different neighbors before data aggregation, effectively eliminates relational camouflage (Dou et al., 2020). Similarly, approaches like PC-GNN employ label-balanced sampling and proximity measures to prevent minority fraud nodes from being drowned out by benign majority nodes (Liu et al., 2021). Furthermore, the GraphConsis model attempts to resolve context and feature inconsistencies by

calculating quantitative consistency scores to filter mismatched neighbors (Liu et al., 2020). While these models deliver profound developments in topological intelligence, they operate in isolation. In the context of the present research, the third pillar of the proposed architecture applies NetworkX's graph-heuristic capabilities to achieve a similar relational mapping. However, unlike isolated academic models, this project embeds topological intelligence directly into a more extensive enterprise ecosystem, guaranteeing that relational data operates as a supporting feature rather than a standalone bottleneck.

Even though GNNs manifest superior skill in detecting camouflage, the term paper identifies a severe limitation: processing latency. The "neighborhood explosion" inherent in the message-passing framework brings hundreds of milliseconds of latency, rendering standard GNNs entirely unsuitable for live, high-velocity transactional environments (Lu et al., 2022). To reduce this, the BRIGHT structure introduced a decoupled inference mechanism, calculating heavy entity embeddings offline while executing lightweight online inference to diminish latency (Lu et al., 2022). Similarly, given the considerable volume of present-day financial data, recent studies point out the need to migrate forecasting analysis to elastic cloud computing environments to prevent on-premises hardware bottlenecks (Amirineni, 2024). Other researchers present integrated architectures that first use unsupervised K-Means clustering to structure the data before feeding it to supervised classifiers such as Random Forests (Chaudhry et al., 2024). These structural concerns directly inform the second pillar of this project's architecture. Rather than relying on computation-intensive, real-time graph aggregations, the proposed system uses Polars, Apache Kafka, and TimescaleDB to build an ultra-low-latency pipeline. By separating elaborate data structuring from the real-time scoring engine, driven by LightGBM, this project achieves sub-millisecond throughput that academic GNNs struggle to provide.

Methodology

The overarching goal of this study is to evaluate the real-world effectiveness of the proposed four-pillar fraud detection architecture as compared to academic state-of-the-art baselines; thus, the core research hypothesis that will be tested is that a decoupled multi-modal framework that combines generative AI defenses with ultra-low latency machine learning pipelines will achieve higher performance both in terms of throughput and security than isolated uni-modal graph neural networks. This hypothesis has been formulated based on the drawbacks of the existing literature identified in the previous sections; namely, while traditional graph models are mathematically sound, they are often infeasible to use in practice due to the neighborhood explosion problem, which causes debilitating latency issues (Lu et al., 2022). By proposing an experimental systems architecture, this research uses a mixed-methods design to test the efficacy of the proposed enterprise architecture as compared to academic state-of-the-art models.

To perform this comparison, a simulated high-velocity e-commerce environment was emulated using a synthetic multimodal dataset that tests both tabular transaction logic and unstructured media defenses. The underlying tabular data comprises benchmark financial logs, such as the IEEE-CIS Fraud Detection dataset, which include features such as anonymized user identifiers, transaction amounts, internet protocol routing, and device signatures. To test topological disguise attacks identified in the prior literature (Huang et al., 2022), synthetic relational edges were injected into the tabular data to simulate collusive fraud rings. Additionally, to test the generative AI detectors at the authentication boundary, the dataset was further augmented with

synthetic unstructured payloads, including simulated prompt injection texts, generated deepfake audio requests, and synthetic phishing chat logs.

Prior to model training, the raw transaction logs were extensively engineered using the Polars library. As discussed previously, given the high-velocity nature of the simulated environment, most data manipulation frameworks introduce unacceptable latency. However, due to its multi-threaded design, Polars efficiently extracted complex temporal features, such as transaction velocity and geographic routing distances, over sliding time windows. Moreover, rather than applying traditional synthetic oversampling methods to address class imbalance, which, as noted in the literature, can introduce severe data leakage in time-series datasets, the extreme class imbalance was handled directly using asymmetric classification weights in the LightGBM algorithm.

The experimental setup is designed to compare the proposed decoupled architecture against academic state-of-the-art baselines such as the PC-GNN (Liu et al., 2021) and the CARE-GNN (Dou et al., 2020) frameworks. To analyze the data and build the proposed high-throughput system, a highly specialized set of software tools and open-source packages was employed. The proposed system's real-time risk engine uses LightGBM and PyOD, with hyperparameters tuned for extreme class imbalance via asymmetric class weights and a restricted tree depth, to achieve sub-10-millisecond inference latency. The streaming data pipeline was built using Apache Kafka with consumer groups feeding into Polars for multi-threaded feature engineering and TimescaleDB for continuous time-windowed aggregations. Hardware resources were also deliberately restricted in the experimental design to mirror real-world, cost-efficient production scenarios rather than theoretical, limitless supercomputer implementations. The 4-stage funnel was restricted to 16GB of system memory and 8GB of virtual VRAM. To meet the inference-time requirements within these hardware constraints, the generative-AI interceptors from Stage 1, such as Fast-DetectGPT, used model weight quantization, batching limits, and an explicit CPU-fallback flag. The proposed architecture demonstrates the ability to achieve state-of-the-art results without supercomputer-level hardware costs. The baseline models were implemented in standard PyTorch environments with multi-hop message-passing layers and deep reinforcement learning thresholds, which require substantial relational data consolidation prior to scoring. To evaluate the statistical significance of results, given the highly imbalanced nature of digital fraud, the quantitative metrics used were the area under the precision-recall curve and the macro F1-score. Standard accuracy was ignored. To avoid data leakage and lookahead bias, a common pitfall for static graph models, a time-split cross-validation scheme was implemented to guarantee that all models were trained only on historical data and evaluated on purely future time segments. Robustness in deployment was quantitatively evaluated through strict latency profiling, measuring the P90, P95, and P99 inference latencies in ms to assess the system's ability to meet the stringent 50ms SLA of modern payment gateways. Qualitatively, algorithmic explainability was assessed by analyzing the stability and local accuracy of SHAP values during the classification phase.

Despite the robustness of the experimental design, several limitations in the methodology can be identified. Firstly, topological graph processing is asynchronous because the NetworkX heuristic mapping is CPU-bound and performed either in an offline setting or in micro-batch periods. There is, therefore, a minor time delay before newly detected camouflage topologies are cycled back into the real-time scoring funnel. Secondly, using synthetic media to test the authentication perimeter introduces an unavoidable bias since synthetic media may not fully capture the unpredictability or rapid evolution of 0-day attacks by highly resourced real-world adversaries. Finally, the multimodal interceptors at the perimeter may introduce false positives when processing low-quality legitimate media or non-standard user inputs. However, the results of

this study remain extremely generalizable to real-world production-scale environments. Software tools such as Apache Kafka, Polars, and TimescaleDB are industry standards for high-throughput distributed systems, enabling the data pipeline to scale horizontally to process millions of transactions a day. Moreover, the modular design of the 4-stage funnel guarantees domain-agnostic adaptability; while tested in an e-commerce simulation, the decoupled machine learning engine and the lifecycle management framework can be rapidly adapted for insurance claims processing, banking authentication, or forensic accounting without modifying the architectural logic (Amirineni, 2024).

Results and discussions

The simulated high-velocity electronic commerce environment was employed to evaluate the proposed architecture, producing clear results regarding the overall hypothesis. The experiments confirm that a decoupled multimodal architecture significantly outperforms unimodal graph neural networks in both performance and overall security surface. The core real-time scoring engine, with the LightGBM classifier and Polars data structure, achieved single-digit millisecond latency even at the highest percentiles. However, the baseline models such as the CARE-GNN model (Dou et al., 2020) and the PC-GNN model (Liu et al., 2021) significantly exceeded the 50 millisecond SLA due to the extreme computational cost of online relational data gathering. Apart from performance and accuracy, significant qualitative differences were observed between the two systems; whereas unimodal graph models returned a probability value, the proposed multimodal architecture provided a deep contextual understanding. With the use of tools like MLflow and Streamlit, the system was able to illustrate the complete attack pathway from network intrusion to transaction execution, greatly reducing the cognitive burden on analysts.

However, an unbiased analysis of the topological detection reveals an important practical trade-off that directly follows from the identified system limitations. Previous academic literature has shown that models such as the AO-GNN (Huang et al., 2022) are mathematically superior at identifying and removing camouflage edges online through native deep reinforcement learning. In the proposed system, the use of NetworkX for heuristic graph mapping was found to be slightly less dynamic in practice as it is performed in offline micro-batch cycles. This introduces a temporal lag that implies a subtle impact on the interpretation of the results; the system may be vulnerable for a small window of time to instantaneous swarm attacks before the offline graph update. This means that the exceptionally high area under the precision-recall curve could potentially overstate real-world results in the very first few minutes of a novel topological attack. Nevertheless, this architectural compromise is necessary to preserve the ultra-low latency that is required for online transaction routing while incorporating the structural intelligence that is valued by previous researchers.

A significant departure from previous academic literature was observed at the authentication perimeter. The proposed first pillar was shown to successfully detect unstructured threats before any transactional data was even generated. To ensure practical relevance, the synthetic media payloads that were used for testing, such as deepfake audio and prompt injections, were generated using state-of-the-art open-source language models and text-to-speech engines that mimic complex attack vectors that have been observed in recent threat intelligence reports. These payloads were sampled through human-in-the-loop to ensure that they would plausibly bypass standard heuristic filters. Previous academic models that operate exclusively on structured tabular clustering (Chaudhry et al., 2024) were unable to detect these generative artificial intelligence

attacks. Although the multimodal interceptors introduce the risk of returning false positives when analyzing low-quality legitimate media, they were found to be significantly more effective at preventing zero-day synthetic attacks.

Finally, the proposed experiment highlights the practical applicability and extensive generalizability of the proposed conceptual framework. Certain design choices were made during the system architecture to ensure that the results are generalizable beyond electronic commerce. Features were designed as general transactional constructs such as velocity, distance, and time since last login as opposed to item-specific metadata. Because the data engineering pipeline employs agnostic data ingestion through Apache Kafka and TimescaleDB, the proposed system can horizontally scale across various industries. This serves to validate the theoretical propositions put forth by Amirineni (2024) regarding cloud-native adaptability, showing that the decoupled machine learning engine can easily be adapted for insurance claims processing or banking authentication. Moreover, the incorporation of SHAP values consistently returns localized tabular reason codes that fulfill the rigid regulatory compliance requirements that are frequently overlooked in purely algorithmic research.

Overall, the results indicate that a shift in academic focus from a singular algorithm to a composite software system is necessary for practical cybersecurity. To overcome the limitations that were identified during the evaluation, future research should investigate the use of streaming graph databases to replace the static NetworkX to remove the lag in topological updates. Future research should also investigate further tuning of the generative artificial intelligence interceptors on live, proprietary datasets to reduce the false positive rate at the authentication perimeter. By successfully filling the knowledge gap between theoretical topological intelligence and practical low-latency deployment, this research provides a verified and deployable blueprint for mitigating the next generation of polymorphic digital crime.

Table 1. Tool Selection Comparison Matrix for Ai Fraud Detection Components

Component	**Selected Tool**	**License**	**GitHub Stars**	**Last Updated**	**Primary Use Case**
Video Deepfake Detection	DeepfakeBench	MIT	600+	March 2026	Face swap & lip-sync detection
Audio Deepfake Detection	AASIST	MIT	N/A	February 2026	Voice cloning & TTS detection
Image Deepfake Detection	SigLIP (HF)	Apache 2.0	N/A	January 2026	GAN & diffusion model detection
AI Text Detection	Binoculars	MIT	N/A	January 2026	Zero-shot LLM content detection
Anomaly Detection	PyOD	BSD 2-Clause	8000+	February 2026	Real-time transaction anomalies

ML Classification	XGBoost	Apache 2.0	25000+	March 2026	Gradient boosting classification
Explainability	SHAP	MIT	22000+	February 2026	Model interpretation & transparency
Dashboard	Streamlit	Apache 2.0	34000+	March 2026	Interactive web dashboard

Source: Authors' own research results

Table 2. Performance Metrics of Anomaly Detection Models on Credit Card Fraud Dataset

Model	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**Training Time (s)**	**Inference Time (ms)**
Isolation Forest	0.96	0.82	0.88	0.94	0.8	0.9
Local Outlier Factor (LOF)	0.89	0.91	0.90	0.92	2.4	12.3
COPOD	0.92	0.87	0.89	0.93	0.3	0.5
AutoEncoder (Deep Learning)	0.94	0.89	0.91	0.95	45.2	3.7
XGBoost Classifier	0.98	0.95	0.96	0.99	12.8	1.2

Source: Authors' own research results

Table 3 summarizes the hardware requirements and resource utilization for each major component of the fraud detection system. These specifications ensure the system can operate on consumer-grade hardware with an RTX 4060 GPU (8GB VRAM) and 16GB system RAM.

Table 3. Hardware Requirements and Resource Utilization by Component

Component	**VRAM (GB)**	**System RAM (GB)**	**Disk Space (GB)**	**CPU Fallback Available**
DeepfakeBench (Video)	6.0	4.0	2.5	Yes (10x slower)
AASIST (Audio)	0.5	1.0	0.2	Yes (4x slower)
SigLIP (Image)	1.2	2.0	0.5	Yes (3x slower)

PyOD + XGBoost	N/A	3.0	0.5	CPU-only (no penalty)
PostgreSQL + Redis	N/A	4.0	20.0	CPU-only
Total System	**7.7**	**14.0**	**23.7**	—

Source: Authors' own research results

Table 4. Dataset Specifications for Model Training and Evaluation

Dataset Name	**Use Case**	**Size**	**Class Balance**	**Source**
Kaggle Credit Card Fraud	Transaction fraud	284,807 records	0.17% fraud	Kaggle/ULB
FaceForensics++	Video deepfakes	5,000 videos	50% real/fake	Technical University Munich
ASVspoof 2024	Voice anti-spoofing	50,000+ utterances	Variable by track	ASVspoof Consortium
HC3 (Human ChatGPT)	AI text detection	87,000 Q&A pairs	50% human/AI	HuggingFace
PaySim	Mobile payment fraud	6.3M transactions	0.13% fraud	Kaggle/Synthetic

Source: Public dataset repositories (Kaggle, HuggingFace, ASVspoof.org, FaceForensics++)

Conclusion

The accelerating proliferation of generative artificial intelligence and the blistering speed of online transactions have motivated a re-examination of the current state of fraud detection. In particular, this paper addresses a significant knowledge gap between academic research (which typically focuses exclusively on graph neural networks) and practitioners' deployment needs. As such, the novelty of this paper lies in the design and evaluation of a four-pillar multimodal architecture that deploys generative artificial intelligence interceptors at the authentication perimeter and an ultra-low-latency machine learning engine (via Polars and LightGBM) that enables sub-ten-millisecond inference while fully protecting against zero-day synthetic media attacks. As such, the significance of this paper lies in its departure from academic research (which

focuses on unimodal algorithms that detect relational camouflage post-hoc) toward a fully integrated, deployable software package that can actively neutralize threats in real time and satisfy regulators through tabular explainability.

Beyond e-commerce, the proposed architecture has far-reaching potential in a variety of high-throughput applications. Since the system utilizes an agnostic, enterprise-grade data streaming infrastructure (such as Apache Kafka and TimescaleDB), the machine learning engine and the entire lifecycle management framework can be trivially translated. For example, in the United Kingdom, large banks are frequently targeted by deepfake audio attacks that aim to circumvent voice authentication systems. By deploying the proposed architecture, these organizations can intercept and neutralize such attacks in real time, thereby preventing unauthorized access to customer accounts. As another example, insurance companies can leverage SHAP values and the system's heuristic graph mapping to explain collusive claims fraud to regulators; indeed, recent investigations into fraudulent auto insurance rings in the United States revealed that machine learning models identified non-obvious relational connections among claims.

Future work on this paper should explore incorporating decentralized threat intelligence feeds into the alert management dashboard, thereby allowing the entire pipeline to learn about global social engineering patterns and novel prompt-injection vectors on its own. Overall, the proposed architecture works well, but several limitations of the present research must be noted. Firstly, the use of NetworkX for topological graph mapping introduces an asynchronous temporal lag; since this heuristic processing is CPU-bound, it occurs in micro-batch chunks rather than in real time, the system will remain vulnerable to instantaneous swarm attacks until the relational context has been fully updated. Secondly, deploying zero-shot models at the authentication perimeter entails the risk of false positives; for example, the system may inadvertently flag low-quality (but genuine) user media or nonstandard text inputs. Future research should thus prioritize the use of continuous streaming graph databases to eliminate lag in heuristic processing, as well as the tuning of generative artificial intelligence interceptors on proprietary domain-specific datasets in order to ensure maximum precision and minimal user friction in live production environments.

Amirineni, S. (2024). Leveraging machine learning, cloud computing, and AI for fraud detection and prevention in insurance. *Preprint*.
Chaudhry, R., & Kaur, S. (2024). Fraud detection and prevention for a secure financial future using artificial intelligence. *IEEE Xplore*.
Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 315–324.

Huang, M., Liu, Y., Ao, X., Li, K., Chi, J., Feng, J., Yang, H., & He, Q. (2022). AUC-oriented graph neural network for fraud detection. *Proceedings of the ACM Web Conference 2022*, 1311–1321.

Ikumapayi, O. J., & Ayankoya, A. (2025). AI powered forensic accounting: Leveraging machine learning for real-time fraud detection and prevention. *International Journal of Research Publication and Reviews*, 6(2), 236–250.

Kopperapu, R. (2025). AI-powered fraud detection and prevention system. *International Journal of Engineering Sciences and Advanced Technology*, 25(1).

Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. *Proceedings of the Web Conference 2021*, 3168–3177.

Liu, Z., Dou, Y., Yu, P. S., Deng, Y., & Peng, H. (2020). Alleviating the inconsistency problem of applying graph neural network to fraud detection. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1569–1572.

Lu, M., Han, Z., Rao, S. X., Zhang, Z., Zhao, Y., & Shan, Y. (2022). BRIGHT: Graph neural networks in real-time fraud detection. *arXiv preprint arXiv:2205.13084*.

Muhammad, A., Dina Tbaishat, Amril Nazir, Seif Yacoub, Mustafa Abdul Razek, Mohamed Ahmed Abo El-Enen & Ahmed T. Sahlol. (2025). Fraud detection and explanation in medical claims using GNN architectures. www.nature.com/scientificreports.

Odeyemi, O., Nwankwo, C., . (2024). Reviewing the role of AI in fraud detection and prevention in financial services. *World Journal of Advanced Research and Reviews*.

Odufisan, O. I., Abhulimen, O. V., & Ogunti, E. O. (2025). Harnessing artificial intelligence and machine learning for fraud detection and prevention in Nigeria. *Journal of Economic Criminology*, 7, 100127.